SCIENTIFIC AMERICAN

KALEIDOSCOPES

$2.50

December 1985


THE NEW SAAB 9000 HAS ACHIEVED AN IDEAL BALANCE.
75% SPORTS CAR AND 75% LUXURY SEDAN.

ell your Saab 9000 what temperature makes you feel the most chipper
nd it, through microprocessors, keeps you happy.

Size: People in government (the EPA) charged with such things
ave declared the Saab 9000 a "large" car.

Considering the legroom, the elbowroom, and the carrying space
up to 56.5 cu. ft.) that may even be a mite bit of an understatement.

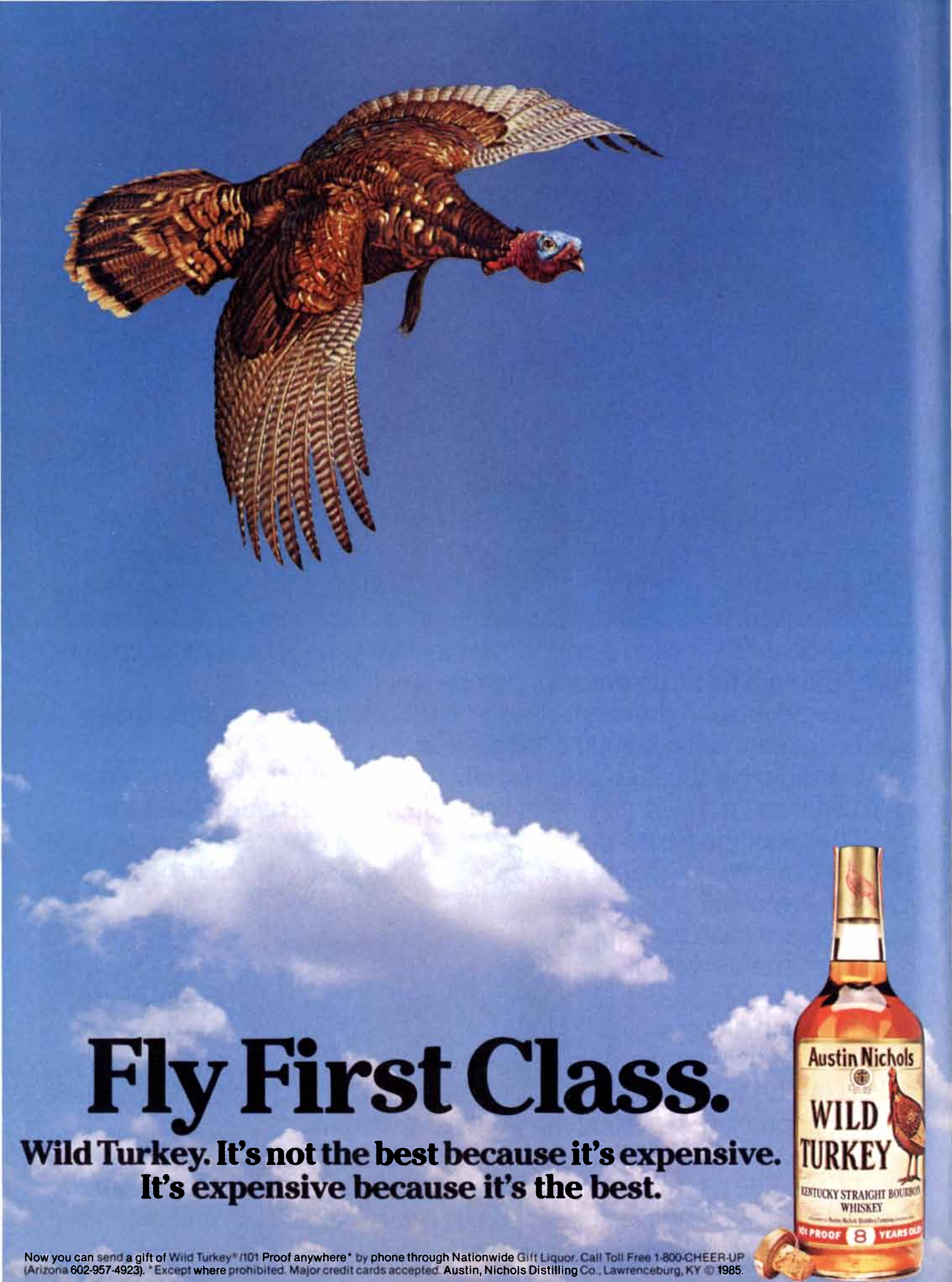
The new Saab 9000.

The touches: Seats
nd a steering wheel that
djust to even the most
xtraordinary anatomy.

Instruments within
our reach and so well-
hought-out that your
yes, so used to being as-
aulted on the road, will
ome to appreciate the soothing green illumination.

The exterior? Well, check the pictures on these pages for a few sec-
nds. Then see a new Saab 9000 for yourself at a Saab dealer, where a
erfect balance always exceeds the sum of its parts.

The most intelligent cars ever built. SAAB


Fly First Class.
Wild Turkey. It's not the best because it's expensive.
It's expensive because it's the best.
Austin Nichols
WILD
TURKEY
Now you can send a gift of Wild Turkey® /101 Proof anywhere* by phone through Nationwide Gift Liquor. Call Toll Free 1-800-CHEER-UP (Arizona 602-957-4923). *Except where prohibited. Major credit cards accepted. Austin, Nichols Distilling Co., Lawrenceburg, KY © 1985.

Established 1845 **SCIENTIFIC AMERICAN** December 1985 Volume 253 Number 6

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address

Street ______________________

City ______________________

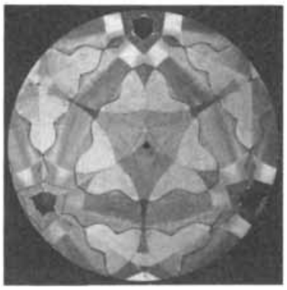
State and ZIP ______________________

Old Address

Street ______________________

City ______________________

State and ZIP ______________________



THE COVER

The photograph on the cover shows an image generated by a polarizing kaleidoscope. Bits of clear and differently stressed plastic at the end of the kaleidoscope tube farthest from the eyepiece change the polarization of the light entering the instrument through a polarizing filter. A second polarizing filter is at the end nearest the eyepiece. As the tube is rotated, different pieces of stressed plastic come into view. Because each piece changes the polarization of light in a distinctive way, some of the colors in the spectrum of light get through the second polarizing filter and some are blocked. The viewer sees a changing panoply of colors. Other modern kaleidoscopes combine imaginative design with electronic technology. In one version sound impulses control an array of light-emitting diodes; the display shimmers and changes in response to sounds (see "The Amateur Scientist," page 134).

THE ILLUSTRATIONS

Cover photograph by Quesada/Burke

Page	Source
8	SCIENTIFIC AMERICAN
16–22	Andrew Christie
47–52	Hank Iken, Walken Graphics
53	U.S. Navy
55	A. V. Mamaev and N. A. Melnikov
56	Ian Worpole
57	A. V. Mamaev (*top left*), Ian Worpole (*top right and bottom*)
58–59	Ian Worpole
60–61	Theo Weber, Max Planck Institute for Behavioral Physiology, Seewiesen, West Germany
62–68	Patricia J. Wynne
84	Ilil Arbel
85	George Janossy, Royal Free Hospital, Hampstead, England
86	Jeffrey Laurence and Powers Peterson, Cornell Medical College (*top*); Ira Schierem, Cornell Medical College (*bottom*)
87–91	Ilil Arbel
92	Luc Montagnier, Pasteur Institute, Paris
93	Ilil Arbel
95	Peter H. Schultz
96	Planetary Image Center, Lunar and Planetary Institute, Houston (*top*); U.S. Geological Survey, Flagstaff (*bottom*)
97–99	Planetary Image Center, Lunar and Planetary Institute, Houston
100–102	Andrew Christie after Peter H. Schultz
105	Gabor Kiss
106–107	Edward Bell
108	Gabor Kiss
109	Edward Bell
110–112	Gabor Kiss
114–115	Edward Bell
117–118	Vaclav Smil, University of Manitoba
120–124	Joan Starwood
127	Lothar Haselberger (*top*), Raymond V. Schoder, S.J. (*bottom*)
128	Alan D. Iselin
128A	George Niemann (*top*), Alan D. Iselin (*bottom*)
128B	Alan D. Iselin (*left*), Hans Georg Bankel (*right*)
129	Alan D. Iselin
130	Lothar Haselberger (*top*), Alan D. Iselin (*bottom*)
132	Alan D. Iselin
135	Quesada/Burke
136–145	Michael Goodman

LETTERS

To the Editors:

On the subject of "Great Guns" ["Science and the Citizen," SCIENTIFIC AMERICAN, October] it should be noted that the electromagnetic railgun and coilgun were conceived, studied and tested by various organizations over a number of decades before the present researchers entered the scene. As one example, M. Fauchon-Villeplé of France developed the concept of a railgun in about 1918 as a method of launching winged projectiles, using twin bus bars as launching rails. Before the end of the war a working model had apparently been assembled and tested.

As to the coilgun, or solenoid gun, two examples from World War II are those of Otto Muck, a consultant to Siemens, and the Hänsler gun of the Gesellschaft für Gerätbau. These were the subject of a considerable amount of theoretical work. The Luftwaffe issued a contract for development of the Hänsler type, with work to begin in February, 1945.

On the subject of chemically powered weapons, the statement that limitations on safe bore pressures restrict such guns to a projectile weight of roughly 100 kilograms is not a general truth. A simple matter of history: the 80-centimeter Dora Gerat, a German railway gun of World War II, fired a 7,100-kilogram projectile to a 38-kilometer range and a 4,800-kilogram one to 47 kilometers. American work with a converted Navy 40.6-centimeter (16-inch) gun some years ago clearly showed that a gun-type launcher designed for the purpose could probably place a projectile in orbit.

Note that artillery barrels seldom approach the state of the art in the matter of peak bore pressures because they are expected to have a considerable lifetime, which would be reduced by the temperatures accompanying the rather higher pressures that could otherwise be used. Another type of "chemically powered" weapon, the shaped charge, can produce a jet with velocities in excess of 11 kilometers per second.

It is true enough that the molecular weight of conventional gun propellant gases limits the maximum velocity of the usual single-stage gun system, but it can be shown that for velocities of up to about .7 to .8 kilometer per second such guns are more efficient than rockets in terms of projectile kinetic energy obtained per unit of propellant mass. In boosting from this level to higher velocities the rocket has an advantage that increases with greater velocity. A hybrid gun-rocket system greatly decreases the deadweight problem inherent in the pure rocket, because a great proportion of the fuel in a pure rocket is expended quite early in accelerating to the one-kilometer-per-second level.

The "chemically powered" launching system should not be underrated, whatever the enthusiasm for the concept of the electromagnetic gun.

JAMES H. JOHNSON

Tampa, Fla.

To the Editors:

In "Slips of the Tongue" [SCIENTIFIC AMERICAN, September] Michael T. Motley makes several errors (I would be wrong to call them slips of the word processor) about Freud's understanding of *lapsus linguae*.

Motley dismisses Freud's discovery that all slips have hidden meanings, and refers to his daughter's mealtime request to "help cut up my meef." Another of Freud's discoveries was the wish of all three-to-six-year-olds to supplant the parent of the same sex in the sexual attentions of the parent of the opposite sex. Motley is quite right in assuming that his six-year-old most likely merged "meat" and "beef," but he misses the sexual implications of these words.

Further, Motley asserts "the 'Freudian slip' is difficult to examine in the laboratory." He ignores the laboratory of the psychoanalytic method of free association, in which the analysand is encouraged to report the spontaneous flow of his or her thoughts, no matter how irrelevant, unimportant or unpleasant they seem. Inferences gleaned by the trained observer are neither less nor more valid than those taken out of any other laboratory.

The use of this method to understand slips of the tongue, pen or word processor, as well as jokes, dreams and certain painful mental symptoms, has both provided the treatment of choice for a limited number of mental illnesses and immeasurably enriched mankind's understanding of itself.

STEVEN A. AGER, M.D.

Philadelphia, Pa.

To the Editors:

The difficulty of examining "Freudian slips" in the laboratory has not been in attributing Freudian interpretations to slips after the fact. With enough imagination one can read coincidental hidden meanings into verbal slips at will. What has been difficult to test is Freud's hypothesis that psychodynamic conflict (hidden meaning) actually causes or contributes to the production of the slips in the first place. Testing this hypothesis requires that one begin with knowledge about the speaker's psychodynamic status and then make a priori predictions about the resulting verbal slips.

As I reported in my article, recent tests of the hypothesis have been conducted under the controlled laboratory conditions required. With speaker anxiety manipulated as an independent variable and with predicted effects on verbal slips observed, evidence has been obtained in support of Freud's hypothesis that conflict can indeed influence the production of verbal slips (at least in some cases).

To find that some slips can be motivated as Freud suggested does not mean that all slips are so motivated, however. On the other hand, neither does my having asserted that most slips are not so motivated make it the case that they are not. Certainly, however, the spreading-activation explanations I discussed can account for the genesis of most verbal slips without involving nonlinguistic or message-independent influences, and thus provide a more elegant theoretical explanation than does Freud's hypothesis. This does not mean that Freudian interpretations such as those of Dr. Ager are definitely wrong; rather, it means that they are unnecessary. For example, it may be that I missed "the sexual implications" of *meef,* or it may be that Dr. Ager invented the sexual implications of *meef.* The post hoc argument that *beef* and *meat* can serve for slang sexual references is hardly persuasive in this case, given that six-year-olds would not be expected to know those slang usages. (The phallic connotation of *meat* or *beef* is not innate, since it is absent in many other cultures.)

Since some psychoanalysts are as defensive about Freud's omniscience as some fathers are about their daughter's innocence, debates over subjective interpretations such as these are not easily resolved. There is, of course, no reason that psychoanalytic and psycholinguistic views of verbal slips cannot be complementary. For example, some clinical psychologists have begun to supplement traditional clinical diagnostic and therapeutic methods, such as the one advocated by Dr. Ager, with the verbal-slip-induction and double-entendre techniques developed for the research I reported.

MICHAEL T. MOTLEY

Davis, Calif.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

DECEMBER, 1935: "Thirty years ago Einstein announced the principle that mass and energy are equivalent. This means that one gram of matter is equivalent to 9×10^{20} ergs of energy. On this view matter is a latent or 'frozen' form of energy. If we could convert the energy into a useful form such as kinetic energy or heat energy or electrical energy, enormous quantities would be available from small amounts of matter. All of this has been the object of much speculative peering into the future."

"There are two stages of success in the solution of a scientific problem. Success in the first means that, knowing the facts by observation, we are able to interpret and explain them upon general principles. Success in the second would signify that, starting with the general principles alone, we could predict the facts themselves. It may be fairly claimed that astrophysics is now advancing to the second stage. We not only can explain the major properties of the stars in terms of those of atoms but also, starting with the general properties of atoms and the law of gravitation, can say, 'If there are masses of matter scattered through space, only the large ones will be hot enough to be visible at the distances measured in light-years.'"

"Observers of the Astrophysical Observatory of the Smithsonian Institution have finally proved that the sun is a variable star. Its heat varies from day to day and month to month."

"Prof. A. V. Vallois has surveyed knowledge relating to the diseases of ancient man. The general conclusion is that it is an error to suppose that our ancestors, living a wild and savage life, had acquired a greater resistance to disease than we have. There is, however, a difference in the diseases that were most prevalent, and this distinction is to be observed not only as between modern man and Neolithic man but also as between Neolithic man and Paleolithic man."

"Chemistry has come to the aid of the oil industry since it was discovered that a dormant or exhausted oil-well can be revived by pumping hydrochloric acid down the pipes to the floor of the oil pocket. The acid attacks the calcareous formations, developing considerable gas pressure and ejecting oil not readily obtained by pumping. But here the chemist is called upon to perform another bit of legerdemain: the acid must not attack the iron of the pipe or pumps. Here the chemist uses substances that he calls inhibitors. Just as there are catalysts that accelerate certain chemical reactions, so these substances retard the action."

"Five years ago Thomas Midgley, Jr., of the General Motors Research Laboratories reported the discovery of the nontoxic refrigerants di- and trichlorodifluoromethane and predicted that the application of these compounds to air conditioning would greatly broaden the scope of that industry. The prediction has been verified. The total air conditioning in use in the U.S. is now more than 16 times what it was in 1930."

SCIENTIFIC AMERICAN

DECEMBER, 1885: "The entire civilized world has for some time past been watching with intense interest the experiments on the treatment of hydrophobia conducted by the celebrated French scientist Dr. Louis Pasteur. These researches have now been so far completed that the results have been presented by the investigator to the French Academy of Sciences. The first step in these investigations, as reported by cable to the *Herald,* was the inoculation of a rabbit with a fragment of tissue taken from the spine of a rabid dog. The incubation of the poison occupied fifteen days. As soon as the animal died a portion of its spinal marrow was in turn inoculated into a second rabbit, and the process continued until sixty rabbits had been treated. Each inoculation increased the power of the virus, so that the last incubation occupied but seven days. Since dried air diminishes the power of the virus, the spinal marrow of the inoculated rabbits was kept in bottles of dried air. In beginning his experiments, therefore, M. Pasteur inoculated his subject with the old tissue and finished the operation by the injection of tissue that had been bottled only two days. These experiments have been very successful, for after such an inoculation the subject is found to be entirely proof against hydrophobia."

"Major John W. Powell, Chief of the Geological Survey, who has been about a month in the field, has discovered in New Mexico, near California Mountain, what he pronounces to be the oldest human habitation upon the American continent. The mountains in this vicinity are covered with huge beds of lava, in which prehistoric man and his comrades excavated square rooms, which were lined with a species of plaster made from the lava, and in these rooms were found numerous evidences of quite an advanced civilization."

"It has truly been said that we talk in a metaphor of 'the harvest of the sea,' but we have only lately been able to realize all that the metaphor means. The Fisheries Exhibition in London in 1883 did a great deal to encourage the study of marine biology, and it is with no small degree of satisfaction that in this much needed work the United States ranks second to no other country. The people of the United States should understand more about the food, habits, spawning and propagation of our fishes, in rivers, lakes and the sea, in order that the harvest may not grow less as the demand becomes more urgent."

"Time has fully justified the enterprise of Dr. Ferdinand V. Hayden in urging upon Congress the project of the creation of the Yellowstone Park. It is certain that in the future the park will become one of the most popular resorts for tourists of all nations, who will be amply repaid by an inspection of one of the few remaining regions of the earth where thermal activity still reaches the surface, and of the grand and impressive scenery which surrounds it."

Louis Pasteur observing an inoculation

Now business

an feel positive again about communications.

We've brightened our customers' outlook. They've found the companies of Ameritech make the complex world of communications easier to deal with.

Nobody understands communications better. Our companies send more messages to more people in more ways than any other company in our region.

Nobody offers a broader scope of full-ranging services. There's virtually nothing we can't provide for the movement of information. Everything from fiber optics and digital technology to Yellow Pages directories that reach millions of readers every day.

You can count on us to meet your future needs as well. Nobody is moving faster than Ameritech to bring more and newer communications services to our region. Nobody is making a bigger investment to accomplish that purpose. Some $3.7 billion during the last two years.

We're the Bell companies of Illinois, Indiana, Michigan, Ohio and Wisconsin. We also serve you through leading companies in: Cellular mobile phone service. Voice and data products and systems. Lease financing. Yellow Pages directory publishing. And the development of technology that creates new business opportunities and improves our communications network.

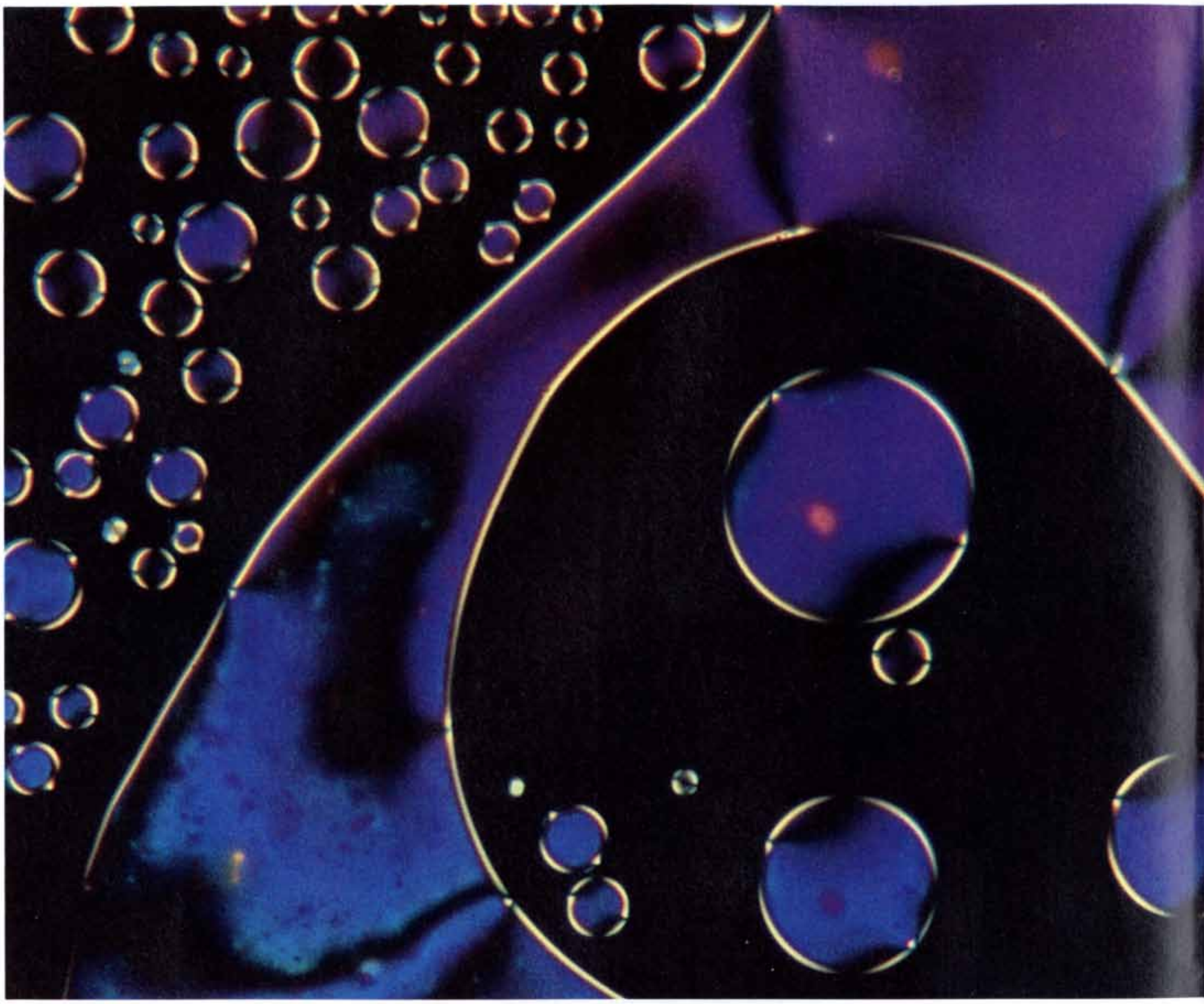
Through products and services that perform reliably and efficiently, Ameritech is taking the worry out of communications. And that's positively a plus for our customers.

Let us help you, too. Contact any of our companies directly or phone us: 1 800 562-2444.

AMERITECH

AMERICAN INFORMATION TECHNOLOGIES

Helping you communicate.SM



This microphoto reveals the fluid nature of liquid crystals—a property Hitachi has put to colorful use in digital clocks, computer/television video screens, and a variety of innovative display devices.

THE AUTHORS

HERBERT LIN ("The Development of Software for Ballistic-Missile Defense") is a postdoctoral research fellow at the Center for International Studies of the Massachusetts Institute of Technology. He is a graduate of M.I.T., which awarded him an S.B. in 1973 and a Ph.D. in 1979. After teaching physics at M.I.T. and the University of Washington he moved to Cornell University, where he was an instructor in physics and a visiting fellow in the Peace Studies Program. He returned to M.I.T. in 1984. Lin's main research interest is the relation of technology to national security policy; in his leisure he indulges a fondness for international folk dancing.

VLADIMIR V. SHKUNOV and BORIS YA. ZEL'DOVICH ("Optical Phase Conjugation") work at the Institute for Problems in Mechanics (IPM) of the Soviet Academy of Sciences. Shkunov was educated at the Moscow Physical-Technical Institute and earned a Candidate of Science degree in 1979. He then joined the staff of the IPM, where he investigates problems of nonlinear and linear optics. Zel'dovich was graduated from Moscow State University in 1966 and was granted a Candidate of Science degree by the Institute of Theoretical and Experimental Physics in Moscow in 1969. He later worked at the P. N. Lebedev Physical Institute in Moscow, where he got a doctor of science degree in 1980. He took his present position in 1981. Zel'dovich and a group of colleagues were awarded the State Prize of the U.S.S.R. for their research in optical phase conjugation.

FRANZ HUBER and JOHN THORSON ("Cricket Auditory Communication") are respectively department director and consultant at the Max Planck Institute for Behavioral Physiology in Seewiesen, West Germany. Huber is a graduate of the University of Munich, which awarded him a doctorate in 1953 for a dissertation on the nervous system of crickets. From 1954 to 1960 he was assistant professor of animal physiology at the University of Tübingen. In 1963 he was appointed professor of zoology and animal physiology at the University of Cologne. Before moving to Seewiesen he served as director of research at the International Institute for Insect Physiology and Ecology in Nairobi. Huber attributes his lifelong interest in animal behavior to a childhood spent on a farm. Thorson holds a B.S. (1955) and an M.S. (1958) in physics from the Rensselaer Polytechnic Institute. He took a position at the General Electric Company, where one of his projects involved biophysics. The experience led him to pursue graduate studies in biology at the University of California at Los Angeles; he got a Ph.D. in zoology in 1965. From 1967 to 1969 he was on the faculty of the University of California at San Diego as assistant professor and research scientist. Since 1969 he has held visiting appointments and consulting contracts at the universities of Oxford and Frankfurt and at the Max Planck institutes. He has collaborated in research on insect and human vision, muscle contraction, sensory transduction and insect behavior. When he is not consulting on the Continent, Thorson writes, "my neuroscientist wife and I are based in our Oxfordshire cottage, where we write, solve nonlinear differential equations and mend old clocks."

JEFFREY LAURENCE ("The Immune System in AIDS") is assistant professor of medicine at the Cornell University Medical Center and assistant attending physician at New York Hospital. He earned a B.A. at Columbia University and an M.D. at the University of Chicago Pritzker School of Medicine. While he was in medical school he was elected a Rhodes scholar and a Henry Luce Foundation scholar. He accepted the latter honor and spent a year investigating immunologic defenses against tumor growth at the Institute for Cancer Research of Osaka University. After returning to the U.S. he completed his medical training at New York Hospital and went on to do research at Rockefeller University. He joined the Cornell faculty in 1982. Laurence has written *Many Happy Returns,* a "medical-detective murder mystery" play produced in 1982.

PETER H. SCHULTZ ("Polar Wandering on Mars") is associate professor of geological sciences at Brown University. He got his B.A. at Carleton College in 1966 and his Ph.D. from the University of Texas at Austin in 1972. From 1973 to 1975 he was a research associate at the Ames Research Center of the National Aeronautics and Space Administration. After a year as a research associate at the University of California at Santa Clara he joined the staff of the Lunar and Planetary Institute in Houston, where in 1981 he was made a senior staff scientist and director of the Planetary Image Center. In 1984 Schultz accepted his position at Brown; he also serves as science coordinator for the NASA-Ames Vertical Gun Range, a national facility where impact cratering is modeled, and as director of the Northeast Planetary Data Center.

DANIEL GORENSTEIN ("The Enormous Theorem") is Jacqueline B. Lewis Professor of Mathematics at Rutgers University. He was educated at Harvard University, where he received an A.B. in 1943 and a Ph.D. in 1950. He began his career at Clark University as assistant professor and became full professor there in 1959. In 1964 he moved to Northeastern University. After spending a year at the Institute for Advanced Study he joined the faculty at Rutgers in 1969. In 1972 he was named a Guggenheim fellow and a Fulbright scholar, and in 1978 he served as Sherman Fairchild Distinguished Professor at the California Institute of Technology. He has been studying finite, simple groups since 1960, and he played a key role in the development of the classification proof of all finite, simple groups. Gorenstein is now formulating a more concise, "second generation" proof.

VACLAV SMIL ("China's Food") is professor of geography at the University of Manitoba in Winnipeg. After getting a degree at Carolinum University in Prague he worked in a regional planning office as a consultant in energy and environmental affairs. Following the Soviet invasion of Czechoslovakia in 1968, he came to the U.S. He earned his doctorate from Pennsylvania State University in 1972 and then joined the faculty of the University of Manitoba. Since his student days in Prague, Smil has pursued his interest in the interrelations of food, energy and changes in the environment.

LOTHAR HASELBERGER ("The Construction Plans for the Temple of Apollo at Didyma") is an archaeologist who specializes in Hellenic architecture. He studied architecture, city planning and the history of architecture at the Technical University of Munich and, as a Fulbright scholar, at Harvard University. Under the sponsorship of the Technical University of Munich he spent two years studying an unusual class of ancient towers that are found on the Greek islands; he was recently given his doctorate for that work. Since 1980, with the support of the German Archaeological Institute and the German Research Society, Haselberger has been recording and cataloguing the construction plans he discovered at Didyma.

COMPUTER RECREATIONS

The search for an invisible ruler that will help radio astronomers to measure the earth

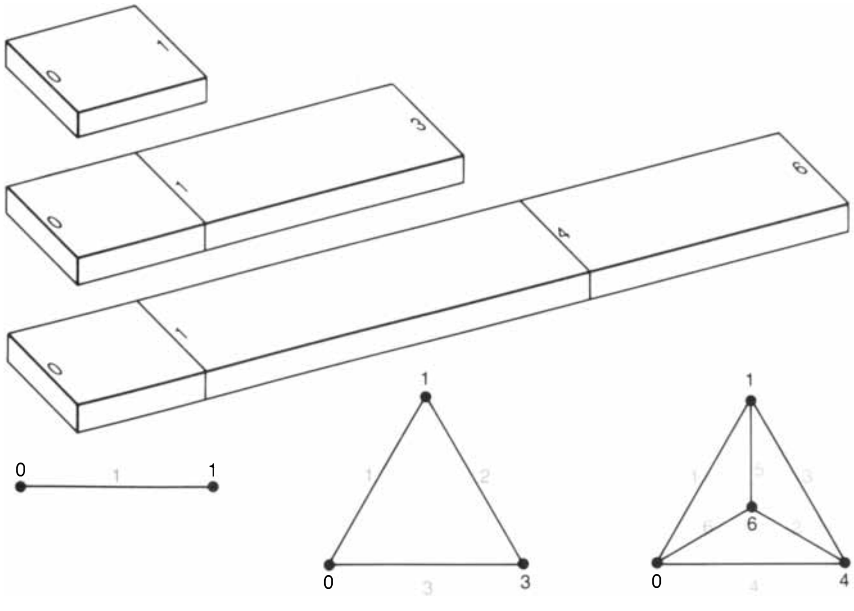
by A. K. Dewdney

A simple ruler one foot long, bearing 13 inch marks, can measure 12 discrete lengths. Is it possible to improve this familiar device so that it measures more lengths than there are marks on the ruler? The answer is yes: it is possible to remove all but five marks from the standard ruler and still measure 10 distances with it. Each distance will be found between some pair of marks as the difference between the integers that label them. It is even possible to achieve the same result with an 11-inch ruler. Readers who puzzle over this exercise and finally succeed will have created a Golomb ruler.

The search for such rulers is an engaging task in which the computer can be useful. What elevates the project from a curiosity to a first-class conundrum is that the need for Golomb rulers emerges from a variety of scientific and technical disciplines.

The devices are the invention of Solomon W. Golomb, professor of mathematics and electrical engineering at the University of Southern California. For two decades he and a handful of colleagues have sought the rulers and studied their properties. The rulers may be applied in coding theory, X-ray crystallography, circuit layout and radio astronomy.

Among the investigators whose work Golomb rulers enlarge and expedite is Douglas S. Robertson, a geophysicist who works for the U.S. National Geodetic Survey of the National Oceanic and Atmospheric Administration in Rockville, Md. He uses the radio-astrometric technique known as very-long-baseline interferometry (or VLBI for short) not to map radio sources but to make finely tuned measurements of the earth. Having sought the rulers himself for a number of years, he unhesitatingly appeals to readers to widen the search. The result may be both a more accurate determination of the size, shape and motion of our planet and a more intriguing time spent thereon.



The three perfect Golomb rulers and their distance diagrams

Before trying to answer Robertson's appeal it is worth mastering the principles that underlie Golomb rulers. Although the rulers come in all sizes, only the smaller ones are known. The first three rulers can be described somewhat abstractly by three sequences of numbers:

0, 1
0, 1, 3
0, 1, 4, 6

To construct the physical ruler mark the left end of a blank ruler that is n units long with the smallest number (0) in the sequence. Inscribe the right end with the largest number (n). The largest number can be 1, 3 or 6. Intervening integers should accompany marks placed at appropriate intervals from the 0 end of the ruler [*see illustration on this page*].

A simple way to check which possible distances from 0 to n can be measured by one of these small rulers is to draw the ruler's distance diagram. For each integer on the ruler mark a point on a sheet of paper and label it with the integer. Then join each pair of points by a line that is labeled with the difference between the integers at its endpoints. If every integral distance encompassed by 0 through n appears on only one line of the distance diagram, the Golomb ruler is said to be perfect. The three rulers in the illustration are all perfect, a fact that can be verified by a glance at their distance diagrams. In each diagram no distance appears more than once and every possible distance between 0 and n is present.

Perfection is rare among Golomb rulers. In fact, the only perfect rulers that exist have just been described. For values of n higher than 6 imperfection is manifested in one of two ways: either a distance occurs more than once or it does not occur at all.

This is a cruel reality to face so early in the search for bigger (and better) rulers. How do we know that no larger perfect rulers exist? Golomb has supplied a proof that is short as well as charming.

His suggestion is that one consider not the marks on a ruler but the intervals between them. If the ruler is perfect, it turns out that the intervals between consecutive marks must provide the distances 1, 2, 3,... up to $m - 1$ in some order (m is the number of marks). Golomb asks: "Where is the one-unit space?" If it is next to any space whose length is less than $m - 1$, then the two spaces together yield a distance that is less than m. This dis-

It's designed to before you

It began with the swift wedge shape of the Honda Prelude. Then, we souped it up.

Introducing the Prelude Si. It has a new front air dam. A rear spoiler. Dual exhausts. A bold taillight strip. Alloy wheels. Michelin steel-belted radials.

But head-turning looks are only the start.

Under the hood of the Prelude Si is a new 12-valve, 2.0 liter engine. It has 110 horsepower and timed sequential, multi-port programmed fuel injection.

Translation: It has more kick.

The Prelude Si also has more in the way of comfort. Settle into the contoured driver's

HONDA
2.0Si
PRELUDE

tance must already occur as a space elsewhere because all distances from 1 to $m - 1$ occur between consecutive marks. Such reasoning forces us to accept the somewhat startling conclusion that the one-unit space is next to the space that is $m - 1$ units long. Moreover, there is no space on the other side of the unit space. It lies at one end of the ruler.

The foregoing argument constitutes one turn of the crank on what some mathematicians call a sausage machine. "Where is the two-unit space?" Golomb asks. The crank is turned again: if the two-unit space is next to any space whose length is less than $m - 2$, the two spaces together produce a distance that already occurs elsewhere. This time we cannot conclude that the two-unit space is next to a space of length $m - 2$. Their combined length, m, is already measured jointly by the one- and $(m - 1)$-unit spaces at one end of the ruler.

The sausage machine grinds to a halt, producing the conclusion that there is only one space the two-unit space may lie next to, namely the one whose length is $m - 1$. Since any ruler has only two ends, a perfect ruler has at most three spaces, 1, $m - 1$ and 2. The proof is complete when we realize that three spaces require four marks: $m = 4$. The spaces are therefore 1, 3, 2; the corresponding marks are 0, 1, 4, 6.

Faced with a complete lack of perfect rulers that have more than four marks, a mathematician will cut the losses by constructing a new definition. What might be called the "next-best syndrome" is thereby demonstrated: the next-best thing to an impossible perfect five-mark ruler might be one that contains each distance only once but does not contain all the distances a perfect ruler of the same length would have. Since this condition is easily met by allowing a ruler to be long enough, a rider is attached. Among all five-mark rulers that contain each distance at most once, determine the shortest one. Such a ruler is called a Golomb ruler of order five. Golomb rulers of order m are defined in the same way. Since the definition includes the first three rulers as a special case, it bridges the awkward discontinuity in perfection beyond four marks.

Herbert Taylor, a colleague of Golomb's, has summarized the current state of information about Golomb rulers in a table [*see illustration on this page*]. From two to 24 marks there is both certain knowledge and some guesswork about the size of Golomb rulers. What I call the zone of perfection extends from two to four marks. Thereafter the zone of knowledge embraces the Golomb rulers having up to 13 marks. All the rulers here are known to be minimum. That is to say, in each case there is no shorter Golomb ruler that has the same number of marks. A Golomb ruler of five marks has length 11. A Golomb ruler of 13 marks has length 106.

Beyond 13 marks lies what I call the twilight zone. Dignified as the zone of research, it contains only rulers not yet known to be Golomb. For each number of marks there is a ruler that has the length given in the table. But shorter rulers may exist. Indeed, there is a formula that provides a lower limit for these lengths. A steadily widening gap between formula values and rulers so far found attests either to a weakness in the formula or to increasingly poor rulers.

Robertson is responsible for extending the knowledge zone to include 13-mark rulers. In a computer run that lasted for a month his program exhaustively searched through all potential Golomb rulers bearing 13 marks and found the shortest one. It would probably interest very few readers to search for Golomb rulers that have 14 or more marks if runs longer than this are needed.

Instead it seems reasonable to suggest some probing techniques, computational raids into the research zone that promise some return in the form of better rulers. Basic to any such effort is a program called CHECKER. CHECKER addresses the following task: Given an array of integers, what is the most efficient way to determine whether the differences between them are all unique? The simpleminded approach generates all possible pairs of integers and stores their differences in another array. Then it checks the file for duplicates relying on an awkward and time-consuming algorithm.

Rarely does the faster way to do a job require a shorter program, but here is a case. Since the differences themselves are supposed to be distinct, they can be used as addresses in a special array called *check*. Initially only 0's are stored in *check*. Each time a new difference is calculated the value stored at the appropriate address is changed from 0 to 1. Thus as CHECKER proceeds with its computations it may find a 1 already stored at a particular address, implying that the "new" difference has actually been seen before. In such a case the ruler cannot be Golomb, because it does not pass the fundamental test of Golombicity: each distance must be generated only once.

The technique of using differences as addresses constitutes a primitive form of what computer scientists and programmers call hashing. In many information-retrieval settings, hashing is the fastest way for a computer to recall a file.

	NUMBER OF MARKS	SHORTEST RULER KNOWN	LOWER BOUND
ZONE OF PERFECTION	2	1	
	3	3	
	4	6	
ZONE OF KNOWLEDGE	5	11	
	6	17	
	7	25	
	8	34	
	9	44	
	10	55	
	11	72	
	12	85	
	13	106	
ZONE OF RESEARCH	14	127	114
	15	155	133
	16	179	154
	17	199	177
	18	216	201
	19	246	227
	20	283	254
	21	333	283
	22	358	314
	23	372	346
	24	425	380

Lengths of Golomb and near-Golomb rulers

In more detail for those who require it, here is the essence of CHECKER. Two nested loops are used to generate all possible pairs of integers from the input array. If the first loop generates i and the second loop generates j, the program computes the absolute value of their difference and stores it in a variable called *diff*. In the next step CHECKER uses the value of *diff* as a kind of hash code: in algorithmic language one can write the following:

```
if check(diff) = 1
   then output "non-Golomb" and exit
   else check(diff) ← 1.
```

If the program never says "non-Golomb," the ruler has passed the main test. But how short is it? There are a number of ways to find out.

First, it is possible to use CHECKER in the stand-alone mode. I can imagine a reader hunched over the keyboard running only that program. He or she is exploring the research zone at an altitude of 14 marks, looking for a ruler shorter than 127 units, the best ruler currently known. The reader, flying in IFR weather, has no idea which way to turn. He has just submitted a sequence of 14 marks. The largest integer in the sequence is 124 and the excitement is almost too much as the display screen springs to life: "Congratulations. The set is OK." In programming this message he vowed never to try CHECKER on anything but potentially record-breaking sets.

Perhaps the reader found his record-breaking set by following Golomb's advice and exploring only those rulers in which the largest space appears in the middle. The spaces on such a ruler become smaller toward the ends of the ruler but they do so at the reader's discretion. Golomb assures us that many good rulers, if not necessarily the best ones, follow this pattern.

CHECKER can be modified to suit a more tentative style of inquiry. In STEP CHECKER the integers are typed in one at a time. After each entry the program generates the differences between the integer just entered and those already stored. In fact, STEP CHECKER is simply a version of CHECKER in which an input statement replaces the outer loop. The sequence is successful if the last integer has been digested and the program has not printed "non-Golomb."

Finally, the program STEP CHECKER can be incorporated into an automated search of the kind undertaken by Robertson. His program (which I may as well call EXHAUST because it is exhaustive) generates new rulers by adding one space at a time systematically. After each addition STEP CHECKER decides whether the ruler currently under construction is valid.

Robertson constructed his program by visualizing a ruler to which new spaces (and so new marks) are added left to right. Readers who followed this trail of prose through the byway of Golomb's argument (proving the nonexistence of perfect rulers) will remember that the spaces that must occur had the lengths 1, 2, ..., $m - 1$ in some order. Although this is true only of perfect rulers, something similar is true of Golomb rulers in general. Most but not all of the lengths from 1 to $m - 1$ between consecutive marks on less than perfect rulers occur in some order. Yet some spaces even longer than $m - 1$ can be found within such rulers.

Robertson generates new spaces in a stepwise manner. He maintains them in an array I shall call, appropriately, *spaces*. EXHAUST traverses the array adding one space after another. Naturally there are some simple tests that ease the labors of EXHAUST. One of these is to be sure that when a new space is generated it does not already occur in *spaces*. A second test is to sum all the spaces making up the current ruler to confirm that their sum does not exceed the shortest length known.

The EXHAUST program in operation seems eager to find rulers. It sets the first element of *spaces* to 1 and adds units to the second space so that it is different from the first. Then it adds units to the third space so that this distance not only is different from the first two distances but also satisfies the requirement set by STEP CHECKER, namely that all distances contained in the ruler must be different from one another. Each time a new entry of *spaces* is decided in this way, EXHAUST adds up the array and compares the sum with the length of the shortest ruler yet known. If the sum is less, the program continues to the next entry. If it is not less, EXHAUST returns to the preceding entry and continues to add spaces there.

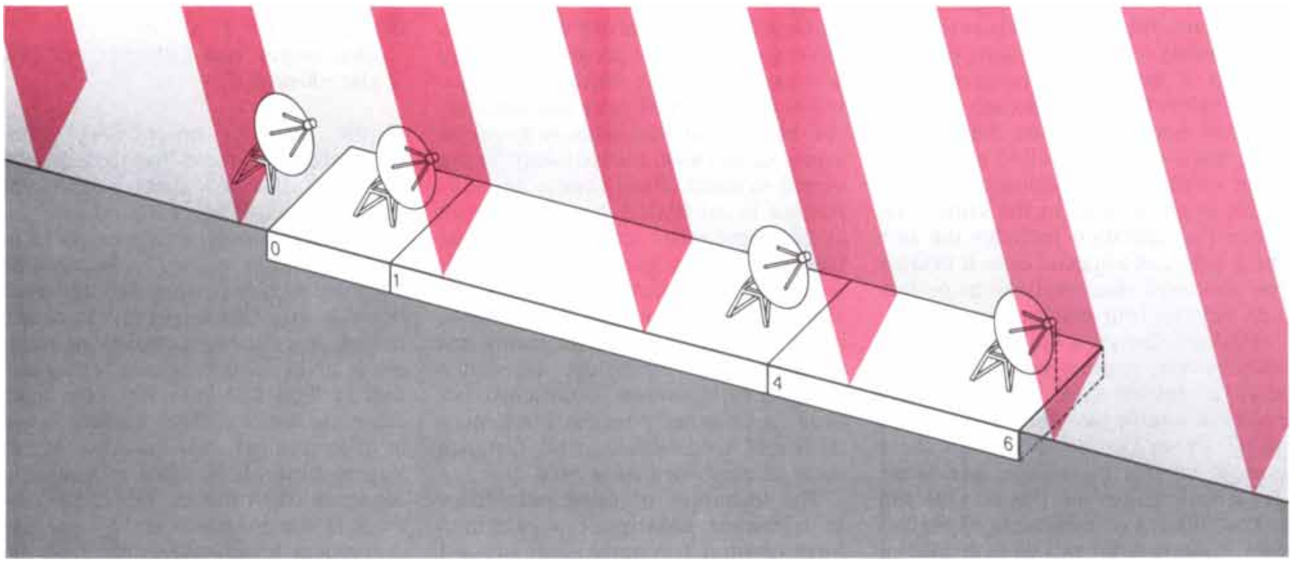
Robertson's program will run marginally faster if the first element of *spaces* is set to 2 instead of 1. Indeed, a one-day run will be shortened by a few hours. Readers may want to ponder why the search is still exhaustive.

Surely the effectiveness of an exhaustive search program depends on the inclusion of further tests and heuristics. Additional limitations on the values assumed by various entries in the *spaces* array would particularly enhance efficiency. Perhaps there is an incrementing procedure that uses much smaller ranges of such values. Processing the array is akin to counting. The count is reached much sooner if the number of possibilities at each entry of *spaces* is reduced. In any event, the readers who found busier beavers, new glider guns and other benefits to research will doubtless make their own way into the Golomb research zone.

New rulers should be sent to Golomb at the University of Southern California, University Park, Los Angeles, Calif. 90089. The most remarkable finds will be published in a future column.

Golomb has offered a prize of $100 to the first person who finds two different rulers that have more than six marks and yet measure the same set of distinct distances. Rulers that are mirror images of each other are not regarded as different.

A positive result would ring the death knell for a "theorem" propounded by Sophie Piccard, a Swiss mathematician, in 1939. Piccard's theorem states that two rulers measuring the same set of distinct distances must be the same rulers. The theorem was embraced by X-ray crystallographers because it helped them to resolve ambiguities in diffraction patterns. Unfortunately the theorem fails for numerous pairs of rulers that have six marks. Perhaps it is true for all rulers



Pairs of radio-telescope antennas set up on a Golomb ruler can reveal phase differences between incoming signals

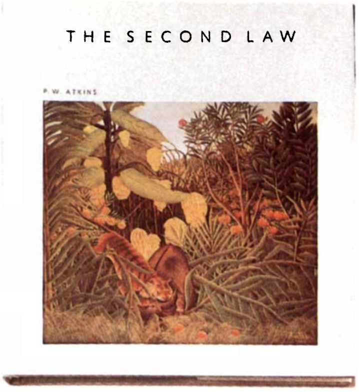
of higher order. Readers may pursue Golomb's prize without venturing into the research zone. It is a question that can be investigated for rulers bearing as few as seven marks.

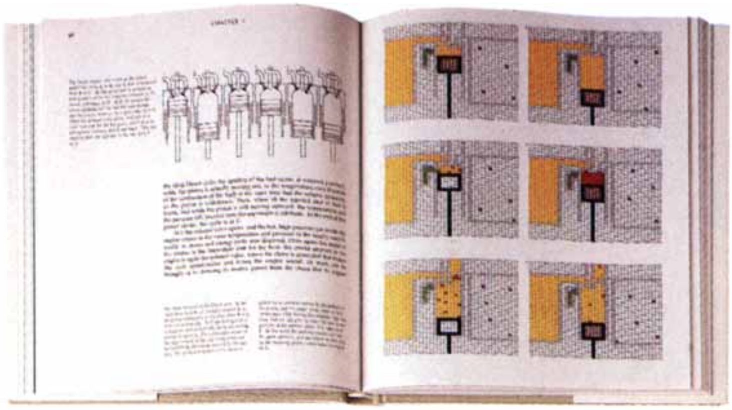
How does all this relate to helping Robertson? Radio astronomy makes occasional use of Golomb rulers in the resolution of distant radio sources and in the measurement of our own planet. In the first case a number of antennas are placed along a straight line several kilometers long. The antenna positions correspond to the marks on a Golomb ruler [*see illustration on page 22*]. To locate a distant radio source, it is essential to determine the angle between the antenna baseline and the direction of wave fronts arriving from the source. The antennas are all observing at a given wavelength. The precise time at which each wave in the incoming signal arrived at each antenna can be determined by analysis of the tape that captures the incoming signal. The total number of wavelengths between a given pair of antennas is called the total phase difference. It is normally composed of an integer and a fractional part called the phase difference. If the total phase difference can be reconstructed, the sought-for angle between the source and the baseline is easily calculated from the observing wavelength and *c*, the speed of light. Each pair of antennas, however, can only yield the phase difference itself, not the total phase difference.

In truth, it is Fourier analysis that recaptures the total phase difference from the many pairs of antenna recordings. But if the distance between one pair of antennas is the same or nearly the same as the distance between another pair, the two pairs provide the same phase-difference information. Redundancy of information means its loss. The accuracy of the source-angle computation is greatest if each antenna pair records a different phase difference; this condition is achieved by in effect placing the antennas on the marks of a Golomb ruler.

Another way to locate a distant radio source is to use just two receivers, each scanning an entire set of wavelengths simultaneously. Observing with two antennas a distant source at several different wavelengths yields the same information about the total phase difference between the antennas as would the use of several antennas tuned to the same wavelength. Here the same threat of redundancy looms again. No two pairs of scanning wavelengths should be the same distance apart, so to speak. The Golomb ruler is invisible but nonetheless present.

Robertson uses the second technique not to map radio sources but to





YOUR PREMIER VOLUME

THE SECOND LAW

P.W. Atkins

"The interested person can do no better than to peruse this fascinating new volume by Peter Atkins . . . . Even those who feel they already know all about the Second Law can hardly fail to find new angles and fresh insights into what is going on in the world around them. I hope that this remarkable book will be widely read by scientists as well as by the non-scientists for whom it is primarily intended."
NEW SCIENTIST

"A lovely book, beautifully illustrated and presented, and clearly commensurate with its companion volumes in the Scientific American Library. *It should engender affection as a sophisticated latterday Hogben, complete with social allusions implicit in the call to conserve, not energy, but entropy, and to use the virtuous entropy-conserving heat pump rather than crude and profligate combustion..."*
NATURE

In his famous essay on the two cultures, C.P. Snow made the Second Law of thermodynamics the litmus test of scientific literacy: Not to know the Second Law is the same as not having read a work of Shakespeare.

This is the law, of course, that explains why hot objects grow cool whereas cool objects do not spontaneously become hot; why a bouncing ball must come to rest and a resting ball cannot, of its own, begin to bounce. To some people, the Second Law conjures up visions of clanking steam engines, intricate mathematics and incomprehensible physical theory. It is better known to many in its restatement as Murphy's second law: "Things, if left alone, will gradually go from bad to worse!"

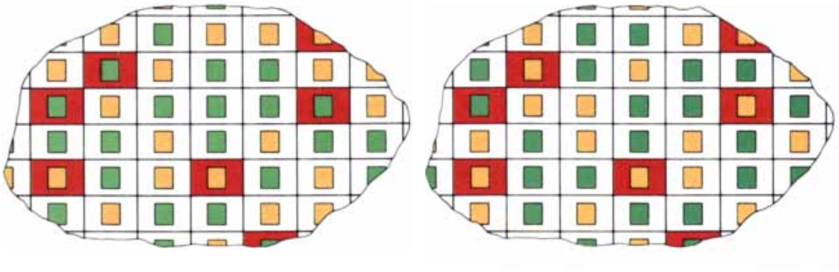
That restatement captures, in its way, the universality and power of the law that inspired Arthur Eddington to call it "the arrow of time."

In *The Second Law* P. W. Atkins, of Oxford University, breaks through the mathematical barriers to understanding of the law. By vivid example he shows it at work in heat engines, refrigerators, and heat pumps; the cosmos, in the ecosphere and in the living cell; in the physical, chemical, and life processes that drive the changing world in all its richness and variety.

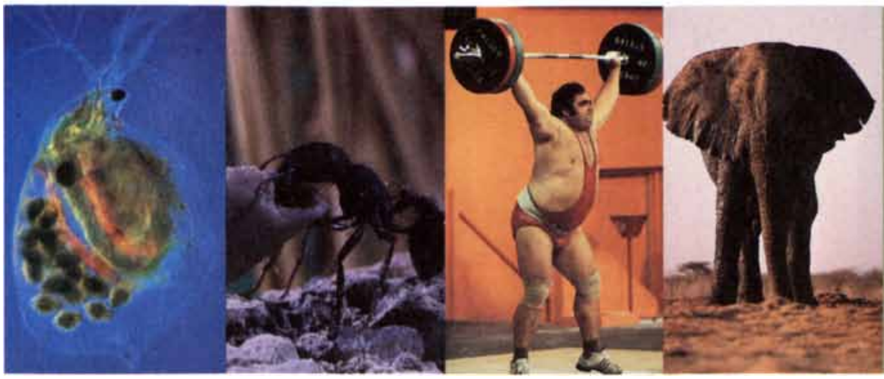
A special appendix equips readers who have computers with a kit of tools to put the Second Law to work.

230 pages, 104 illustrations

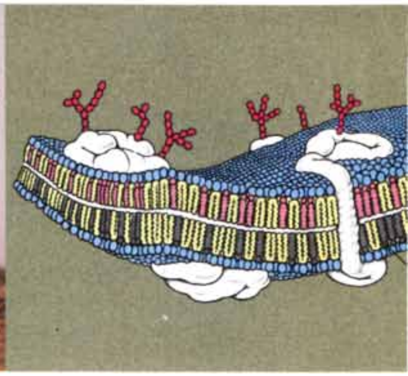
P. W. ATKINS is a lecturer in physical chemistry at the University of Oxford and a fellow of Lincoln College. He is the author of numerous distinguished works in his field, including *Physical Chemistry,* Second Edition, *The Creation,* and *Molecular Quantum Mechanics,* Second Edition.



The Mark II universe model of a paramagnetic material at some arbitrary temperature. In the absence of a magnetic field, the opportunity for electrons to be UP (yellow) or DOWN (green) is an additional contribution to the entropy of the sample. Atoms may still be ON (red) and OFF (white), but the spins they carry may be UP or DOWN at random. The two illustrations show the same distributions of ON and OFF, but different distributions of UP and DOWN.



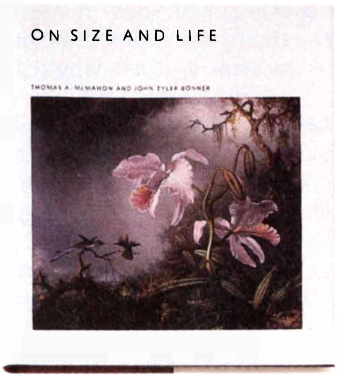
The largest forms of life reproduce slowly, move ponderously, and live much longer than smaller organisms. On Size and Life *examines why.*



A molecular view of the plasma membrane. Ele through an axopod of the heliozoon Echinospl The Living Cell.

ON SIZE AND LIFE

Thomas A. McMahon and John Tyler Bonner



The size of living things reaches across 22 orders of magnitude, from the 100 ton (100,000,000 grams) blue whale down to the most primitive bacterium at .0000000000001 gram. Such differences in size compel the differences in form, proportion and complexity that give us the near-infinite diversity of life on this planet. Upon close and systematic study, however, it turns out that "certain beautiful regularities in nature, describing a pattern in the comparison of animals (and plants) of different size," give order and plan to that diversity.

These regularities and that pattern explain why there are flying squirrels but no flying horses, why ants can lift 50 times their weight while humans must strain to lift just their own weight, why the smallest multi-celled organism is of the same size as the largest single-celled one.

In this book a biologist and an engineer have joined to give their readers command of the concepts that explain how it is that living things are the way they are.

THOMAS A. McMAHON is Gordon McKay Professor of Applied Mechanics at Harvard University. JOHN TYLER BONNER is professor of biology at Princeton University and a fellow of the American Academy of Arts and Sciences.

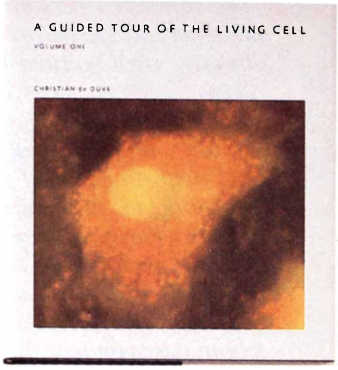
255 pages, 207 illustrations

"A remarkable book—both absorbing and accessible .... I know of no other book that introduces [dimensional analysis] so thoroughly or so well in a biological context .... It should give [readers] a much deeper understanding of why living things are the size they are."

TIMES HIGHER EDUCATION SUPPLEMENT

A GUIDED TOUR

Chris



with the exterior. In the second, cells own organs—and observe t turing their own special produc nucleus itself, where we see the d prints encoded in the genes to ste and to reproduce the cell itself.

CHRISTIAN DE DUVE is Andrew W. Me sity in New York. He was awarde jointly with Albert Claude and Ge the structural and functional orga

2 volumes, 422 pages, 425 illustration

"It is hard to imagine a better intr means to everyone."

JOHN BARKHAM REVIEWS

"A splendid introduction to cell b enjoyment."

NATURE

OTHER BOOKS IN THE SCIENTIFIC AMERICAN LIBRARY

THE DISCOVERY OF SUBATOMIC PARTICLES
Steven Weinberg

POWERS OF TEN
Philip and Phylis Morrison and the Office of Charles and Ray Eames

HUMAN DIVERSITY
Richard Lewontin

THE SOLAR SYS
Roman Smoluch

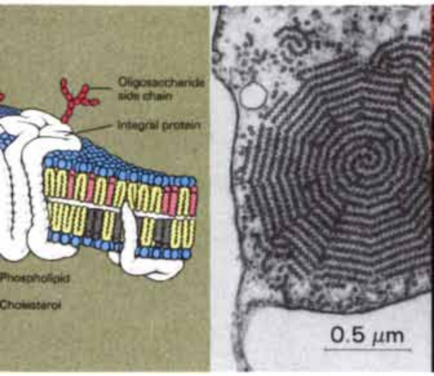
PERCEPTION
Irvin Rock

MATHEMATICS AND OPTIMAL FORM
Stefan Hildebrandt and Anthony J. Tromba

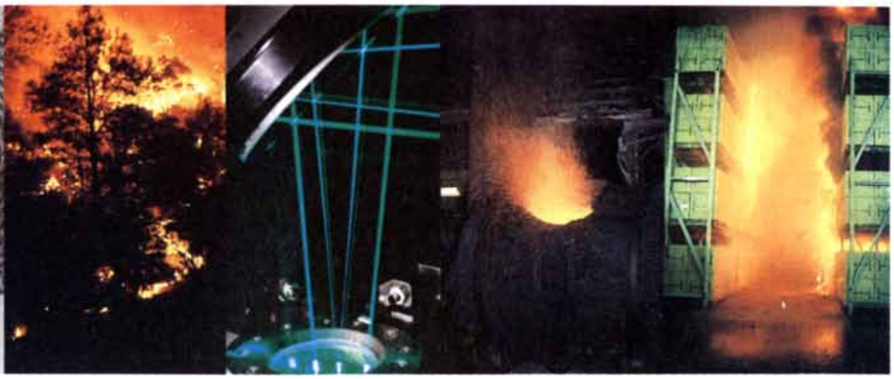
SUN AND EARTH
Herbert Friedman

EINSTEIN'S LEGACY
Julian Schwinger

GRAVITY
John Archibald W



micrograph of a section
nucleofilum. *From*



The "crowning" of a forest fire. Laser-beam spectroscopy. The copper converter. An experimental warehouse fire. From Fire.

THE LIVING CELL

de Duve

n *A Guided Tour of the Living Cell,* Christian de Duve invites us to accompany him on an engrossing ex-
lition through a world far removed
everyday experience yet central to
very being. It is the world of the
lions of cells that make up the human
ly.

he journey is divided into three itin-
ries. In the first, we examine a cell's
er and inner membranes, their intri-
foldings and two-way commerce
call at all the major organelles—the
transforming energy and manufac-
Our third itinerary takes us into the
helix of DNA reading out the blue-
biochemical machinery of the cell

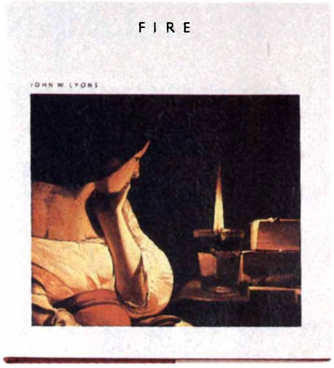
Professor at The Rockefeller Univer-
Nobel Prize in medicine in 1974,
e Palade, for discoveries concerning
ation of the cell.

ction to the world of the cell and all it

gy... should be read with profit and

FIRE

John W. Lyons



Fire contained in furnaces and combustion chambers serves each American with the electrical and mechanical energy equivalent of 100 human slaves. To unwanted fires, however, Americans lose more life and property each year than any other people in the world. Fires no longer burn down cities, it is true, and fires are confined nowadays to the dwellings in which most of them start. But anyone who has experienced the horror of a house on fire must find it hard to believe that there has been any progress in the prevention, control, and fighting of fire.

In *Fire,* John Lyons shares the lore and understanding he acquired in a career of study of the subject at the National Bureau of Standards. He develops both the friendly and unfriendly aspects of fire. Out of his own deep fascination with it he makes fire a compelling object of instruction in the physics and chemistry within. He shows the flame in its generality from the candle to the fireplace, the furnace, the combustion chamber and finally in the primal violence of fire out of control.

As director of the National Engineering Laboratory, JOHN W. LYONS is responsible for the National Bureau of Standards research in applied mathematics, building technology, electronics and electrical engineering, manufacturing engineering, and chemical engineering and fire research.

170 pages, 127 illustrations

PHOTO CREDITS—Daphnia: James Bell, Photo Researchers, Inc. / Ant: Daniel Adams / Elephant: Peter Johnson, National History Photographic Agency / Illustration drawn by Neil O. Hardy / Electron micrograph courtesy of Dr. L. E. Roth and Dr. Y. Shigenaka / Fire in trees: L. Burr, Gama-Liaison / Copper Smelter: John Moss, Photo Researchers, Inc. / Laser: Courtesy of Sandia Labs / Plastics Fire Test: Courtesy of Factory Mutual Research Corporation

THE SCIENCE OF MUSICAL SOUND
John R. Pierce

FOSSILS AND THE HISTORY OF LIFE
George Gaylord Simpson

CONSTRUCTING THE UNIVERSE
David Layzer

ENZYMES
David Dressler

EXTINCTION
Steven M. Stanley

DRUGS AND THE BRAIN
Solomon Snyder

locate the antennas themselves. For his purpose it is not enough to know that the second antenna is in Westford, Mass. He needs to know its position to within a few centimeters. The precision of such location is possible if a very distant, pointlike radio source such as a quasar is used. To locate a single point on the earth's surface with respect to a distant radio source is tantamount to the precise determination of such fundamental earthly variables as diameter, spin orientation and length of day. At the level of centimeters or microseconds such variables are truly that, posting annual, seasonal and even meteorological variations that are sometimes meaningful and at other times mysterious.

The September "Computer Recreations" column described CRABS, a benign terminal illness. Hordes of marine crustaceans occasionally and without warning descend hungrily on the Blit terminals of scientists at the AT&T Bell Laboratories in Murray Hill, N.J.

Almost every aspect of the Blit terminal's multiprogramming environment relies on an important instruction called bitblt. Bitblt transfers a rectangular set of bits from one area of the Blit terminal's memory to another. Two such rectangular sets can be combined by various logic operators before the transfer is effected.

Readers were challenged to solve two bitblt puzzles: erasing a picture from the screen and rotating a picture by 90 degrees. The first puzzle is easy to solve. To erase a picture occupying a given rectangular set, combine the set with itself using the XOR operator. The result consists of nothing but zeros. A single bitblt command replaces the original picture with a blank screen.

Rotating a picture is much harder than erasing it. So far only one reader has submitted a workable solution. Thomas Witelski, a high school student in New York City, has found a way to rotate an $n \times n$ picture on a $2n \times 2n$ screen. Using only $3n - 1$ bitblt operations, he slides $n - 1$ rows, copies an $n - 1 \times n - 1$ subpicture, slides $n - 1$ columns, copies another subpicture and then slides $n - 1$ more rows. Witelski's method is faster than the standard rotation algorithm, which requires $4n - 2$ operations.

There have been heartfelt cries from readers who do not feel competent to write their own version of the MANDELZOOM program but would like to run it on their personal computer. When readers write to ask for listings or disks of the programs I describe, there is a temptation to comply. But it would simply take too long to fill all the orders I get. Moreover, the time a reader spends learning the few elements of programming necessary for most of the projects I describe is time well spent. The teacher in me is pleased at the thought.

When other readers generously offer their own programs for distribution, however, I am tempted beyond my ability to resist. Following are the names and addresses of six individuals who are willing to supply programs under varying conditions and in varying states of accuracy and beauty. Caveat emptor:

Mark W. Bolme
Token Software
P.O. Box 3746
Bellevue, Wash. 98009
(IBM PC and Apple II family)

Pete Gwozdz
21865 Regnart Road
Cupertino, Calif. 95014
(IBM PC)

Will Jones
609 Rochester Avenue
Coquitlam, British Columbia
CANADA V3K 2V3
(IBM PC)

Bradley Dyck Kliewer
3001 East 24 Street
Minneapolis, Minn. 55406
(IBM PC with or without 8087 chip)

Charles Platt
9 Patchin Place
New York, N.Y. 10011
(IBM PC)

Richard A. Tilden
10 Thurston Street
Somerville, Mass. 02145
(Zenith-100)

Just as some readers have trouble getting started in the fine art of programming, so others, more expert, have an urge to share. It has occurred to me that a form of computer buddy system might be implemented with a minimum of administrative detail. It would link two classes of people: readers who expect to need help with the programs in "Computer Recreations" and readers who are willing to be of help. Am I wrong in supposing that the latter are in adequate supply? Send me a postcard with your name, full address, telephone number and category (tyro or adviser). Readers willing to advise should specify the maximum number of tyros they feel capable of helping. I shall endeavor to make matches within reasonable geographic distances.

Here are 152 reasons to buy at Elek-Tek, not to mention the fastest delivery anywhere.

Save 30% to 43%

off Manufacturer Suggested Retail prices on

EPSON • Okidata • Star • TOSHIBA • COMREX • OLYMPIA

IOMEGA BERNOULLI BOX (Modular)

1. 10 meg ½ height Drive for IBM-PC/XT/AT & compatibles 1675
2. 20 meg ½ height Drive for IBM-PC/XT/AT & compatibles 2335
3. Non-Bootable Interface Card ... 104
4. Bootable Interface Card ... 234
5. 10 meg cartridges for above (3 pak special) 125
6. Head Cleaning Kit 59

EPSON®

16. LX 80 .... $ 215
17. LX 90 220
18. RX 100+ 310
19. FX 85 340
20. FX 185 475
21. LQ 1500 parallel .... CALL
22. DX10 Daisy Wheel 10CPS .. 230
23. DX20 Daisy Wheel 20CPS ... CALL

TOSHIBA

46. Toshiba P341 .... CALL
47. Toshiba P351 .... CALL
48. Toshiba P1340 430

OKIDATA

49. Oki 182 .... CALL
50. Oki 192/193 .... CALL
51. Oki 84 .... CALL

star

52. SG 10 .... $ 235
53. SD 10/15 .... CALL
54. SR 10/15 .... CALL

HEWLETT-PACKARD

75. HP-71B Computer .... $ 399
76. HP-82400A Series 70 card reader 125
77. HP-82401A HP-IL interface 95
78. HP-82402A 4K memory module 60
79. HP-82700A 8K memory module 150
80. HP-82441A Assembly/Forth 120
81. HP-82480A HP-71 math pac 75
82. HP-82482A HP-71 finance pac 60
83. HP-82483A Surveying Pac 120
84. HP-82484A Curve Fitting 75
85. HP-82485A Text Editor 60
86. HP-82488A Data Com. Pac 120

HP-ThinkJet

HP-2225B $350

87. HP-41CV .... $ 168
88. HP-41CX 245
89. Optical Wand 95
90. Card Reader 145
91. Printer 283
92. Quad R.A.M. (for HP41C) .. 60
93. Ext. Memory Module 60
94. Ext. Function Module 60
95. Time Module 60
96. HP-IL Loop Module 95
97. Digital Cassette Drive 400
98. Printer/Plotter (HP-IL) 335
99. HP-9114A Disk Drive 600

SLIMLINE Shirtpocket Styled Power Packed Programmable LCD PROBLEM SOLVERS

100. HP-11C Scientific 58
101. HP-15C Scientific 90
102. HP-12C Financial 90
103. HP-16C Programmer 90

HEWLETT PACKARD — PERSONAL COMPUTATION DEALER

HARD DRIVES

Dysan CORPORATION

Internal Hard Disk Subsystems

7. 10 MEGABYTE 500
8. 20 MEGABYTE 600

Each system includes:

- Micro Science Drive
- Western Digital Controller
- 2 cables
- Documentation
- Dysan 1yr limited warranty
- For use in IBM PC, XT or most compatibles

PRINTER BUFFERS

9. DDP-16 64K Par./Par. 70

QUADRAM

10. QRMP8 8K Parallel/Parallel 125
11. QRMPS8 8K Parallel/Serial 139
12. QRMSP8 8K Serial/Parallel 139
13. QRMSS8 8K Serial/Serial 139

POWER PROTECTORS

14. PowerMite PG600-S9 6 outlet Surge Suppressor ... 30
15. ISOBAR 8 8 outlet Surge Suppressor ... 50

PRODUCTS FOR IBM-PC

24. Amdek 310A Amber Monitor .... $ 150
25. Generic Multi Multifunction Board, 64K 129
26. Generic Multi 384K Multifunction Board, 384K 163
27. AST Six Pak + Multifunction Board, 64K 225
28. AST Six Pak + (loaded) Multifunction Board, 384K 279
29. AST Megaplus II Multifunction Board, 64K 270
30. Quadram Quadboard Multifunction Board, 0K 195
31. Quadram Quadboard Multi. Board, 64K/384K .. 210/267
32. Orchid Tech. PC Turbo 186 570
33. Paradise Modular Graphics Card 290
34. Hercules Monochrome Card 329
35. Hercules Color Color Graphic Card 170
36. Novation 490605-1 2400BPS inc. Mite Software .. 620
37. Novation 490603 ½ Card Modem 2400BPS No software 425
38. Novation 490603-1 As above inc. MS-DOS Software . 490
39. Hayes 1200B Internal modem w/software ... 359
40. AT&T 4000 300/1200 Ext. Modem 335
41. Hayes 1200 External modem 380
42. Hayes 2400 External modem 599
43. US Robotics Courier 2400 Ext. 2400B Smart Modem ... 460
44. US Robotics Telpac Telecomm. Software 75
45. Toshiba ND 04D ½ ht. DSDD Disk Drive 90

IBM is a trademark of International Business Machines Corp.

PRODUCTS FOR APPLE

55. Toshiba HND 04DT ½ Height DSDD Disk Drive . $ 70
56. U—Print IIc-16K Par Interface external 85
57. Grappler PLUS Graphics Interface 80
58. Buffered Grappler 16K Buffer (Expand to 64K) .. 150
59. Kensington System Saver Fan and Surge Suppr. 69
60. Super Cooling Fan Fan and Surge Suppr. 29
61. Gamma Disk Controller Card 48
62. Gamma II SS Disk Drive for II, IIe, IIplus ... 125
63. Gamma IIc SS Disk Drive for IIc 125
64. Rana ELITE 2 DSSD Disk Drive 335
65. Rana ELITE 3 DSDD Disk Drive 420
66. ERAM 80 80 col; 64K RAM for IIe ... 115
67. Novation Apple Cat II 300 Baud Modem 200
68. Novation 212 300/1200 Baud Modem .. 390
69. Nov 49059112 Mod. & Soft. for MacIntosh . CALL
70. Hayes Micromodem IIe 300 Baud Modem 145
71. Hayes Smartmodem IIc 300 Baud Modem for IIc .. 240
72. Wico 501030 Analog Joystick 36
73. Amdek Monitors ..... CALL
74. Zenith Monitors ...... CALL

HUGE SAVINGS ON DISKETTES

	Dysan	maxell	3M	SONY	wabash	MEMOREX	BASF
3½″ SSDD	23.00	22.00	25.00	25.00	20.00	—	—
DSDD	29.00	29.00	33.00	37.00	24.00	—	—
5¼″ SSDD	17.00	14.00	14.00	13.00	11.50	11.50	11.00
DSDD	21.00	18.00	18.00	16.00	12.50	14.00	12.00
SSDD96TPI	36.00	33.00	35.00	—	24.00	—	24.00
DSDD96TPI	24.00	24.00	24.00	—	—	—	—
5¼″ DSDDHD (For IBM AT)	33.00	29.00	29.00	—	—	—	—
8″ SSDD**	22.00	29.00	25.00	—	19.00	—	—
8″ DSDD**	26.00	32.00	29.00	—	20.00	—	—

**Unformatted

Call for Quantity pricing for 10 boxes or more.

3M DATA CARTRIDGES

147. DC100 .... $ 14.00
148. DC300A 18.00
149. DC300XL 21.00
150. DC300XL/P 22.00
151. DC600A 23.50
152. DC1000 .... 15.00

Call for Quantity pricing for 10 cartridges or more

CALL TOLL FREE 800-621-1269 EXCEPT Illinois, Alaska **CANADIAN TOLL FREE 800-458-9133**

ELEK-TEK, inc. 6557 N. Lincoln Ave., Chicago, IL 60645 (312) 631-7800 (312) 677-7660

Corp. Accts. Invited. Min. Ord. $15.00. Visa or MasterCard by Mail or Phone. Mail Cashier's Check, Mon. Ord., Personal Check (2 wks. to clear) Add $4.00 1st item. (AK, HI, P.R., Canada add $10.00 first item) $1.00 ea. add'l shpg. & handl. Shipments to IL address add 7% tax. Prices subj. to change. WRITE for free catalog. RETURN POLICY: Defectives Only. Most products replaced within 30 days of purchase with identical merchandise only. Computer and large peripherals replaced only when defective on arrival (within 3 work days of delivery). Other problems covered by mfr. warranty. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE. Delivery subject to availability. DUNS #09-716-0517

On the road, an impending storm presents a special challenge—one the driver of a Thunderbird is well-prepared to accept.

Thunderbird's electronically fuel-injected engine provides the power. Steel-belted radial tires and rack and pinion steering provide the grip. And for further stability and road control, Thunderbird's shape helps reduce front and rear lift.

Inside, you'll find the environment of a true driver's car. Thunderbird's airflow management reduces wind noise and helps keep the windows clean. Seating areas provide lateral support for cornering. And to minimize the time your eyes are off the road, the instrument cluster provides vital information at a glance.

Of course, Thunderbird does have its limits; it can't predict the weather. It can, however, make dealing with a storm a little easier. You can buy or lease a Thunderbird at your nearby Ford Dealer. Have a nice day.

Best-built American cars.

"Quality is Job 1." A 1985 survey established that Ford makes the best-built American cars. This is based on an average of problems reported by owners in the prior six months on 1981-1984 models designed and built in the U.S.

Have you driven a Ford... lately?

Ford

BOOKS

A Christmas wreath of new books about science for younger readers

by Philip and Phylis Morrison

We present here the usual seasonal roundup of books about science and technology for children and young people. This year the list is long, but many of the issues are less than distinguished. Here is a sampling of the best of the crop.

Laboratory Sciences

ON BEING THE RIGHT SIZE AND OTHER ESSAYS, by J. B. S. Haldane. Edited by John Maynard Smith. Oxford University Press (paperbound, $5.95). It would be difficult to appraise this book as well as does its editor, the distinguished British evolutionist. He begins: "I first came across J. B. S. Haldane's essays when, as a schoolboy at Eton, I found that he was the person my schoolmasters most hated. Feeling that anyone they hated that much could not be all bad, I went to seek his books in the school library."

Professor Smith has compiled a wonderful collection of about 20 brief essays J. B. S. wrote during a period that began in the mid-1920's and ended in 1956. They come off as personal, as wryly prescient and original, as iconoclastic, as simple of phrase and as deep in meaning. The essays range over the broad spectrum of concerns that occupy contemporary scientists. The title essay, an account of the importance of scale to natural history, is as classic as the celebrated pieces by Galileo and D'Arcy Thompson. The closing essay offers reflections on nonviolence.

J. B. S., urging Indian biologists to nonviolence against animals, writes: "I am a man of violence by temperament and training... when in 1915 I was first under enemy shell fire, one of my first thoughts was 'how my father would enjoy this.'... I do not condemn those who do experiments on animals that involve their death, or even moderate suffering. But I have never done an experiment on an animal of a kind which I have not previously or subsequently done on myself; and I hope I never shall. I have dissected dead animals. But I have left instructions that my own body should be dissected by medical students. And if I die in India I hope some future Indian doctors will have the unusual experience of dissecting a European.... India has made many contributions to world culture. Perhaps the greatest is the ideal of nonviolence. Europe's greatest contribution is the scientific method. If these can be married, their offspring may raise mankind to a new level."

High school students who come on this honest and challenging book are lucky. If here and there it appears to be unfair to the occult, to religion, to physicians or to the rich, not authority but probable cause buttresses the author's judgment.

THE BIG STRETCH: THE COMPLETE BOOK OF THE AMAZING RUBBER BAND, by Ada and Frank Graham. Pictures by Richard Rosenblum. Alfred A. Knopf, Publisher ($9.95). A little experimental physics, a little history, a little technology and a good deal of humor are extracted here from a commonplace topic for curious youngsters in the grades.

There is even a clear procedure for making your own little rubber band out of backyard latex from dandelions or a milkweed plant. Natural enough, that idea is somehow surprising. You can check on the thermodynamics of fast expansion and contraction without any instruments, and a meter is described that you can make yourself to detect the unexpected contraction of a rubber band when it is heated. (The strangeness of the complementary fact that a stretched band cools when it contracts is not explained or even flagged; friction is no answer.)

Manufacture, testing, statistics and the trade are outlined. There are rubber bands specially made to handcuff lobsters, to preserve flower buds in air transit, to help straighten teeth. (Orthodontists call theirs ligatures; among us therapy tends to be in Latin—and expensive.) A modern rubber band is a slice from an extruded rubber tube passed through a hot vulcanizing bath of fused salts; it lives elastic and someday it dies brittle. "I've seen healthy rubber bands... ten or fifteen years old," a manufacturer says; such a venerable band is perhaps older than the intended reader. A famed rubber-band band is recalled, and rubber-band musical instruments, rubber-power boats and a traditional spool "tractor" are sketched for amateurs to build. Remember the Wright brothers, whose first flier was a rubber-band-powered toy from France. There are even a few rubber-band riddles.

It is a delight to see so much relevant matter drawn out of a modest topic. Perhaps this winning book with all its fun and sense will inspire others.

THE MIRROR PUZZLE BOOK, by Marion Walter. Tarquin Publications. Distributed in the U.S. by Parkwest Publications, P.O. Box A-10, Cathedral Station, New York, N.Y. 10025 (paperbound, $4.95). Puzzles have their devotees, young and old. This little book is wider in its appeal than most papery puzzles. For here the solution is sought with a small mirror, placed on a drawing of a geometrized fish so cleverly that the image you see matches one of a dozen test figures on the page. All of them appear in some way to have descended from the same green-and-purple fish and its twins. The book offers a dozen master drawings that resemble the fish, each flanked by its own dozen test figures. The work is neither as easy as it appears nor severely taxing. The sensory pleasures of the task set it apart from the pencil-and-paper puzzle.

Note that a good number of these plausible matches are in fact impossible. Soon enough you will begin to form a generalized rule for the conditions that allow the match. A summary chart of impossibilities is provided, set out in the form of another puzzle.

Symmetry is of course the name of this game. The little book can be thought of as mathematics, but these days it is surely to be counted also as a lower-grade harbinger of theoretical physics, mirrors and all. To help ensure against loss, a pair of small mirrors are included. Deft use of both mirrors at once opens the world of the kaleidoscope.

Technology

DRAWING WITH COMPUTERS, by Mark Wilson. Perigee Books, the Putnam Publishing Group (paperbound, $9.95). In remarkably little space Mark Wilson, an artist concerned with two-dimensional static works of art and seized with "computer madness," has made a fascinating


This Christmas,
take special care of your fine feathered friends.
This year, let your gift soar to new heights.
Give the Scotch that commands respect:
Johnnie Walker Black Label.
It has every right to be expensive.
Johnnie Walker
Black Label Scotch
YEARS 12 OLD

Here's one supermini that's

vigorous hybrid. On one side he gives an introductory account of the nature, limits and hopes for computer art. On the other he provides a workable example of the generation of such art in concrete detail, for anyone able to follow BASIC listings of modest length.

The general review begins with the synthetic realism of the high technology of Lucasfilm, Ltd. But in a couple of pages it has looked away from those corporate marvels toward abstract drawings open to almost anyone with a microprocessor-based computer who is able to command a plotter or a video screen. There follows an account of the various output means and a distant look at software in the domain of the personal computer.

Then we program a computer so that it renders 100 outlined squares. Transformations follow. By simple changes the squares can dance; they overlap, abut neatly, withdraw to fastidious separateness and transform drunkenly into odd shapes. Random patches of squares from the array acquire tones of gray and appear in 3D by forming perspectivelike distortions. They superpose on one another many layers to build apparent solids, and even gain coats of many colors. It is this bewildering productivity of visual variety that is now within the power of the personal computer and so within the grasp of a beginner. To date computers act more like kaleidoscopes than like artists: what they show an artist, that flood of naive image, is excitement and stimulation never before encountered. Art is still to come.

For good sense and a delightful start, this book is one of a kind. Young computer adepts will find much in it: rich visual enjoyment and perhaps the path no one quite knows toward the new art we hope for one day.

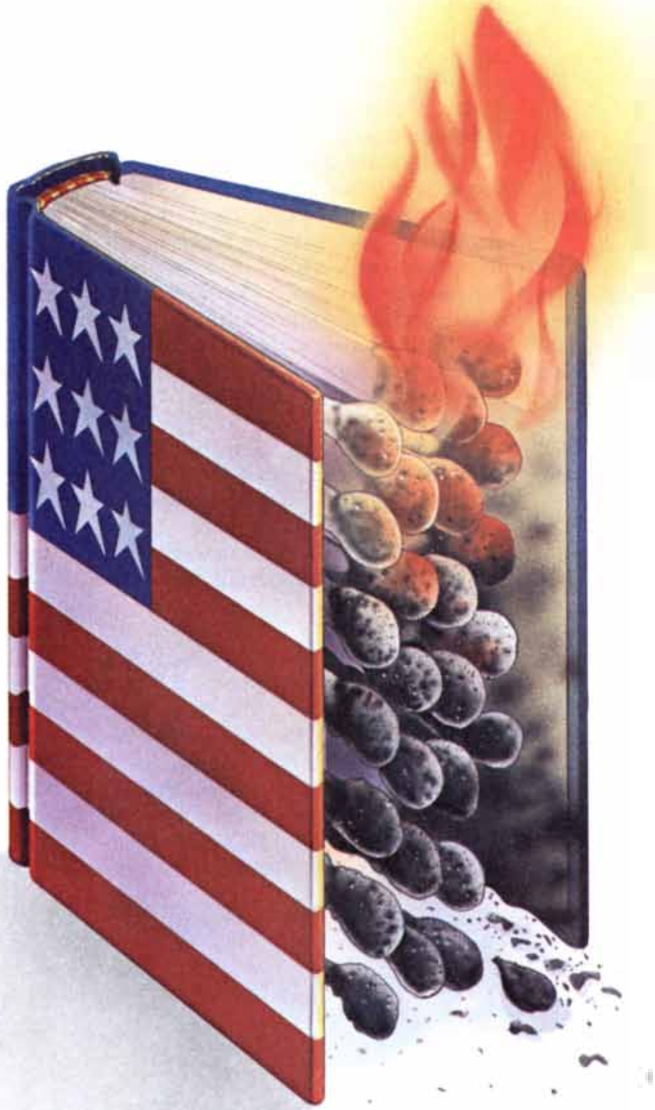
WORKS OF MAN, by Ronald W. Clark. Viking Penguin Inc. ($24.95; $29.95 after December 31). Nearly 100 color photographs (others too) brighten this well-written history of engineering. The author, an experienced popular biographer of scientific worthies from Einstein to Darwin to Freud, has not attempted a comprehensive book to compete with the scholarly tomes in many volumes that stand imposingly on library shelves. His is a more personal effort; its scope and its lively, readable details reflect his own interests. The core is the story of civil, mechanical and electrical engineering, old and new. Soda, mauve, celluloid and nylon do appear in the tale, but glass, the magnetic compass, McCormick's reaper, fixed nitrogen, photography, the linotype and television are not treated. Nuclear reactors, digital computers and space flight stand for the world today.

A useful outline of the book is provided by its illustrations, some fresh and some virtually obligatory. We see the Great Wall of China, Hadrian's Wall, Machu Picchu and a fine view of that classical iron-arch bridge, the work of the Darbys of Coalbrookdale, with an assist from "Iron-mad Jack" Wilkinson (who sought to be buried in an iron coffin). We see the canals grow, from the Regent's in London to Suez and Panama, and tunnels too. Less expectedly, there is a painting of the Battle of Lepanto fought in Greek coastal waters in 1571, the Turkish galleys losing to the bigger ones of Don John. Here is Isambard Kingdom Brunel natty in his plug hat, as his *Great Eastern* is launched. A good photograph and a fine few text pages treat that late-Victorian masterpiece, the grand set of tubular iron cantilevers that still carries the railway across the Firth of Forth.

Electricity enters at smaller scale: there are a Faraday dynamo and an Edison lamp as well as a close-up of a memory board (those are Nippon Electric Company chips made in Ireland). The big Hoover power dam appears in a tricky shot taken looking down its steep curve. The tallest freestanding tower, that of the C.N. at Toronto, is much more impressive in a night view than the factorylike protoype fast reactor at Dounreay in Scotland. Flight through the air is well summed up, from the first manned flight of a hot-air balloon, past Otto Lilienthal in his hang glider, all the way to a Sea Harrier. In space the images progress from a V-2 rocket on its pad to the untethered space walk of astronaut Bruce McCandless in 1984.

Ronald Clark's enthusiasms are clear; he is an optimist and an admirer. His anecdotal material is good and his facts are quite reliable. Even though a few myths enter they are famous and delightful ones. Take the story of the

The iron bridge at Coalbrookdale in England, from **Works of Man**



"sailor from San Remo" who dared to shout from the crowd for water to be poured on the stretching ropes, even as all stood silent in St. Peter's Square under the threat of hanging, while the trumpet paced Domenico Fontana's big crew as they heaved to raise the obelisk. If it is not true, it is surely artfully told. The entire book is a palpable hit for hopeful high school engineers with a bent for history.

THE PLANE. Illustrations by John Bradley. Design and Paper Engineering by Ray Marshall. Idea and text by Sadie Fields Productions. Viking Penguin Inc. ($13.95). Open the cover of the thin book and up toward your nose pops a four-inch pasteboard model of the flight deck of a Boeing 747, walls and panels studded with instruments named at the side, pilot and copilot wheels unmanned and ready, the dark airport edge outside. The small book presents two more pop-ups, one a generic plane labeled at aileron, elevator and so on, the other a retractable undercarriage strut with its clever nesting wheel. There are plenty of movable tabs to come, one set keyed for yaw, pitch and roll of little moving planes, another tab that pushes cutout planes along for landing and takeoff, one more that rolls out the flaps from a wing section. A final tab draws the air through a cutaway of a three-spool Rolls turbofan engine (this is a British book) and animates a convincing diagram of the flow and of how it is fed the flaming energy of combustion. Cabin pressurizing is not mentioned, although it is a familiar experience.

Pop-ups are always attractive and ingenious; these carry enough of the useful task of making clear the basic spatial geometry of flight to justify their presence in a book good for young passengers, airport visitors, even dreamers about jet travel. There is a closing page picturing airport terminals and their swarm of vehicles, and even a six-inch cutout model of the big jet, glueless but intricate, to assemble. The book would give any curious boy or girl aloft something to do over a couple of thousand miles of flight, certainly from JFK to O'Hare, perhaps all the way to LAX or Heathrow (particularly given a lucky escorted visit to the flight deck itself).

FILL IT UP! ALL ABOUT SERVICE STATIONS, by Gail Gibbons. Thomas Y. Crowell ($9.95). The tow truck comes back from the first call of the day, dragging in a little yellow car. It overheated, and sure enough the mechanic finds that it needs a new fan belt: easy. At John and Peggy's full-service station the day crew is busy. Two knowing mechanics are on duty, the owners are in all day, three young people pump gas, give directions, sell tires and point out the vending machines. We see the action underground too; the big gas storage tanks are there to be filled from the tank truck, and the hydraulic-lift mechanism dwells below. Two pages show just how a flat can be fixed when you know what to do and have the tools.

Evening comes and the mechanics go home. Repairs will have to wait. After midnight only one attendant stays on; he greets the passing police officer cheerfully. Another busy day comes with morning. The final page amounts to a catalogue of drawings of service-station tools. Here is a bright slice of American suburban working life served with a light heart for imaginative children in the early grades.

Natural History

THE MYSTERY OF THE BOG FOREST, by Lorus J. Milne and Margery Milne. Photographs by the authors and by Fred Bavendam. Dodd, Mead & Company ($9.95). The central eye of the bog looks like a shallow froggy pond, where harmless midges—but, oddly, few mosquitos—hum and open waters reflect the blue sky. But there is no muddy bank, no rocky shore, no stream, in or out. Instead the still pond is bordered everywhere by a mat of floating moss too thin to bear the weight of even the youngest explorer. Just outside that is a sturdier moss mat decorated with unusual flowers and low shrubs, where a light foot may penetrate, although hardly stay dry. Beyond that ring is a zone of shoulder-high dwarf spruce, larch (with good luck some highbush blueberry), the moss floor underfoot quaking as you walk gingerly among the woody plants. Only at the outermost edge of the wetland do you come at last to the mature forest.

It is easy to lose your way within the flat, watery world. "Bring along a yard or so of bright red ribbon, and, as you enter the bog, tie the ribbon to a tree at eye level." It is even easier—although much less scary—to lose yourself for a while in this fascinating book about one widespread class of wetlands and the special web of life they house.

The living basis of this small world is the peat moss, a term that embraces some species of *Sphagnum*. The moss turns the soaking-wet and soilless bog into an astringent little desert: what grows there can only be those species of animals or plants (100 are listed) that husband fresh rainwater, make do with scant nutrition and can tolerate groundwaters turned acid by the slowly rotting moss (no mosquito larvae). The topographic basis of the bog is a pool of water virtually free of circulation. Barely measurable drainage arises chiefly from glacial processes, as on the utterly flat silt of meltwater lake bottoms (Minnesota) or in gravel-banked depressions left by chunks of melting ice (Maine). North of the Carolinas there is no spot in the eastern U.S. or Canada more than 50 miles "from a growing or a dead peat bog."

Not many people live or work there, although surface mining of old peat bogs for fuel is an important industry across northern Europe. Here we drained nearly half of our eastern boglands so that their surface could be plowed. There they delved more into old mossy depths for peat. Perhaps that helps to explain why more than 700 human bodies, well embalmed by the tannic waters, have been recovered from European bogs, whereas only a few have been found here.

Any able grade school reader would enjoy this small book. A map representing and locating some particular bog might have added a bit more concreteness, but the work is a model of eyewitness natural history.

THE DAYWATCHERS, written and illustrated by Peter Parnall. Macmillan Publishing Company ($16.95). "Children dream as lions and eagles. I still do." So writes this amateur naturalist and distinguished author-illustrator. The experiences he recounts in this handsome book through word and image were not dreams. Rather they are dreams realized, dreams of birds of prey seen in their power and beauty, dreams that turned real for him once along the Maine coast, in the Colorado plains or even in Manhattan. He has recorded them all in lively and personal essays and in vigorous and free ink drawings, often at full-page size, to convey his feelings about a dozen species of birds and the boy or man who watched for them.

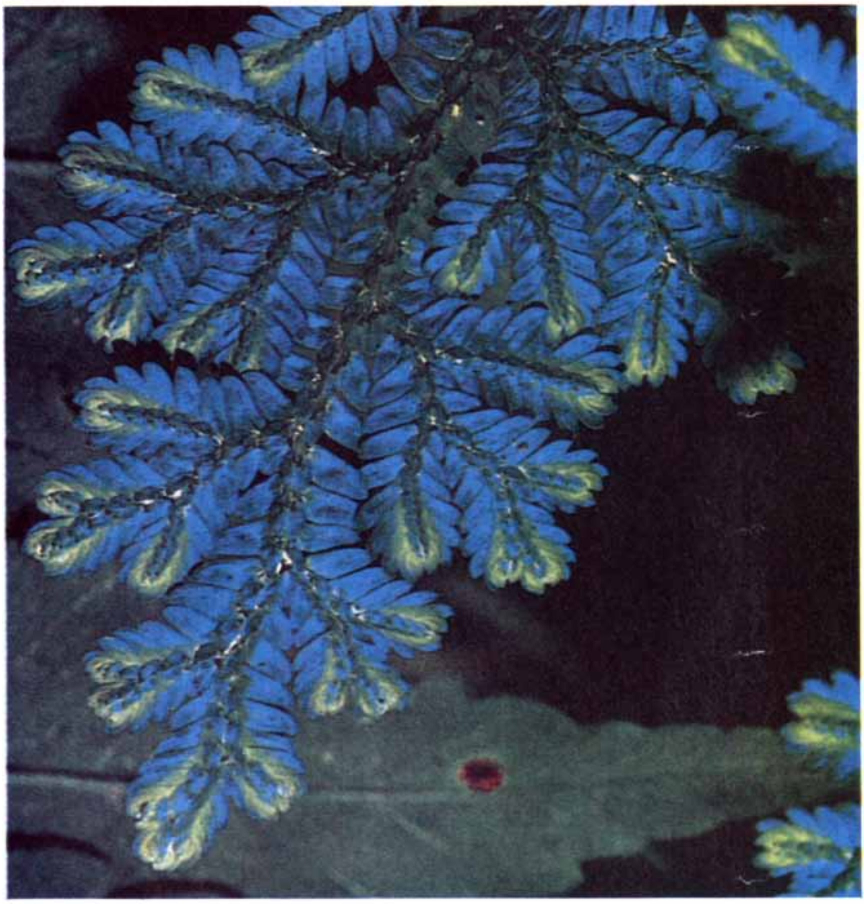
One crisp Sunday morning years ago he and his Irish wolfhound were jogging in Central Park behind the Metropolitan Museum of Art when "two dark missiles shot into the flock" of wheeling, banking pigeons overhead. They were a pair of peregrine falcons, and each took its prey. Peregrines feed on plump birds and dwell on high ledges, safe, lonely spots with a long view. Both the natural cliffs of the Hudson and the palisades midtown have sheltered their peregrines.

In a few pages Peter Parnall reviews the recent restocking of high Manhattan roof ledges (now sites in a dozen other cities as well) with a few of these swiftest of falcons, bred at Cornell since the early 1970's by the biologist-falconer Heinz Meng and his colleagues. The birds flourish now among the well-fed flocks of city parks, the pigeons themselves of course formerly birds of natural cliffs. The falcons replace the wild peregrines almost all lost in the 1950's, when DDT residues slowly contaminated their food and rendered eggshells too fragile to hatch chicks. Eagles, osprey, goshawk ... a dozen hawk species are recalled in the same spirit as part of a fulfilled life.

Mr. Parnall has never seen a gyrfalcon in the wild. He spends a few pages on the bird all the same, in memory of a schoolboy's dreaming and "because its elegance demanded it. I've not seen them, but I do feel them." There are plenty of facts about hawks in the book. Even though it is no field guide, it should engage many a young birder at quite a different level of interest.

LEAVES: THE FORMATION, CHARACTERISTICS, AND USES OF HUNDREDS OF LEAVES FOUND IN ALL PARTS OF THE WORLD. Photographs by Kjell B. Sandved, text by Ghillean Tolmie Prance. Crown Publishers, Inc. ($35). Leaves "color our world green." So remarks the peripatetic photographer from the National Museum of Natural History, and some 300 of his fine pictures of leaves color the vivid pages of this encyclopedic visual anthology. Why should flowers have so many celebrations in picture books and leaves none at all? Are we readers as blossom-drunk as the insects or can we grasp the importance of leaves? The point is argued here in interesting detail in the brief informed text by Dr. Prance of the New York Botanical Garden.

The varieties of leaf shape? There is a one-inch sedge leaf from Thailand spread across a page at eight times natural size. It looks like a wind-turbine wheel bearing a dozen narrow spiral blades. Some leaves float? Half a dozen shots show the giant platters of the royal water lily of the Amazon, whose young rolled-up leaves unfold into mature back-ribbed structures sometimes eight feet across. Prance is shown here in a study of that remarkable plant made by wading in the rivers where it grows. The color of leaves? Observe the strange little rain-forest plant, a fern ally, *Selaginella,* its leaves a glowing peacock blue, thin-film interference filters giving the leaf the ability to absorb the desirable red component remaining in the filtered light of the forest floor. Two fossil leaves are remarkable: a gingko print in rock unmistakably matches the living leaf, and another fossil leaf from Oregon, 30 million years old and still autumn red, peels freely from its rock matrix. The sight of leaves in mass is not omitted. A glorious scene from fall in Vermont closes the section in a crimson glow.



An iridescent leaf of the fern ally Selaginella, *from* Leaves

A book that serves as a door to meaning and offers a world of botanical examples, this fresh compendium is in no way intended for children. Yet it is certainly a volume many of them will enjoy consulting frequently. The book will attract even young readers, once they are able to work out the short but somewhat technical text.

SPRING DAYS; SUMMER DAYS; AUTUMN DAYS; WINTER DAYS, by Patricia Casey. Philomel Books, a division of the Putnam Publishing Group (four miniature boxed books, each $4.95). Take autumn: a delicately painted scene runs across about 20 accordion-folded tiny pages without a word. Leaves and berries are full of color; the apples are ripe. Eight or 10 of the pages bear a lift-up tab. Behind the red barn door one tab reveals a speckled white cow waiting to be milked, and a sentence to the point appears on the reverse of the tab in firm type. Lifting the tab that shows a dilapidated shed off in the corner of the garden even changes the point of view, to disclose (and describe in a sentence) a scary close-up of the hidden fox, hungrily watching the geese.

All four seasons follow the same pleasant pattern. The books are about two and a half inches square; there is a printed title on each cover and on its matching box. The paintings are full of charm, and the game with the tabs accents the flow of the countryside scenes with repeated small surprises. These should draw in very young readers indeed, those who are still getting ready for books.

LARGE AS LIFE: NIGHTTIME ANIMALS LIFE SIZE, paintings by Kenneth Lilly, text by Joanna Cole. Alfred A. Knopf ($9.95). Across the big double spreads run two or three sentences of simple text that identify the animal brightly painted in loving textural detail at closeup range, literally life size. They are not commonplace creatures: here are 10 forms, including a fennec fox, all ears; a family of European hedgehogs, appealing in their jaunty spininess; and a warty two-pound giant toad of the Caribbean, a formidable hunter of mice, birds and even snakes, itself safely wrapped in a stinging poisonous skin. The big toad contrasts with another page that shows a tiny

The tiny lesser mouse lemur is depicted at life size in **Large as Life**

royal antelope, a rabbit-size hoofed leaper, known in West Africa as "king of the hares." A fuller paragraph on each animal is tucked into a final section. It is a lesson in scale, a kind of paper menagerie of little animals amidst big flowers, and a bargain for both the read-to set and better readers. There is a matching book of daytime animals as well, although toad and hedgehogs here were winners (not to mention a quizzical pair of elf owls nesting in a saguaro).

Space

THE THIRD EXPERIMENT: IS THERE LIFE ON MARS?, by David E. Fisher. Atheneum ($12.95). COSMIC QUEST: SEARCHING FOR INTELLIGENT LIFE AMONG THE STARS, by Margaret Poynter and Michael J. Klein. Atheneum ($10.95). Professor Fisher is a Miami geochemist who commands a chatty writing style, full of rhetorical questions and the clichés of the day. One chapter ends with "Sorry about that"; schoolroom quips occur, even the expressive if stereotyped expletive "Yuck!" In that everyday form he has made for kids in the higher grades (and for adults who tolerate their jargon) an exceptional review of the wonderful question that is his subtitle.

Fisher's short chapters begin with the Mars of the imagination and later present the Mariner images. He outlines the strongest ideas we hold about life, including optical activity, proteins, cells and evolution, and he surveys the long contest among the life-seeking experiments for a place in Viking's costly one-foot box. The climax is a fascinating logical analysis of what Viking found. "The worst thing any ... could envisage was that they would go to Mars ... and still wouldn't be able to say for sure whether there was life up there or whether the planet was barren and dead. So guess what happened? Right."

Three complementary biochemical experiments all showed more or less lifelike processes. Yet for good reasons most experimenters agree that the copious reactions they found were not biological at all but some "weird chemistry" of long-irradiated iron red soils rich with reactive superoxides, stable only on bone-dry Mars. Such dirt would greedily destroy all trace organic gases, whose measured scarcity is the most unexpected feature of the atmosphere of Mars.

The third experimental team (Gilbert Levin and Patricia Straat) remain unconvinced. In an Antarctic cold dry valley, closest of all earthly locales to Martian conditions, they found in 1981 both the active label release and an absence of trace organics, a match to their own result on Mars. They see in the data "a biological response." There is a fourth experiment too, a visual one; all the data are here (in color on the jacket). A small green patch on one red rock moves and shrinks an inch or two between two Lander close-ups taken half a Martian year apart. Some wondrous Mars lichen enduring on water vapor? Or windblown dust?

Viking got an answer that raised new questions, which is what we ought to have expected all along. The fifth experiment? It is costly but clear: "We have to go back to Mars."

In simple but less trendy language, an experienced writer and a working radio astronomer (SETI manager at JPL) join to explain an even more tentative exploration in science. In this clearly and briefly outlined story the focus is on galaxies, radiation and star stuff. The relevant scale is explained and a general account is presented of the chemistry of life. A fair-minded but sensible approach is taken to the pseudowonders of UFO's, Nazca figures and the like. The ideas of field of view and frequency band are nicely clarified by marked photographs and apt chart recordings. Radio telescopes are given reality, not simply by familiar photographs of the big dishes but by presentation of the feed, the maser, equipment racks and the display screen. The arguments are clearly and personally outlined. This chain of logic has not been drawn taut as it has with Viking, but the exciting search is set out in a friendly and reasoned narrative.

THE RETURN OF THE COMET: AN ACTIVITY BOOK FOR SKYWATCHERS FROM 9 TO 14 WITH ADULT TEACHING GUIDE, by Dennis Schatz. Illustrations by Yasu Osawa. Pacific Science

Center, 200 Second Avenue North, Seattle, Wash. 98109 (paperbound, $8.95, postpaid). COMET HALLEY: ONCE IN A LIFETIME, by Mark Littmann and Donald K. Yeomans. American Chemical Society, 1155 16th Street N.W., Washington, D.C. 20036 ($19.95; paperbound, $12.95). From apt title and descriptive subtitle to the three-foot orbital map dotted with past events and future conjectures (Viking reached Mars in 1976; commercial moon flights begin in 2010—maybe) and empty boxes for filling in events of your very own, the enthusiastic authors have prepared an agreeable menu for comet-caught youngsters who want more than just reading about our icy visitor. Very properly stressing not the doubtful spectacle but the delight of its time-binding return, the book arises from the active milieu of the hands-on museum. A model, a flip-book of motions, comet picture and period matching, a greeting card to alert some friend by mail and of course hints and maps concerning the look of the sky during the next few months occupy most of its pages. Given a few nimble-fingered young people to share the tasks, families and friendly neighbors will find here a vacation resource. A faint shakiness of historical detail is entirely tolerable.

Mark Littmann and Donald Yeomans have produced a remarkably compact and yet comprehensive text on P/Halley and its kind. Their many illustrations in both color and black and white include the obligatory Giotto, the starry Comet Egg of 1680, the cover of the Comet Rag of 1910, spectra and the Siberian forest. The writing is light but expert, with footnotes to make it plain that Bayeux is really a tapestry and that Halley signed himself Edmond. Inserts on biography and on astronomical ideas further serve younger readers, and lists of facts (the features of this comet and its predicted returns, cometary meteor showers, split comets, spacecraft on the way) make reference easy for anyone. Theories are clearly presented and the presentation is broad. It spans the subject from the origin of comets to the death of the dinosaurs writ in iridium. Richer, thicker Halley books of style and quality are out there on the shelves, but this one, a real bargain in paper, should be high on any list for clarity in small compass. Littmann is an astronomical writer of long experience from planetarium to Space Telescope Institute. His partner Yeomans, of the Jet Propulsion Laboratory in Pasadena, is the current computer custodian of the perturbed dynamics of this comet; he is also a bibliophile and historian by avocation.

People

THE MYSTERY OF THE ANCIENT MAYA, by Carolyn Meyer and Charles Gallenkamp. Atheneum ($11.95). Pacal, Lord of old Palenque, looks at you, it would seem. What we see is the magnificent jade mosaic portrait mask, eyes of shell and obsidian, buried 12 centuries before Alberto Ruz (prompted by an insightful hunch more than 30 years ago) excavated the pyramid tomb. And that is not the only thrilling image in this fine introduction to the Maya. Here are Frederick Catherwood's wonderful pioneer images, the stones overgrown, Tatiana Proskouriakoff's spacious reconstruction, and a neatly drawn stele, its date glyphs translated into English in the adjoining column. The authors have written a brief book that tells the story of the ancient Maya and of how we have come to know this culture: we hear of Bishop Diego de Landa and his tales of the Well of Sacrifice and learn about the current enigma of the Classic collapse.

The thatched huts of Yucatán are drawn as they are today, identical in form with those whose images were carved on temple facades. There are two million people who now live in Mayaland, their culture continuous with (albeit of course much changed from) the cruel and glorious past under priests and nobles. The Maya farmer may eat up to 20 tortillas at a sitting; the housewife, who has made them all with patience and skill, dines later. Nowadays a metal press will save her much hand labor. No doubt about it, this fine account exposes its young readers to the contagion of *Mayismo,* an enthusiasm that has captured many another. It lacks a reading list, but it does have a good glossary and index.

FABLED CITIES, PRINCES & JINN, FROM ARAB MYTHS AND LEGENDS. Text by Khairat Al-Saleh, illustrations in color by Rashad N. Salim, line drawings by Peter Dennis. Schocken Books ($15.95). The huge ifrit dominates the cave. Buried to his armpits until the Day of Judgment, the three-eyed creature, enjoined against working any harm to men, has two great wings and four arms; two of them are human, two are lion paws. Through the opening of the cave we look out with the awed travelers to the distant minarets of the City of Brass, silent in its opulence of charmed death.

That painting in deep colors and dizzying perspective illustrates a story in *The Thousand and One Nights.* But this volume of striking paintings and brief text contains only a few tales from the famous treasury. Its stories come from the Arab world old and new, both the pre-Islamic time of isolation and the Golden Age. You will read of black-and-white serpents in fierce combat, of a man whose years were measured by the life spans of seven falcons, of trees of emerald and hyacinth. Arabia Felix echoes here, not only the life in old irrigated cities of the Incense Road but also the lore of the nomads of the austere desert.

Cities and princes we all know somewhat, angels and jinn much less. The angels are ordered in their holy ranks to serve God. Jinn are but mortal creatures like ourselves, fallen long before Adam, losers to the angelic host. They are long-lived; some are believers, beautiful, good, whereas others are infidels, evil, ugly. Those ifrits are both malicious and powerful.

The magical fantasies, the powerful images and the richness of background matter make this an introductory mythology of unusual authenticity and interest for imaginative readers. Reading skills will help; the writing is clear but not simple. The list of sources is impressive, although the citations do not seem full enough to make finding those books from Beirut and Cairo very easy. The author, Khairat Al-Saleh, is poet, painter and dramatist at once. Born in Jerusalem, she has studied and worked in Damascus, Cairo and now in Britain. The talented artist, Rashad N. Salim, is an Iraqi, a young man of cosmopolitan experience and style, a recent graduate of the Institute of Fine Arts in Baghdad.

Mathematics/Language

ANNO'S HAT TRICKS, text by Akihiro Nozaki, pictures by Mitsumasa Anno. Philomel Books ($11.95). Anno does not let us travel with him this time. He keeps us thoughtfully within the tiny world of one problem of logic. Here are small Tom cheerful in a white hat and Hannah in a red one, against a buff ground. A shadow—it is that of Shadowchild, the reader—stretches up from the bottom of the page. The trick is to guess the color of the Shadow's hat; we see that the hat is in place, but of course we cannot judge its color.

Naturally we have enough information to decide, from the statement that there are three red hats and two white ones from which to choose, from the two hats we see and from the fact that Hannah, who raises her hand to show it, knows what color her own hat is, although of course she cannot see it. The game is pursued over 20 pages, working from the simplest version to more challenging ones. A final few pages of text offer a clear account of


Grand Enchantment.
PRODUCT OF FRANCE
BOTTLED IN FRANCE
Triple Orange
Triple Orange
Grand Marnier
Liqueur
PARIS — FRANCE
750 ml (25.4 fl.oz) · 80° Proof
Gallardo
For gift delivery anywhere call 800-CHEER-UP (except where prohibited by law).
Product of France. Made with fi 1985 Carillon Importers, Ltd., N.Y.

the general case, put as an introduction to the powerful word "if."

The ingenious text is the work of a professor of mathematics at the International Christian University. The artist Anno has taken part because he holds that "the mathematical laws that underlie nature" are beautiful and can be presented so as to appeal to very young children. The two partners have done that trick too, if perhaps not for every child.

IS IT LARGER? IS IT SMALLER? by Tana Hoban. William Morrow & Company, Inc. ($11.75). About 30 beautifully composed and eye-catching photographs appear here with not one word of text, once title page and dedication are past. The organizing concept is absolutely clear, and the varied real world is seen so sharply that the single theme is enriched. The big mother sow nurses seven eager piglets; next to the big schooner its dinghies float quietly; one of the band blows a big tuba, the others toot smaller brasses, and a tiny hand rests trustfully on a big palm. The covers have two more photographs, one of pumpkin and squashes in the golds and reds of harvesttime. "This one is for my father," our author of a dozen such books writes in dedication, "who told me I was wonderful." How right he was!

WHAT'S INSIDE? THE ALPHABET BOOK, by Satoshi Kitamura. Farrar, Straus and Giroux ($11.95). A bunch of boxes stand against a brick wall. There is an open crate of strawberries, a big pineapple in a flat and a loose scatter of lemons and pears. Two almost-closed boxes, a hint of color showing from within, are labeled *a* and *b*. In the next picture we see the two boxes, lids now open. Red apples and yellow bananas appear, to no one's surprise. But what of that garbage can around the corner that bears the letters *c* and *d?* Again, a black paw at the edge of the lid offers a clue. These are easy, but by the time this interlocked chain of wildly romantic but careful imaginings reaches the snowy cemetery with its yellow tiger, the identity of the *u* and *v* so plainly marked on the coffin is less evident. Both whimsical and deductive, this funny book by a Japanese artist at work in London is a true original in its much traveled domain.

Perception

THE VERY BUSY SPIDER, by Eric Carle. Philomel Books ($14.95). MARMALADE'S PICNIC, by Cindy Wheeler. Print/Braille edition of National Braille Press Inc., 88 St. Stephen Street, Boston, Mass. 02115 ($5.95; a new title, costing from about $6 to $12, is available monthly from Children's Braille Book Club at this publisher). LOADS OF CODES AND SECRET CIPHERS, by Paul B. Janeczko. Macmillan Publishing Company ($10.95). Image and text both consist of symbols, of course. Still, they differ: the image has a certain loose geometric relation with its subject; the printed word is coded, all but arbitrary in form, although there are spoken words whose sound evokes their meaning. These books, each in its distinct way, exemplify this perceptual issue.

Eric Carle's spider spins her web step by step along a dozen pages of this very easy book for the youngest readers. Meanwhile the horse neighs, the cow moos, the dog woofs, a rhythmic sentence or so of text to the page. But the busy spider, preoccupied with her task, says never a word, although she ends up with a lovely intricate orb web. What is unusual is that the spider herself and her web at every stage of construction are represented to two senses. The spider is drawn bright red and green, and the web threads are gray. So much is plain to be seen. But just as you can hear those animal sounds when someone reads the text aloud, or you read aloud to yourself, so your fingers can follow by touch the form and positions of the spider and her work (and a pesky fly). Those particular lines are applied in raised ink. The result is a book of multisensory meaning, good for a variety of reading and speaking games, for coming to know books and for children with impairments of sight.

The second book deals with the orange cat Marmalade, who comes along one nice day to share the picnic (no cat food, please) and to ride home in the basket. Cindy Wheeler herself owns a black cat, although she has prepared several popular books for young readers about bright Marmalade. Her publisher, Random House, Inc., generously donated copies of the Marmalade picnic story to the publishers of this Braille-enriched edition. (Marmalade's tale is no longer available, but other children's publishers do the same each month with one of their titles.) Braille text is stamped onto both sides of translucent plastic sheets that are bound spirally as pages into the proper places, enlarging the original book of pictures and text.

The full text is rendered, and the visible pictures are briefly described, in Braille dots. Thus both sighted parents with blind children and blind parents with sighted children can share the serious and happy task with children reading and learning to read. Children and their elders who have no trouble at all seeing will certainly see even more, once they hold such an example of easy meanings recoded into patterns meant not for keen eyes but for sensitive fingers.

The third book is a very good first book on codes and ciphers of every kind. It begins with the simplest forms, code books, pictographs, pig Latin and its kindred, eve nth espa ceco de. It goes on to substitutions, cipher wheels, Morse telegraphy (in fact, Continental), with hints and frequency tables for cryptanalysis, the most important being to find a partner! There are examples and solutions throughout. A minimal Braille table (just the basic 26 letters) is to be found here, although getting through Marmalade's day demands a little more, namely Braille syllable signs and punctuation.

SPLIT SECOND: THE WORLD OF HIGH-SPEED PHOTOGRAPHY, by Stephen Dalton. Salem House, 250 Commercial Street, Manchester, N.H. 03103 ($17.95). Half a dozen strobed images down the page show a striped cat, launched upside down, successfully landing on its feet. This photograph was made in France; it is included here in a fine compilation by a leading English specialized photographer. A hundred or so color photographs and another 70 in black and white make up the book.

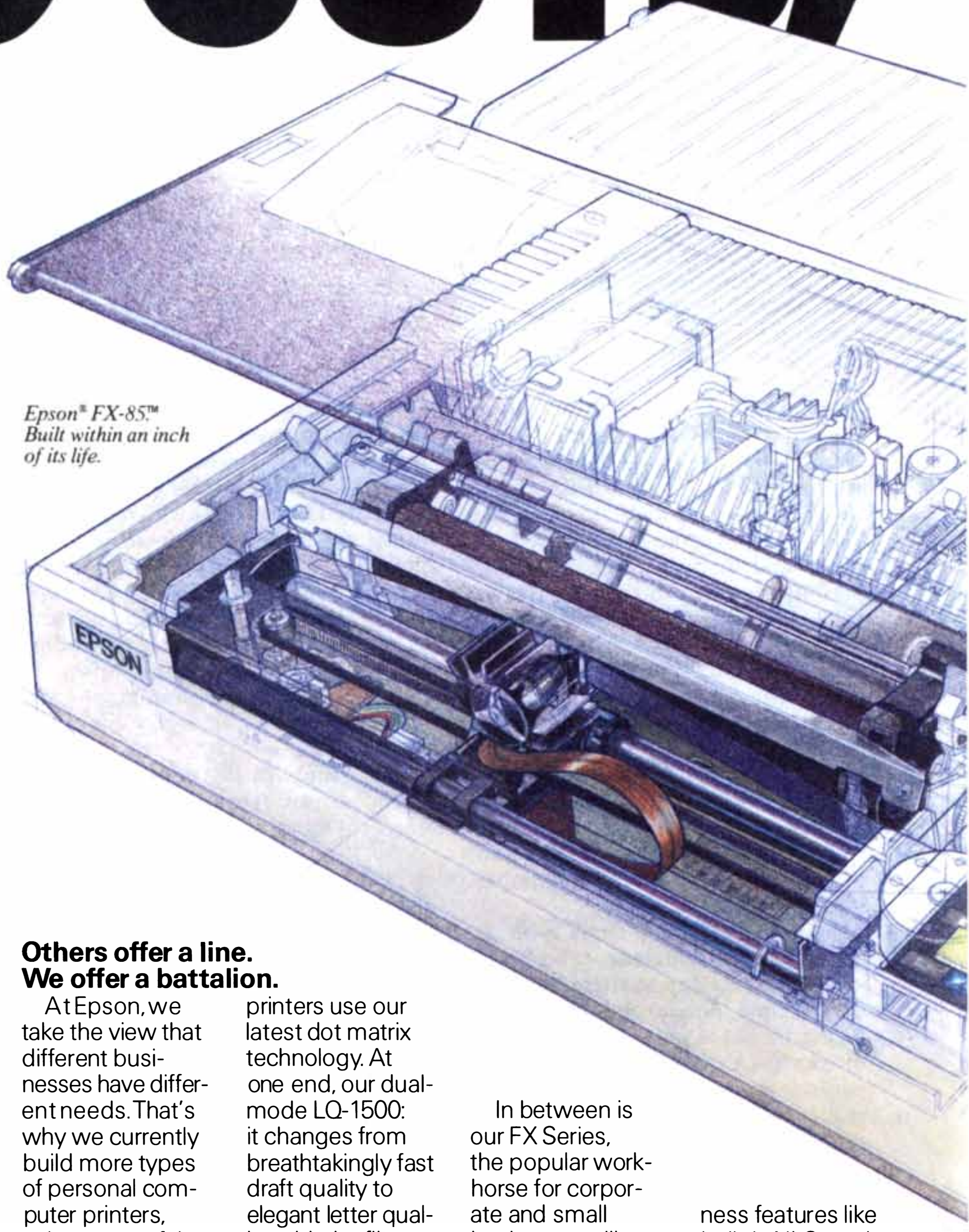
The history of this extension of our vision is not forgotten; we find the classical images from old E. J. Marey and Eadweard Muybridge, and those pioneer Edwardian splashes of A. M. Worthington. We see a dozen or so of the exploding apples, squashed tennis balls, milky coronets and many-armed dancers that come from Papa Flash himself, Harold E. Edgerton of M.I.T. But many others are new. A raindrop sequence, its fall into a deep puddle and a soft rubber ball abouncing are freshly instructive.

Unlike Edgerton, Stephen Dalton is no research photographer. He presents a wider variety of shots. They include the more contrived, showy and multiple images: the unexpected splashing wine, floating balloons and illusions of scale, all in a week's work for photographers whose clients include the trendy London music industry.

Most of the nearly 50 pictures here that were made by Dalton himself stop motions that are part of natural history, puffballs, beetles, birds and bats. A striking rattlesnake all agape and three shots of a diving kingfisher are hard to forget. Quick mechanical shutters, flashes of many types, light-beam and acoustic triggers, and the elaborate second-splitting techniques of the engineers are all outlined here; most are illustrated with examples. The emphasis is on still photography, although high-speed cinematographic techniques are described. This volume is a fine buy in its indispensable genre; it is pretty hard to imagine any youngsters who will not find much to ponder and argue over in these big pages.

A bullet tears through a playing card in a photograph from **Split Second**

NO GUTS,



Epson® FX-85™
Built within an inch of its life.

Most people credit the personal computer for the stunning transformation in business creativity.

But hold on.

It's one thing to run down the hall waving a *computer* with your latest brainstorm. Quite another to run down the hall waving a *piece of paper.*

Which is why, in 1964, Epson engineers invented the first dot-matrix printer to use with the computerized equipment they had designed for the Olympic Games. With it, at last, there was a complete computer *system* that could take man's most creative ideas, translate them into words and images, and churn them out the other end...on paper... in all their glory.

Small wonder that we prepare every Epson printer to be the guts of that system.

What being first has to do with being best.

For one thing, it means we've had the most time to refine a fairly sophisticated piece of equipment.

For another, Epson printers have become the industry standard for which virtually all major computer software is written. In fact, most Epson printers in use are hooked up to IBM® and Apple® computers.

Others offer a line.
We offer a battalion.

At Epson, we take the view that different businesses have different needs. That's why we currently build more types of personal computer printers, using more of the popular technologies, than IBM, Apple and one or two others... combined.

Eight of those printers use our latest dot matrix technology. At one end, our dual-mode LQ-1500: it changes from breathtakingly fast draft quality to elegant letter quality with the flip of a switch. At the other, our Homewriter™ 10: Epson reliability, and a very attractive price.

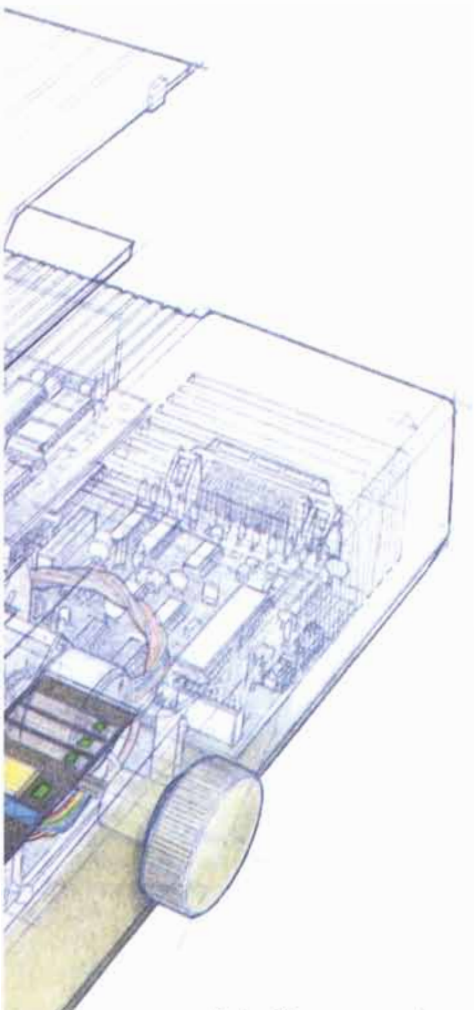
In between is our FX Series, the popular workhorse for corporate and small businesses alike, newly upgraded to offer a built-in Near Letter Quality mode, a standard 8K buffer, and even greater ease-of-use and compatibility with today's most popular PC's.

Also, the LX Series. Designed specifically for home, and home business usage, but with big business features like built-in NLQ mode, and SelecType pushbutton type selection.

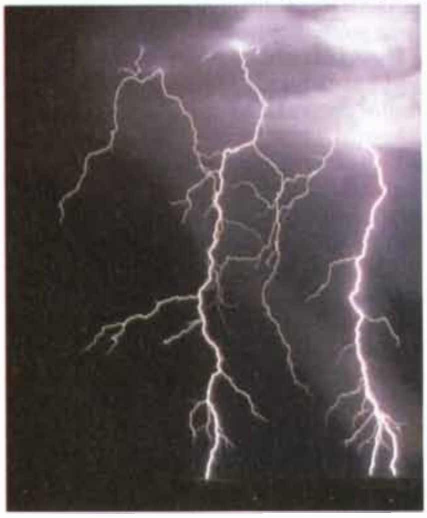
Plus specialized machines like our ink-jets, portables, color plotter, and color printer.

Together, they represent the most comprehensive choice of features and specifications available in the personal computer printer category.

NO GLORY.



A testing procedure some have labeled sadistic.



Every Epson printer goes through a series of tests, personally designed by our Czar of Quality Control, to make life in the real world seem tame.

For instance, we know our printers don't keep banker's hours. They turn it out all day, and often all night, without a break. So, in one test, we let a printer print, 24 hours a day, non-stop, day in, day out, to see if it will fail.

When we're designing a printer, we conduct what is affectionately known as the "run and stun" test, to see if it can withstand the rough voltage equivalent of a lightning bolt, in mid-operation, without a loss of character. If it can, it will certainly withstand minor annoyances like power surge and the static charge you carry from walking around on new carpeting.

The result? A printer that can take it as well as dish it out.

The Czar of Quality Control.

FX-85/185 FEATURES

Print Method:
9 pin, impact

Print Head Speed:
160 CPS draft mode
32 CPS NLQ mode

Character Sets:
96 ASCII set
italic set
11 internat'l sets
3 NLQ sets
11x9 user def. sets

Paper Feed:
FX-85 friction/pin
FX-185 frctn/tractor

Interfaces:
(standard)
Centronics® parallel
(optional)
RS-232-C serial
IEEE-488

Physical Dimensions:
FX-85 4" x 7.5" x 13.5"
FX-185 5" x 23.5" x 14"

Hardware Features:
8K Buffer
DIP sw. selectable
IBM 5152 and
Epson Std. Code.

At last, a warranty worth the paper it's printed on.

Most printers carry a ninety day warranty. Every Epson printer comes with a *full year* warranty. Not because you're ever likely to need it. But because it says something about how much faith we have in the way we build our machines.

No matter how good a warranty someone offers you, though, it's only as strong as the reputation of the people behind it. When you buy a printer from Epson, you're staking your business on how well we know our business. And we think, after twenty years, we have the guts of it figured out.

The glory is all yours.

Epson is a registered trademark of Epson Corporation. FX-85 and Homewriter are trademarks of Epson America, Inc. Centronics is a registered trademark of Data Computer Corporation. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computer, Inc.

For information on the complete line of Epson printers, call (800) 421-5426, or (213) 539-9140 in California.

EPSON
P R I N T E R S

Established 1845 SCIENTIFIC AMERICAN December 1985 Volume 253 Number 6

The Development of Software for Ballistic-Missile Defense

What some call a "Star Wars" defense would depend on computers to control an unprecedentedly complex array of weapon systems. Developing reliable software for such a defense may be impossible

by Herbert Lin

"The ultimate goal of the Strategic Defense Initiative (SDI) is to eliminate the threat posed by nuclear ballistic missiles," according to the interim charter of the Strategic Defense Initiative Organization. To achieve the goal of comprehensive defense, the organization is empowered by the Department of Defense to manage research programs that examine the feasibility of developing technology for a ballistic-missile defense (BMD) to protect cities and military assets. Such a defense would destroy or incapacitate nuclear-warhead delivery systems on the way to their targets. A wide variety of defensive weapons might be employed. They include lasers, particle beams, electromagnetic railguns and nonexplosive "kinetic kill" vehicles.

The Strategic Defense Initiative Organization recognizes that the computer technology that would control individual weapons and coordinate their operation is of comparable importance to the development of the critical new interception technologies. The command-and-control system for a comprehensive ballistic-missile defense must be capable of flawlessly receiving and acting on information pertaining to thousands of missile launches, tens of thousands of warheads and hundreds of thousands of decoys. It must do all this within the half hour it would take for an intercontinental ballistic missile (ICBM) to travel from a launch site in the Soviet Union to a target in the U.S. Because the system must be highly automated, there would be virtually no time for human intervention to correct unexpected failures. For the most part the execution of a computer program would replace human decision making once the BMD system was engaged.

Before the U.S. makes a serious effort to develop software for such a system, three questions should be considered. What is the nature of a BMD system? What are the obstacles to BMD software development? Can these obstacles be circumvented?

The comprehensive defense system most often discussed by Strategic Defense Initiative Organization officials (and the system's critics) consists of four tiers. That is, the ballistic-missile defense would attack hostile missiles in each of the four phases of their flight. The phases are the boost phase, during which a multistage launch vehicle carries the payload through the atmosphere; the postboost phase, during which nuclear warheads in reentry vehicles and "penetration aids" such as decoys are sequentially released above the earth's atmosphere by a maneuverable "bus"; the midcourse phase, during which the reentry vehicles and decoys traverse the greater part of their trajectory, and the terminal phase, during which the warheads in their reentry vehicles penetrate the atmosphere and detonate at their assigned targets.

A four-tiered ballistic-missile defense gives the interceptors several opportunities to destroy the offensive weapons. Within each tier a defensive system must successfully detect and track targets before it can destroy them [see "Space-based Ballistic-Missile Defense," by Hans A. Bethe, Richard L. Garwin, Kurt Gottfried and Henry W. Kendall; SCIENTIFIC AMERICAN, October, 1984]. Computers and appropriate software are needed to coordinate operation of the defense and evaluate its effectiveness. This coordination process is called battle management. Although experts are still undecided as to how a battle-management system should be organized, a hypothetical structure might include "local" computers and software that are responsible for battle management within each defensive tier. Each tier's system would then be connected with the other tiers through a global battle-management system.

The software guiding battle management within a given defensive tier would control the local sensors and weapons. These sensors would locate and track potential targets and distinguish actual targets from decoys. This part of the software might create a "track file" that contains all the known information about each target. The software could then allocate defensive resources in a specific tier by coordinating the track-file information with the available weapons and the programmed rules of engagement, rules that determine under what circumstances targets are to be attacked. The global battle-management system would assess the extent and nature of an attack in progress and specify the rules of engagement for each tier. In order to prepare a local battle-management system to engage warheads that have leaked through preceding tiers, the global system might pass on track-file and sensor information obtained in the boost-phase tier to each succeeding defensive tier.

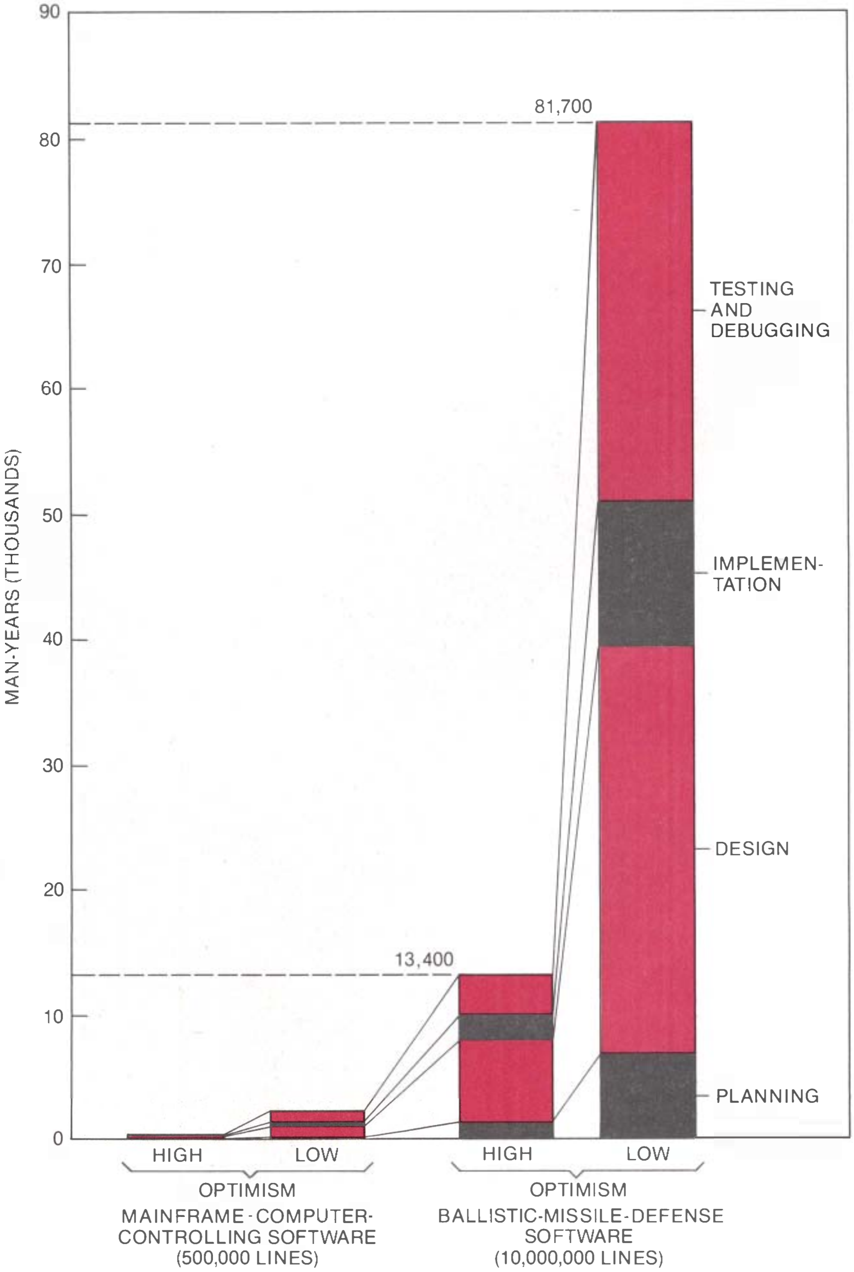
A ballistic-missile defense depends heavily on the software controlling it; defective software might lead to failure. Hence software development is a critical factor in attaining the objectives for which the system is designed.

Software development is an intellectual process that can be divided into distinct conceptual phases. These include planning, design, implementation, testing and debugging. In actual software-development projects the phases are not always carried out sequentially; for example, plans can change after a project has begun. Consequently developers may have to redesign a piece of software.

The first crucial area of software development is planning: the developer needs to determine what functions the software is to perform and to envision the different situations to which the software must respond. Specification, or deciding what action should be taken and at what time, becomes more difficult as a task grows in complexity. A simple example from everyday life makes the point. If the task is to count the number of people in a small audience, it is intuitively obvious how to proceed. If the audience consists of tens of thousands of people in a football stadium, the counting task requires much greater elaboration. For example, what defines the boundary of the stadium? How does one define a person? (Should the fetus in a pregnant woman be counted as a person?) Clearly, factors that appear infrequently and are therefore irrelevant in the first case may complicate the specification of a task as it grows in size.

The precise specification of what a ballistic-missile defense must do is a complicated task. For instance, the statement "Shoot down all Soviet missiles" is sufficient if the world contains only Soviet missiles and every Soviet missile should be shot down under all circumstances. But the world is not so simple. How can Soviet missiles be distinguished from non-Soviet missiles? What if a Soviet missile is headed for a target in East Germany? While these questions only scratch the surface of specification, they broach a fundamental problem developers will encounter, namely that it is often difficult to decide whether some particular aspect of a program is desirable or undesirable.

Moreover, developers will need to accurately predict every contingency and then decide how the software should respond to each. For instance, what procedures should the software follow if the computers in one defensive tier's battle-management system fail? How might a ballistic-missile defense differentiate a Soviet space-shuttle launch from that of a Soviet ICBM? Because of the nearly infinite number of possibilities, developers will be hard pressed to foresee all possible circumstances. Simply writing out all situations a BMD system might have to face and the appropriate actions to be taken in each case might plausibly require tens of thousands of pages. In comparison the U.S. Tax Code, which is the



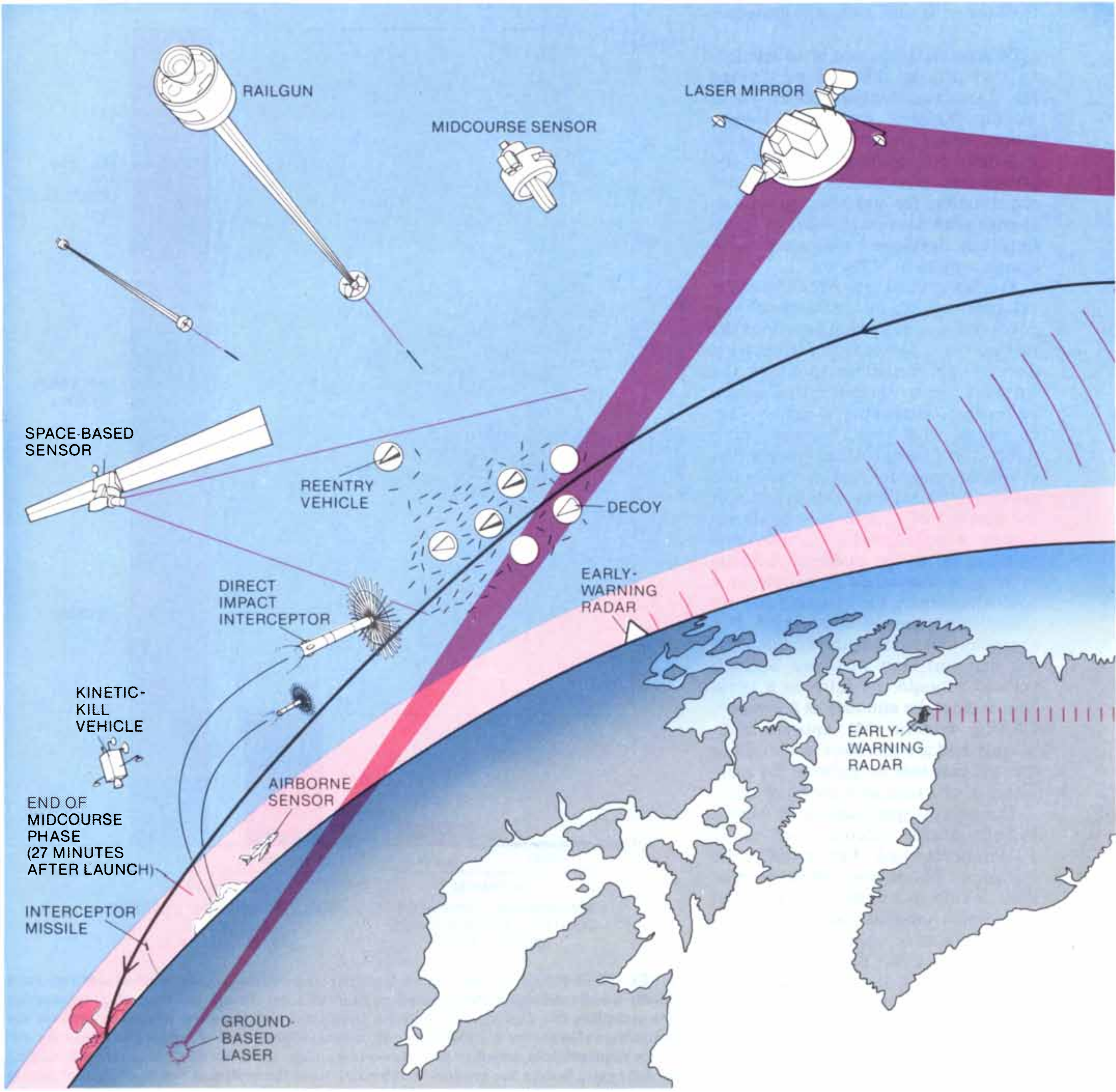
LEVEL OF EFFORT required for each of four stages in the development of software for a ballistic-missile defense (which would amount to some 10 million lines of programming code according to a Government study) is contrasted with the effort needed to produce the controlling software for a mainframe computer (500,000 lines of code). The stages are not strictly sequential; in actuality there is some overlap. Also, the final stage refers to testing and debugging before the product is delivered; once the system is "on line" further testing and debugging would be necessary. For each software-development project the bar at the left represents estimates based on highly optimistic assumptions; the bar at the right represents estimates based on less optimistic assumptions. The scaling function relating level of effort to program size is taken from *Software Engineering Economics,* by Barry W. Boehm.

legislative specification for Federal tax law, fills about 3,000 pages. The possibility that all potential scenarios would be adequately accounted for in the specification of the BMD software is at least as unlikely as the possibility that no unanticipated loopholes exist in the Federal tax laws.

Two examples may serve as further illustration. First, on June 3, 1980, the North American Aerospace Defense Command (NORAD) reported that the U.S. was under missile attack. The report was traced to a faulty computer circuit that generated incorrect signals. If the developers of the software responsible for processing these signals had taken into account the possibility that the circuit could fail, the false alert might not have occurred.

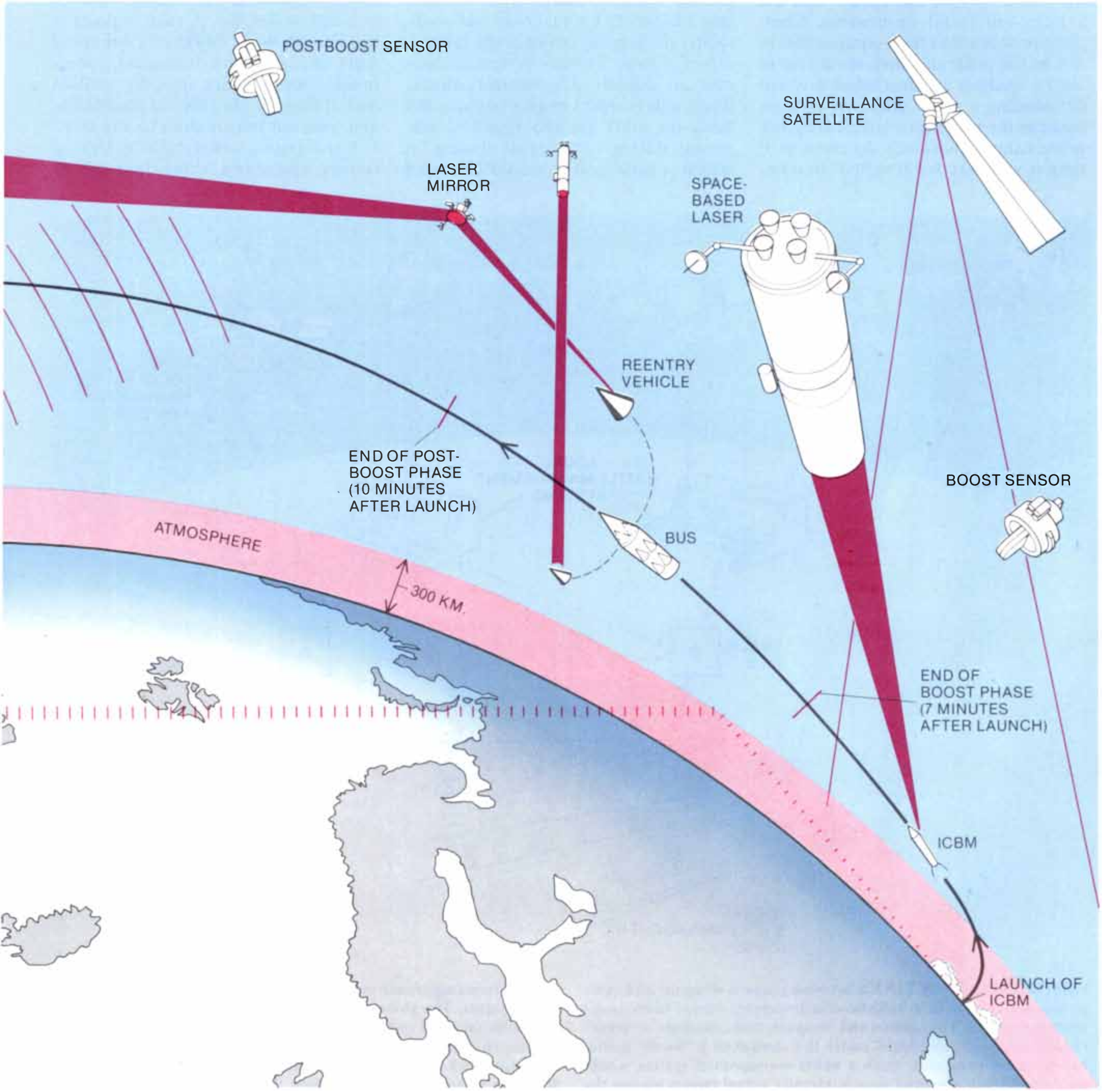
A second example, involving the sinking of the British destroyer H.M.S. *Sheffield,* came to light in the aftermath of the Falkland Islands war. According to one report, the ship's radar warning systems were programmed to identify the Exocet missile as "friendly" because the British arsenal includes the Exocet. As a result the system ignored the transmissions of a hostile Exocet's homing device and allowed the missile to reach its target, namely the *Sheffield.*

The design phase of software development is also rife with potential problems. In this phase developers ask how the specifications that emerged from the planning stage can be imple-



PROPOSED HARDWARE for a ballistic-missile-defense system includes an assortment of intercepting weaponry: high-powered lasers (based in space or on the ground), railguns and missiles. In addition satellite, airborne and ground-based sensors would be needed. Enemy ballistic missiles would be engaged in every phase of their trajectory: the boost phase, during which the missile is under powmented through computers. Developers conceive the algorithms that must be incorporated into the software, the sequence certain actions are to follow, and so on. Conceptually the design phase resembles the task confronting an architect who must draw the blueprints for a house after the requirements have been determined. If the architect incorrectly assesses the ability of the surrounding land to carry off water, the basement can flood, causing damage that is expensive to repair. Similarly, design errors in software development, if not caught in the design phase, can be expensive to correct.

Important software systems have not been exempt from design errors. For example, the manned space capsule *Gemini V* missed its landing point by 100 miles because its guidance program ignored the motion of the earth around the sun. In another case five nuclear reactors were shut down temporarily because a program testing their resistance to earthquakes used an arithmetic sum of variables instead of the square root of the sum of the squares of the variables. The lesson is that neither the nature nor the frequency of planning and design errors is predictable; indeed, these errors can be eliminated only if analysts can perceive them beforehand, a task that becomes more demanding as system size or complexity increases. Detecting and correcting errors is therefore a fundamental aspect of any software-development project. How then can errors



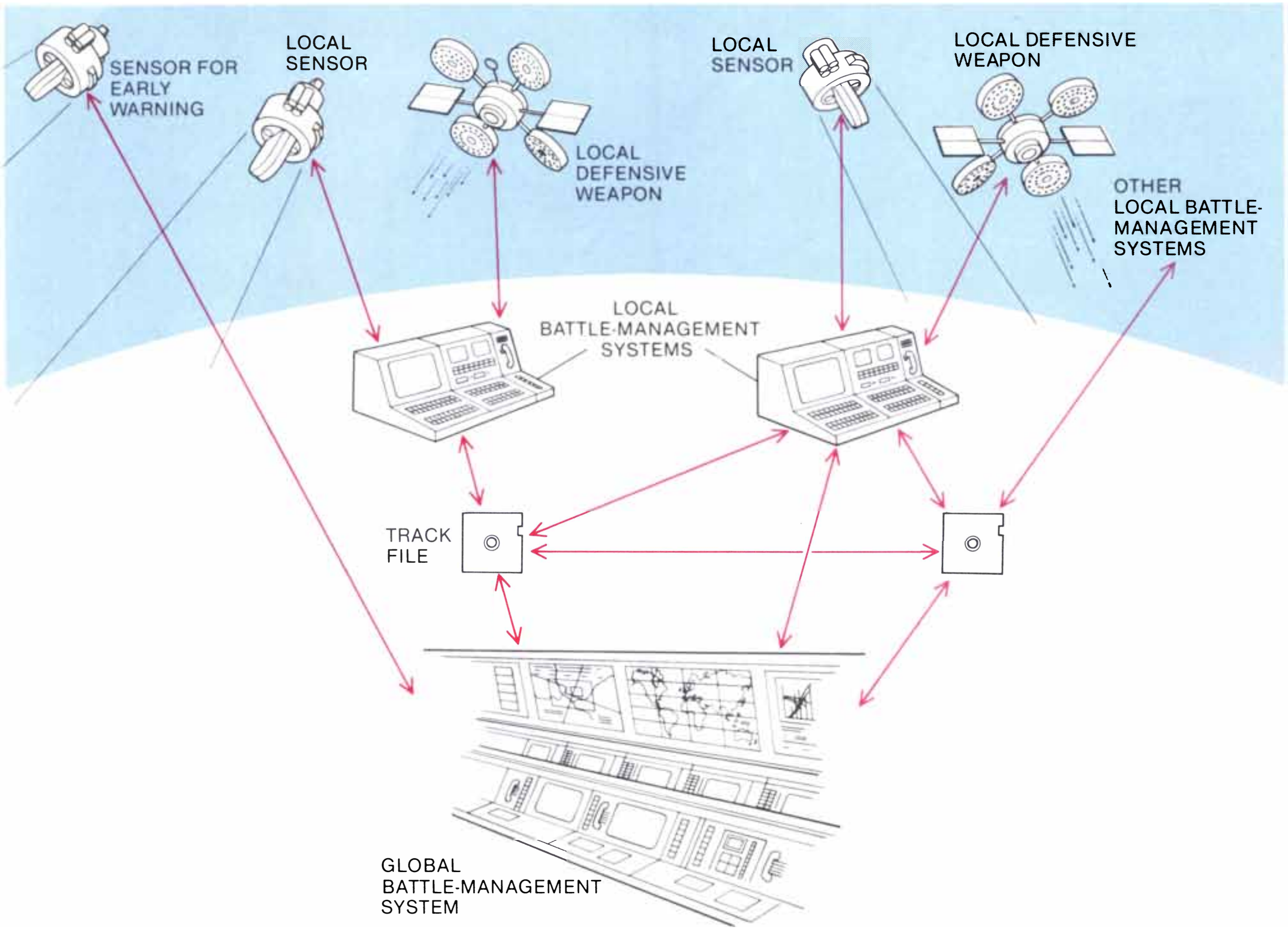
ered flight in the atmosphere; the postboost phase, during which reentry vehicles (equipped with nuclear warheads), decoys and obscuring metallic chaff are released; the midcourse phase, during which the various objects travel the major part of their course in space, and the terminal phase, during which the reentry vehicles penetrate the atmosphere to detonate over their respective targets.

be found and confidence in the operation of the system be assessed?

Two techniques are effective for assessing the trustworthiness of a software system. One technique is analytical. It requires that the program's correctness be proved by mathematically certifying that its output conforms to certain formally specified criteria when expected input data are received. Although proofs of program correctness help to ensure that a software system meets its theoretical specifications, they do not ensure that the system will fulfill its mission. Complete proofs are at least comparable in size to the programs they are trying to verify; analysts are thus faced with understanding a proof that is as complicated as the program itself. A program proof cannot guarantee the amount of time it will take for a system to complete "real-time computations" (a critical factor that will be discussed below), nor can it reveal the nature of the output when the system receives unanticipated input data. Finally, proofs of correctness cannot provide any promise that the actual program specifications themselves are correct.

A more important assessment technique is empirical testing. Clearly a ballistic-missile defense cannot be subjected to a large-scale empirical test under realistic conditions. The cost of such a test would be staggering. Of greater concern, however, is the fact that the Soviet Union could not confidently distinguish between the launching of a large number of test missiles and an actual U.S. nuclear attack. System developers must therefore fall back on more limited types of empirical testing: small-scale testing, in which a BMD system would be tested against a few missiles, and simulation testing, in which a computer would mimic large-scale threats to the targets a BMD system protects. Some experts maintain that such tests make actual large-scale ones unnecessary. Although such tests do increase confidence, they cannot ensure that the mission of a comprehensive ballistic-missile defense will be accomplished.

Problems that arise from integrating individual components into an effective system often appear only when the complete system is tested near its limits, something that cannot be done in small-scale tests. A case in point is the World Wide Military Command and Control System (WWMCCS), a communication network used by civilian and military authorities to coordinate and transmit information to and from U.S. military forces in the field. During routine operations, when the message



COMMUNICATION LINKS between sensors, weapons and computers are crucial in a ballistic-missile-defense (BMD) battle-management system. The sensors and weapons that constitute a "layer" of defense would be placed under the control of a "local" battle-management computer. Such a battle-management system would locate and track potential targets, identify actual targets among the debris and decoys, assign its weapons to specific targets at specific times and assess whether the target was successfully destroyed. All relevant information concerning the targets that have been engaged would then be placed on a "track file" and passed on to the next local battle-management system as well as a global battle-management system. The global system would be in constant communication with all local systems. For each of them it would specify the precise circumstances under which given targets could be attacked, transmit track-file information and coordinate actions to protect the BMD system itself. In addition the global battle-management computer would receive the first warning that an attack was under way. The network diagrammed here could be severely disrupted by an attack on the single vertex: the global battle-management system. Such vulnerability could be minimized through redundancy.

traffic is low, the WWMCCS has performed satisfactorily. When the message traffic is high, however, the performance of the system has suffered. In a 1977 exercise, when it was connected to the command-and-control systems of several regional commands, the WWMCCS had an average success rate for message transmission of only 38 percent.

As a way of overcoming the limitations of small-scale testing, developers often rely on simulators to create likely attack scenarios. Although simulation provides information beyond that achieved by small-scale testing alone, the technique, as applied to the testing of a ballistic-missile defense, is restricted in several ways. First, simulators may not be able to reproduce the "signatures," or parameters, of the physical phenomena associated with various events, such as simultaneous nuclear explosions, rapidly and accurately enough to test the responsiveness of the defense system. Analysts must therefore rely on assumptions about the signatures of each event they choose to simulate; that is, events are "preprocessed" before they are fed into the defense system. As a result of using preprocessed data, there can be no surprises during the simulation that might test how the system reacts to circumstances not anticipated by developers, and so the realism of the test is diminished.

Although advances in computer technology have greatly increased the computational speed of computers, and thereby facilitated the design of simulators that can model more realistic environments, increased speed will not help to simulate data on physical phenomena for which theoretical and empirical understanding is inadequate. For instance, given present knowledge, scientists would not be able to model accurately the simultaneous explosion of several closely spaced nuclear weapons under certain conditions. More important, simulators cannot satisfactorily duplicate all plausible attacks, since a determined and clever opponent ultimately chooses the parameters of an actual attack. Consequently any confidence in a BMD system that is based on simulated tests rests on the assumption that those responsible for the simulation can predict and reproduce in electronic form the range of tactics to which an enemy might resort.

The story of the Aegis air-defense system illustrates the limitations of simulation testing. The battle-management system for Aegis is designed to track hundreds of airborne objects in a 300-kilometer radius and then allocate enough weapons to destroy about 20 targets within the range of its defensive missiles [see "Smart Weapons in Naval Warfare," by Paul F. Walker; SCIENTIFIC AMERICAN, May, 1983]. Aegis has been installed on the U.S.S. *Ticonderoga,* a Navy cruiser. After the *Ticonderoga* was commissioned the weapon system underwent its first operational test. In this test it failed to shoot down six out of 16 targets because of faulty software; earlier small-scale and simulation tests had not uncovered certain system errors. In addition, because of test-range limitations, at no time were more than three targets presented to the system simultaneously. For a sizable attack approaching Aegis' design limits the results would most likely have been worse.

Errors are not unexpected in operational tests; indeed, malfunctions are inevitable during the initial shakedown exercise of a new weapon system. By the time of the next Aegis tests, they had been corrected and none recurred. Future exercises will probably uncover additional errors, which will in turn be fixed. Through this process the performance of the Aegis system will gradually improve.

Unlike the performance of Aegis, the performance of a comprehensive ballistic-missile defense against a large-scale attack will not improve with experience. Because large-scale empirical testing is impossible, the first such test for a comprehensive BMD system would be an actual large-scale attack on the U.S. A more complex battle-management system that must keep track of more targets and operate under tighter timing constraints is unlikely to have a better performance record than Aegis has. In addition a severe software failure in a full-blown battle situation would offer little or no opportunity for system designers to learn from the experience.

Nevertheless, developers of a ballistic-missile defense will attempt to improve the system's performance by testing for errors and eliminating them, by expanding software capabilities in response to newly perceived needs and by adding new hardware and concomitant software to the project. These efforts, along with bringing new workers into the project, come under what specialists call software maintenance. It can account for approximately 70 percent of the total life-cycle costs of a software-development project.

Two major maintenance issues are particularly salient in relation to a system for ballistic-missile defense. First and most important, significant errors discovered after the software has been put into operational use must be eliminated. These errors can range from an incorrect coding symbol to a fundamental design flaw. On June 19, 1985, the Strategic Defense Initiative Organization did a simple experiment. The crew of the space shuttle was to position the shuttle so that a mirror mounted on its side could reflect a laser beamed from the top of a mountain 10,023 feet above sea level. The experiment failed because the computer program controlling the shuttle's movements interpreted the information it received on the laser's location as indicating the elevation in nautical miles instead of feet. As a result the program positioned the shuttle to receive a beam from a nonexistent mountain 10,023 nautical miles above sea level. This small procedural error was of little significance to the test itself, however; a second attempt a few days later was successful. Nevertheless, the event shows that even simple errors can lead to mission failure.

Although the bug in the space-shuttle experiment was easily rectified, removing errors from a ballistic-missile defense will require considerably more effort. In particular the battle-management software must receive and process information in such a way as to keep pace with circumstances external to the computer during an attack. That is, the software must run in "real time" [*see illustration on next page*]. One problem with debugging a real-time software package is ascertaining why it may operate with one specific configuration of equipment but not with another that differs only slightly. For example, a certain missile can be currently launched at supersonic speeds by the F-4G Wild Weasel aircraft but not by the F/A-18 Hornet: a software system that works in the F-4G is not entirely compatible with the avionics of the F/A-18.

A second problem of particular significance to debugging real-time software is that analysts often find it difficult to make errors recur, something that is essential if bugs are to be located. Real-time output is often determined by factors such as the arrival times of sensor inputs. These determinants of a program's behavior cannot always be reproduced with enough accuracy for the error to be found and eliminated. Indeed, finding an error requires an understanding of the precise circumstances leading to it. Because that is not always possible with large real-time systems, software errors may be identified in practice only in a probabilistic sense. Large computer programs, in fact, usually evolve through a series of incompletely understood changes. After a while the programmer can no longer predict outcomes with confidence; instead he can only hope that the desired outcome will be attained.

There is a final complication to debugging real-time software: even if an error can be located, attempts to eliminate it may not be successful. The probability of introducing an error (or more than one) while eliminating a known error ranges from 15 to 50 percent. Moreover, the majority of software-design errors that appear after software is put into service do so only following extensive operational use. Experience with large control programs (ones consisting of between 100,000 and two million lines of code) suggests that the chance of introducing a serious error in the course of correcting original errors is so large that only a small fraction of the original errors should be remedied.

In the context of a comprehensive ballistic-missile defense, one should therefore ask about the consequences of an error that would manifest itself infrequently and unpredictably. The details of the first operational launch attempt of the space shuttle provide an example. The shuttle, whose real-time operating software is about 500,000 lines of code, failed to take off because of a synchronization problem among its flight-control computers. The software error responsible for the failure, which was itself introduced when another error was fixed two years earlier, would have revealed itself, on the average, once in 67 times.

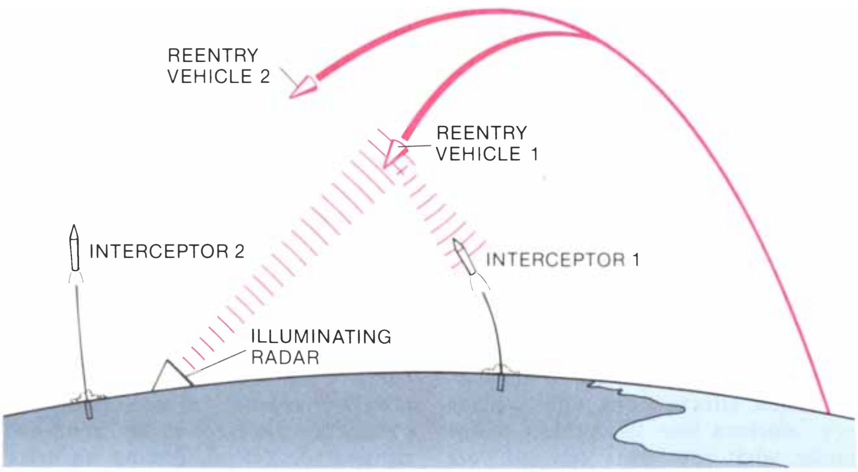
Aside from detecting and removing errors, a second aspect of software maintenance stands as a hurdle to developing a defensive system: the management of product development. The Defensive Technologies Study Team (DTST), chartered by the Department of Defense to examine the feasibility of a comprehensive ballistic-missile defense, estimates that the entire system will require a minimum of 10 million lines of programming code. In comparison, the entire software system for Aegis is an order of magnitude smaller. If the DTST estimate is low by a factor of only two, even a very optimistic software-development project will entail more than 30,000 man-years of work, or at least 3,000 programmers and analysts working for about 10 years. For this reason the project can expect a staff turnover that will reduce the institutional memory of the project. During staff transitions it is conceivable that an essential detail, such as updating a particular subprogram, could be overlooked. A tragic example of management error occurred in 1979, when an Air New Zealand airliner crashed into an Antarctic mountain; its crew had not been told that the input data to its navigational computer, which described its flight plan, had been changed.

It is also possible that Soviet agents might try to sabotage the project. They could, for instance, deliberately introduce hard-to-find flaws into the system that would become evident during a real attack but not during testing. The possibility that a program might contain some hidden "time bomb" that would cause the software to fail at a critical moment cannot be ignored. The obvious preventive step would be to impose security clearances on a "need to know" basis. Such restrictions, however, would inhibit communication among the staff working on different parts of the system and thereby increase the likelihood of serious, unintentional software errors arising.

Tightening security will not prevent the Soviet Union from thwarting the system by several other means. There is no reason to believe that Soviet officials faced with a U.S. BMD system would not try to disguise the observable characteristics of their ballistic missiles, warheads and decoys during tests or in actual use. Without reliable data on Soviet missiles, U.S. developers cannot be sure the defensive system will work. Furthermore, the Russians may develop new tactics or weapons that would force U.S. analysts to reprogram the battle-management system to meet new threats. Accumulated changes over time would probably result in many unpredictable interactions and could ultimately necessitate a total redesign of the system.

It has been suggested that the development of software for exceedingly complex systems could be facilitated by reliance on "expert systems" and automatic programming. For example, a Defense Advanced Research Projects Agency (DARPA) report stated that expert systems may be applicable to a ballistic-missile defense. Expert systems are detailed descriptions, expressed as computer rules, of the thought processes human experts use to reason, plan or make decisions in their specialties. Typically, expert systems use informal reasoning procedures or rules of thumb. For example, an expert system may take the following as a given: If a Soviet ICBM is launched, it is a threat to the U.S. It may then use the converse of the statement as a legitimate rule of inference: If there is a threat to the U.S., it is a Soviet ICBM launch. The validity of this inference tool cannot be proved or disproved using the standard tools of formal logic; the statement is not true all the time but is sensible most of the time. The reliance on informal reasoning procedures in programming expert systems leaves open the possibility that deeply embedded conceptual contradictions in the system could exist that might cause it to fail.

To date expert-system research has focused on well-defined areas such as biochemistry and internal medicine. These are areas in which human expertise is sufficiently developed to allow expert systems some success. Even so,



"REAL TIME" SOFTWARE must be executed at a rate defined by the timing of events external to the computer hardware. A computer controlling the radar illumination of an incoming reentry vehicle so that an interceptor missile can home in on the reflected signal would have to maintain the radar beam on the reentry vehicle until the vehicle was destroyed. If the radar were turned away too soon, the interceptor might "lose" its target, particularly if the reentry vehicle were designed to evade approaching interceptors. Yet the computer must re-aim the radar beam so that a second reentry vehicle could also be intercepted. The computer program has to shift the beam to the second vehicle once the first is destroyed, and it must do so quickly enough to enable a second interceptor to reach its target.

the technology is still limited. Those who recommend applying expert systems to battle management for ballistic-missile defense ignore the fact that human expertise is based on human experience. No one has expert knowledge of massive nuclear missile attacks based on experience.

Equally improbable is the idea that automatic programming, the use of computer programs to write other programs, will provide a solution. Air Force Major Simon Worden, special assistant to the director of the Strategic Defense Initiative Organization, has stated that "we're going to be developing new artificial intelligence (AI) systems to write the software. Of course, you have to debug any program. That would have to be AI too." Statements such as this one are misleading. The primary function of automatic programming is to alleviate the technical difficulties of implementing design specifications and modifying existing code. Roughly half of all software errors, however, are the result of human choices and decisions made during the planning and design phases. These human choices could not be delegated to automatic programming.

Two scenarios based on the above discussion show how the software for a comprehensive ballistic-missile defense might fail. Suppose a battle station that has successfully intercepted two missiles in operational testing using an electromagnetic railgun is faced with a large-scale Soviet missile attack. At first projectiles launched from the battle station destroy their targets. Unfortunately at the design stage developers neglected to take into account the fact that less massive objects have more recoil. As more projectiles are fired the recoil becomes stronger and skews the aiming algorithms for the railgun. As a result later projectiles are too slow and do not reach their targets in time to destroy them. Or suppose that, during the early phase of a nuclear war, U.S. and Soviet leaders agree to a cease-fire. The U.S. leaders realize that one submarine captain will not receive the cease-fire message in time but are confident that U.S. missile-defense satellites will be able to shoot down any missiles launched by mistake. Only after the submarine's missiles are launched does it become tragically clear that no developer had specified that the satellites might have to shoot down U.S. missiles. The missiles explode over the Soviet Union with catastrophic consequences.

Because they are anticipated, neither of these errors is likely to occur. The problem is not whether a system contains a particular error but rather how likely it is that the system will contain any one of millions of potential errors. The primary matter of concern is "unknown unknowns," the potential errors that remain unanticipated.

SM-2 MISSILE is part of the U.S. Navy Aegis air-defense system. The Aegis battle-management system was designed to track hundreds of airborne objects and to carry out up to 20 simultaneous intercepts. During its first operational test it failed to shoot down six of 16 targets because of faulty software. The system's malfunctions have since been repaired. A comprehensive ballistic-missile defense, which is a much more complicated and ambitious weapon system, must perform far better than the Aegis system did, and it must do so in its first try; a severe software failure during the system's first full-scale test would offer virtually no opportunity to rectify the problem. Because large-scale testing is impossible, the first empirical test of the BMD system would be an actual large-scale ballistic-missile attack.

All the problems cited in this article are related to software planned, designed, implemented and debugged by experienced software engineers; most, if not all, were corrected. The general technique for correcting these errors, namely discovering the bugs in operational use and then correcting them, is unlikely to be fruitful in the development of a BMD system. Yet a comprehensive ballistic-missile defense requires not only that software operate properly the first time, in an unpredictable environment and without large-scale empirical testing, but also that planners are positive it will do so. The Reagan Administration has stated that empirical testing is essential for maintaining the reliability of nuclear weapons and has opposed a comprehensive test ban on these very grounds. One wonders how it is possible to have confidence in a comprehensive ballistic-missile defense, which will be at least as complicated as nuclear weapons are now, without requiring it to meet comparable test standards.

Proponents of the Strategic Defense Initiative argue that even in the absence of large-scale empirical testing a ballistic-missile defense is still desirable, because the possibility that the system might work reasonably well would deter Soviet leaders from contemplating an attack. Others argue that the goals of the Strategic Defense Initiative are really more modest than the goal of a comprehensive defense against ballistic missiles, suggesting that a BMD system might be most appropriate for protecting military assets such as missile silos. Of course, for goals more circumscribed than those of comprehensive ballistic-missile defense, perfection is not a requirement and testing is much easier. Although a program that has more limited goals does have a much higher probability of success, these limited goals raise a disturbing issue. In particular they are inconsistent with President Reagan's stated vision of eliminating the threat of nuclear ballistic missiles. Consequently it is proper to judge the feasibility of the Strategic Defense Initiative on the basis of the president's challenge to the scientific community. By that standard no software-engineering technology can be anticipated that will support the goal of a comprehensive ballistic-missile defense.

Optical Phase Conjugation

In everyday experience time always moves forward. The situation is qualitatively different, however, in the case of wave motion: light waves can be "time-reversed" and made to retrace their trajectories

by Vladimir V. Shkunov and Boris Ya. Zel'dovich

Imagine a sportswoman standing on a springboard, preparing to make a high dive. A quick run, a flight—but because of a small technical mistake her body enters the water at a wrong angle, making a big splash and causing waves to spread from the point of contact. How wonderful it would be to reverse time in order to correct the mistake and get a high score! Sprays of water would come together, waves would run backward to the point of contact, the diver would be thrown out of the water and onto the springboard and the surface would become as smooth as it had been before the dive. Unfortunately, although such a scenario can easily be portrayed with the help of a motion-picture projector, the time-reversed process contradicts our everyday experience, and for good reason: the sequence of events violates the second law of thermodynamics (the law that states that systems tend toward maximum entropy).

The scenario can be played successfully, however, if the actor is the wave motion of light or some other electromagnetic radiation. Such a phenomenon is possible because of a remarkable property of light rays that has been known for a long time, namely the reversibility of their propagation. For every light beam that has an arbitrary structure of rays there exists a possible "time-reversed" beam whose rays run along the same trajectories but in the opposite direction, as when a film is run backward. The success in the reversal of wave motion is due to an extreme simplification of the problem: the quantum-mechanical and thermal motions of atoms and electrons that radiate and refract light do not need to be reversed. It is sufficient, for practical purposes, to reverse the temporal behavior of macroscopic parameters describing the averaged motion of a large number of particles.

The existence of reversed beams has important consequences. It is evident, for example, that an ideal beam, or one that is free of distortion and has minimum divergence, can be degraded by transmitting it through inhomogeneities, such as a glass plate of nonuniform thickness. The property of reversibility means it is possible to create an "antidistorted" beam that becomes ideal after being transmitted backward through the inhomogeneities. The technology by which such beams are created and manipulated is called optical phase conjugation. The waves making up such a beam are called phase-conjugate waves.

To describe the properties of a phase-conjugate wave we must first discuss some of the basic concepts of wave motion. As a set of waves moves through space its oscillations arrive at different points at different times. Points at which oscillations are synchronous are said to be in phase. A phase is a stage of a period in relation to some starting position. The surfaces connecting points that have the same phase are known as wave fronts. An important feature of the wave-front surface is that it is perpendicular to the direction of wave propagation. Wave fronts of plane waves are planes and wave fronts of spherical waves are concentric spheres. The wave fronts of actual light beams may have rather complicated shapes and topologies.

The concept of a wave front can be utilized to understand the properties of a phase-conjugate wave. Suppose a photograph of a light wave is made in which the beam is propagating from left to right [*see top illustration on page 57*]. Owing to the reversibility of wave propagation, someone examining the photograph would not be able to tell whether the direction of propagation is from left to right or from right to left. If the beam were propagating from right to left (that is, if the beam were reversed), however, the wave fronts would be reversed with respect to the direction of the beam. That is why in the Russian language the process of generating a reversed wave is called wave-front reversal.

The relation between the wave fronts of two mutually reversed waves is analogous to the relation between the positions of two opposing armies on a military map. The front line of each army coincides with that of the other, and the directions of desirable movement are opposite. One can say that the front lines are mutually reversed: a convex part of one army's front corresponds to a concave part of the other.

Expressed in different language, the phase difference between any two points of the reversed beam has a sign opposite to that of the phase difference between the same points of the original beam. The mathematical operation of changing a phase sign is known as conjugation, and that is why the term optical phase conjugation was coined and subsequently adopted in the English-language scientific literature.

How does one conjugate (that is, reverse) a wave? It is easy to generate the phase conjugate of a plane wave: one simply mounts a plane mirror so that the wave is reflected precisely backward. It is not much more difficult to conjugate a wave that has a spherical wave front. A concave mirror in the shape of a section of a sphere would be mounted so that the center of the mirror corresponds to the source of the wave. Then at every point on the mirror the rays would be incident perpendicularly to the surface and would be reflected precisely backward.

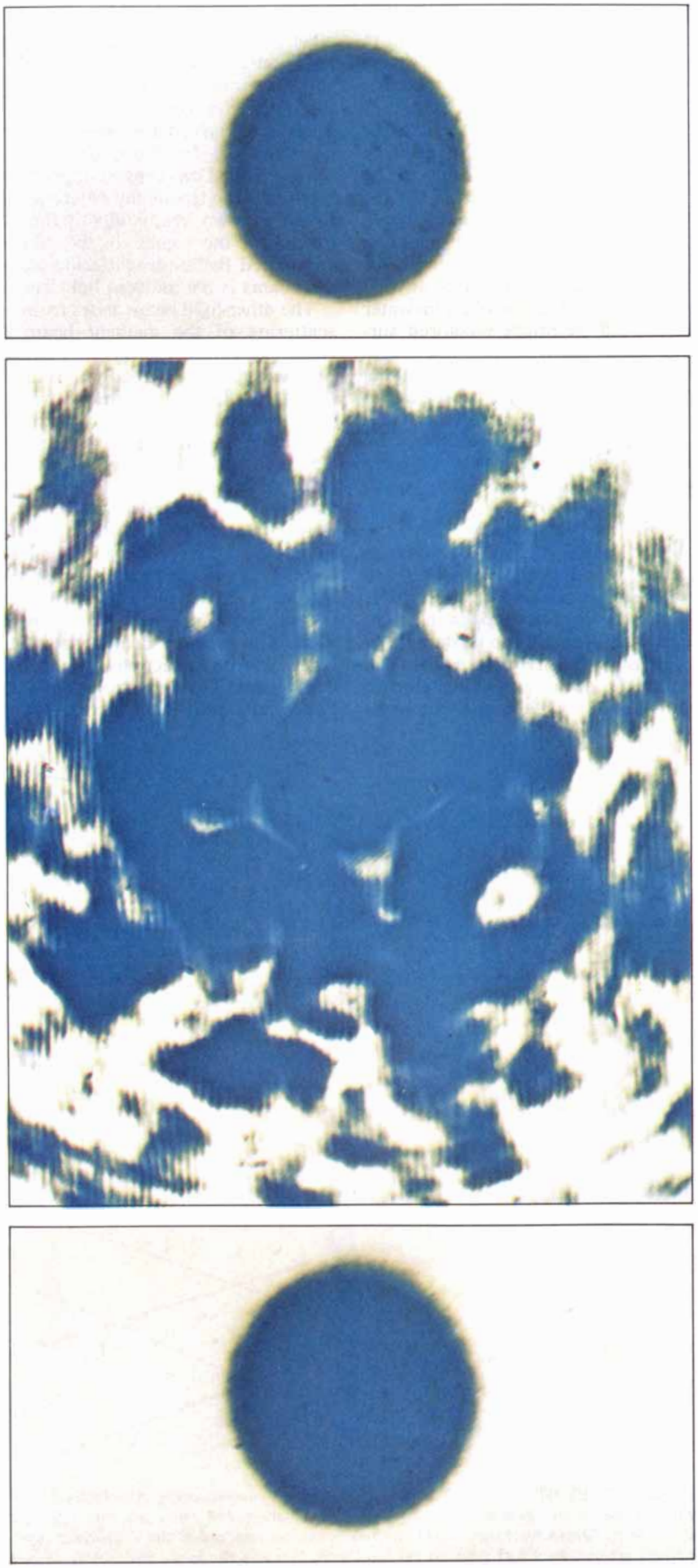
To conjugate a beam that has an arbitrary wave front one could in theory position a mirror whose profile coincides with that of the wave front. Unfortunately such a method is difficult to realize in practice. First, one would have to make a new mirror for each particular incident beam. Second, even the shape of a wave front of a laser beam can change during a brief pulse; one would therefore have to change the shape of the mirror continuously to match the changing shape of the wave. Finally, the precision required to prepare and position such a mirror would be extremely high.

To produce a phase-conjugate wave a medium or surface is required whose properties are affected by the characteristics of waves that are incident on it. That dependence allows the medium or surface to adjust itself with such delicate correspondence to the structure of the incoming beam that a reflected phase-conjugate beam is produced under certain conditions. Fortunately such materials exist. They are termed optically nonlinear.

Two widely used methods of phase conjugation that rely on such materials are stimulated Brillouin scattering and four-wave mixing. Stimulated Brillouin scattering, which is one of the most beautiful effects in nonlinear optics, was discovered in 1964 by Raymond Y. Chiao, Boris P. Stoicheff and Charles H. Townes, who were then working at the Massachusetts Institute of Technology. The effect involves directing a beam of light into a transparent medium such as liquid, compressed gas, glass or crystal. Light of low intensity passes through such a sample with virtually no attenuation. The behavior of a high-intensity light beam, however, is astonishing. Beginning at a threshold power of roughly a million watts the beam is reflected backward almost completely. Although such a power is rather high, it is easily attainable with a laboratory pulsed laser.

The reflected beam is a consequence of events that produce Brillouin scattering (named after the French physicist Louis Marcel Brillouin). In Brillouin scattering a sound wave is directed into a solid, a liquid or a gas. A sound wave causes the density of the material in which it is traveling to change in a periodic way. The resulting pattern, which moves with the wave through the material, consists of alternating zones of compression and rarefaction.

Because the zones of compression are denser than the zones of rarefac-



A PHASE-CONJUGATE, or "time-reversed," light beam can compensate for distortions introduced by an optically inhomogeneous medium (such as frosted glass). A highly directed laser beam (*top*) was distorted, resulting in a degraded beam (*middle*). The degraded beam was then reversed by optical phase conjugation. The backward transmission of the phase-conjugate beam through the inhomogeneous medium restored the quality of the beam (*bottom*).

tion, the behavior of a light wave directed into the material is different in the two zones. Specifically, the index of refraction of the first kind of zone is slightly different from the index of refraction of the second. (The index of refraction of a material is the ratio of the speed of light in a vacuum to the speed of light in the material.) If the separation between the zones is exactly half the wavelength of the incoming light, the light will be reflected. This kind of reflection is familiar to anyone who has seen a thin film of oil on water and noticed its rainbow-colored surface. At each point on the film one color is reflected better than it is at all the others, and it is that color whose corresponding wavelength is half the thickness of the film layer there. Since the thickness of the layer varies, different colors are reflected at different points.

In stimulated Brillouin scattering the sound wave, or pressure-density variation, is not applied externally to the material; it is stimulated internally by pairs of counterpropagating light waves. Just as sound can be thought of as a pressure-density wave, so light can be thought of as a moving electric field. An electric field can compress materials, a phenomenon known as the electrostrictive effect. If an electric field pattern moved at the speed of sound through a material, it could therefore give rise to a sound wave. Such an electric field pattern can be generated by the interference of two optical beams traveling in opposite directions if the frequency difference between the two beams equals the frequency of the sound. In the case of stimulated Brillouin scattering one of the beams is the incident light beam.

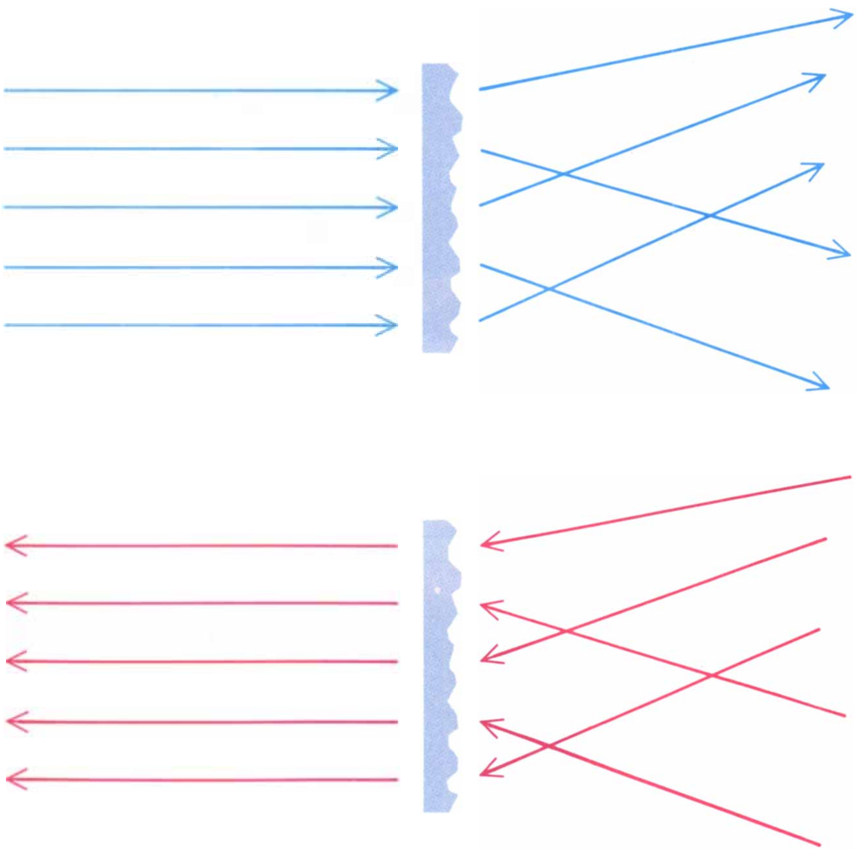
The other light beam arises from the scattering of the incident beam by small, statistically distributed density fluctuations in the medium (that is, thermally fluctuating sound waves). When the frequency and direction of a scattered wave are just right, the wave will interfere with the incident beam and amplify the pressure-density variations in the material. The variations subsequently lead to the reflection of a minute portion of the incident beam. The reflected portion interferes in turn with the incident beam, generating more pressure-density variations. The pressure-density variations lead to more reflections of the incident beam. The reflections build exponentially as distance increases, until a reflected beam emerges from the material. Since the amplification depends on the intensity of the incident beam, however, a prerequisite for producing the reflected beam is that the power of the incident beam must exceed a certain threshold.

Phase conjugation by stimulated Brillouin scattering was first achieved by Valery V. Ragul'skii, Vladimir I. Popovichev, Fuad S. Faizullov and one of us (Zel'dovich) at the P. N. Lebedev Physical Institute in Moscow in 1972. The main trick that allowed phase conjugation to take place was the use of a special glass plate that had been made nonuniform by etching with hydrofluoric acid. A beam of red light from a pulsed ruby laser was distorted by being passed through the plate. The distorted beam was directed into a pipe one meter long, four millimeters wide and four millimeters high that had been filled with gaseous methane at a pressure of 140 atmospheres. Stimulated Brillouin scattering occurred in the pipe, and the reflected beam, when it was passed backward through the same etched plate, emerged undistorted. That is to say, its structure was identical with that of the incident beam.

Phase conjugation by stimulated scattering has now been realized in a large number of scattering mediums and many types of lasers. The main advantage of the technique is that it only requires a cell filled with an appropriate gas, liquid or solid. The simplicity of the setup has prompted Robert W. Hellwarth of the University of Southern California to comment that in stimulated scattering "Nature surely loves the phase-conjugate beam." It is interesting that in 1977 Hellwarth proposed another way to conjugate waves: four-wave mixing.

Four-wave mixing is now the other popular method of phase conjugation. It involves the interference of four light beams in a nonlinear medium. Three of the beams are input beams: one is the object beam whose phase conjugate is sought and the other two are reference beams. The reference beams, which travel in opposite directions with respect to each other and are usually plane waves, have the same frequency as the object beam itself. The object beam may enter the medium from any direction. The fourth beam, an output beam, is the phase conjugate of the object beam and emerges along the same line as that of the object beam, although its direction of propagation is reversed.

The conjugate beam is produced by perturbations in the medium due to the interference of the object beam with



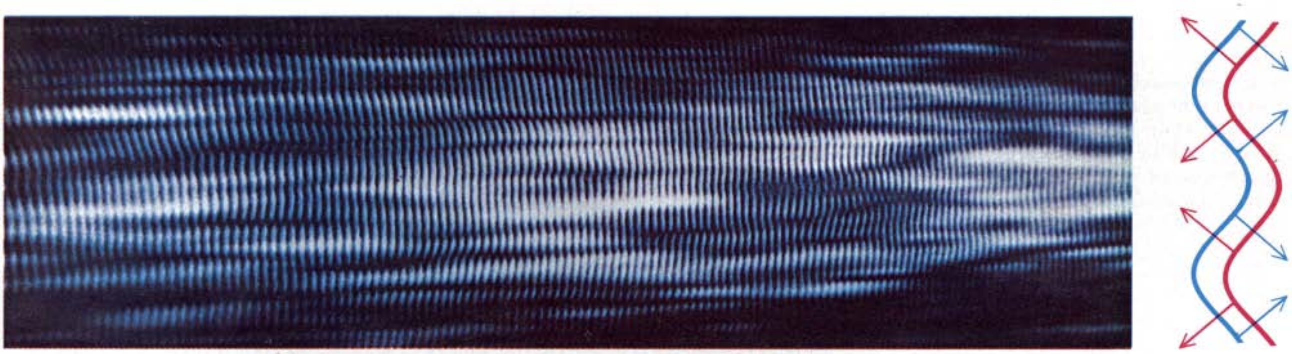
REVERSIBILITY OF LIGHT WAVES has important consequences. An ideally directed beam (one that is free of distortion and divergence) is degraded when sent through a glass plate of nonuniform thickness (*top*). The beam can be restored if the individual rays of the beam are reversed and transmitted backward through the same glass plate (*bottom*).

one of the reference beams. Wherever the electric fields of the waves oscillate with the same phase, the fields are added and the local intensity of light is high. Where the fields oscillate with the opposite phase, the fields are subtracted and the local intensity of light is low. The zones of high intensity are sandwiched between the zones of low intensity. The size, shape and orientation of all the zones are determined by the characteristics of the interfering fields. All the information about the phase of the object beam is therefore stored in what is known as an interference pattern, which manifests itself in the medium as a series of zones that have different refractive indexes.

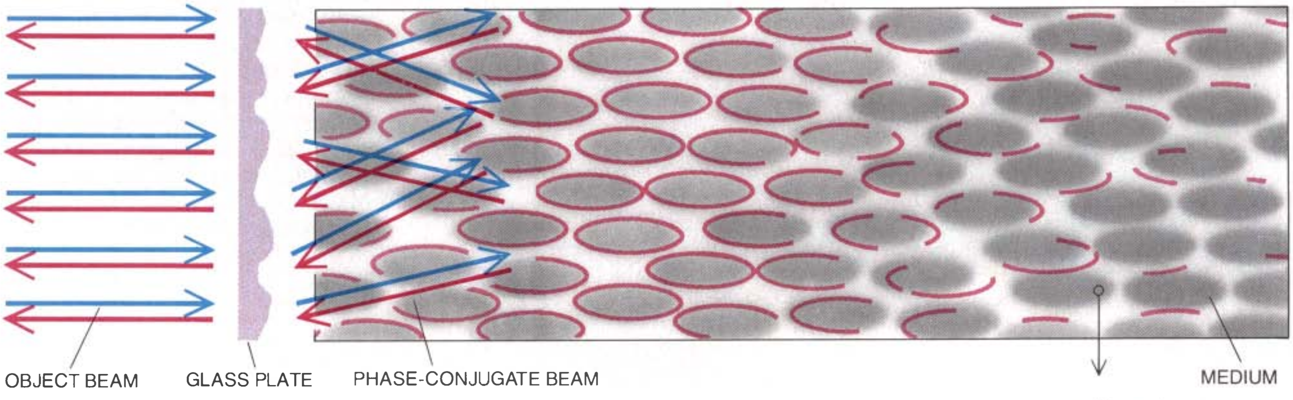
The second reference beam is reflected by the zones of the interference pattern, that is, the beam reads out the information about the phase structure stored in the pattern. Since the second reference beam comes from a direction opposite to that of the first one, the reflected beam is the phase conjugate of the object beam. Although the reflection from each zone of the pattern is weak, the sum of all such reflections is large and a considerable amount of energy can be transferred from the reference beam to the conjugate beam.

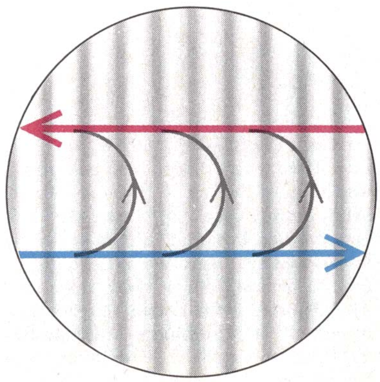
The symmetry in the arrangement of the reference beams suggests that the interference pattern could just as well be created with the second reference beam and the object beam and be read out with the first reference beam. In reality two interference patterns are formed in the nonlinear medium: each reference beam generates a pattern from which the other reference beam is reflected.

What we have just described is, in fact, the recording and reading out of a "dynamic" hologram. A hologram is an interference pattern, made with laser beams and stored in photographic film, that enables one to produce a three-dimensional image. Traditional "static" holography consists of the following three distinct steps: a hologram is recorded by illuminating a photographic transparency with a two-wave interference pattern resulting from an object beam and a reference beam; the film is developed, and the hologram is

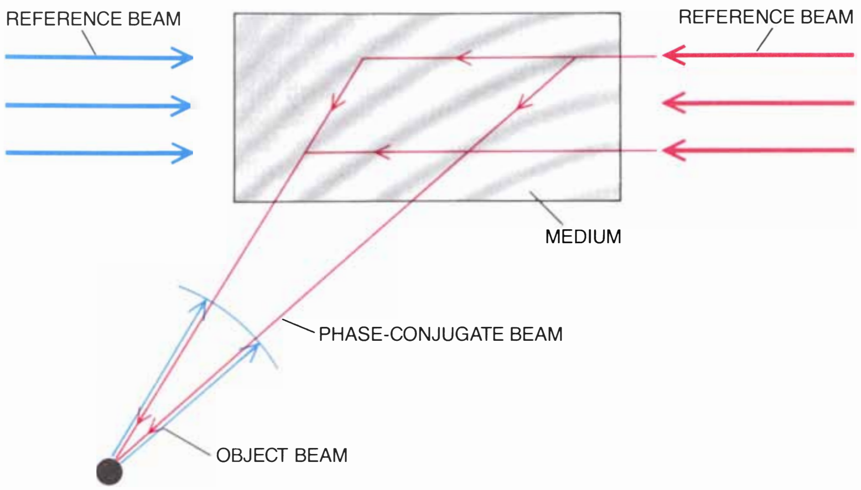


PHOTOGRAPH OF A LASER BEAM suggests the reversibility of light waves: on the basis of the image alone it is impossible to tell whether the beam travels from the left of the page to the right or from the right to the left. (The direction is from left to right.) The series of vertical dark bands resulted from the interference of the laser beam with a "reference" beam; they indicate wave-front surfaces, or places of synchronous oscillation. The wave-front surfaces of two mutually conjugated waves (in this hypothetical scenario the leftward-traveling wave and the rightward-traveling wave) are reversed with respect to the directions of propagation (*right*).

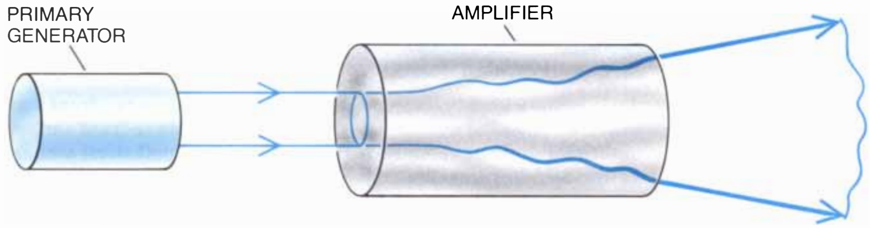


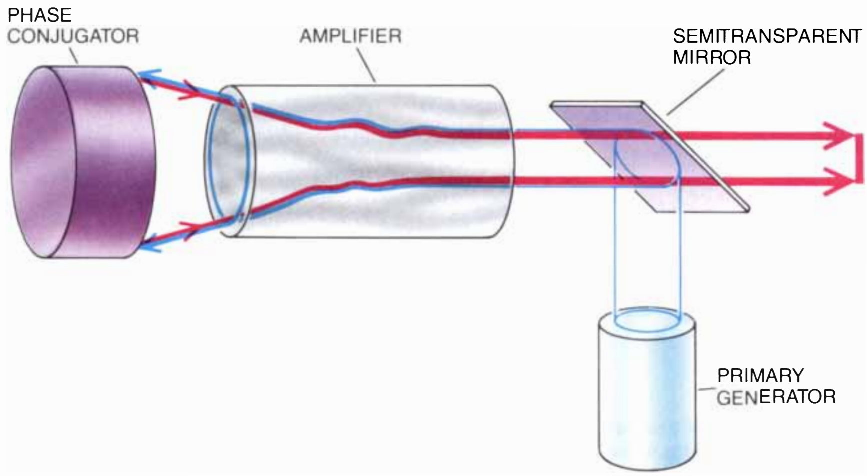


STIMULATED BRILLOUIN SCATTERING is one method of generating phase-conjugate light beams. A powerful, highly directed beam of light (*blue*) is distorted by transmission through a glass plate of inhomogeneous thickness. The distorted beam enters a transparent medium such as crystal, glass, liquid or compressed gas. It scatters from small, statistically distributed density fluctuations (thermally fluctuating sound waves) at the far end of the medium, giving rise to waves of varying spatial configurations (*curved red fragments*). When the frequency and direction of a scattered wave are just right, the wave will interfere with the incident beam and generate more pressure-density variations in the medium (*gray bars in magnified region at right*). The variations subsequently lead to the reflection of a minute portion of the incident beam. The reflected portion interferes in turn with the incident beam, generating more pressure-density variations. The pressure-density variations generate more reflections of the incident beam. The reflections build exponentially until a powerful phase-conjugate beam (*red*) emerges from the material. The high quality of the beam is restored on backward transmission of the conjugate beam through the glass plate.



FOUR-WAVE MIXING is another way of producing phase-conjugate beams. It involves the interference of four light beams in a crystal, glass, liquid or compressed-gas medium. Three of the beams are input beams: one is the object beam (*shown here as a spherical wave*) whose phase conjugate is sought and the other two are reference beams. The fourth beam, an output beam, is the desired phase conjugate of the object beam. The interaction of the object beam with one of the reference beams (*blue*) produces an interference pattern in the medium. The second reference beam (*red*) is reflected by the interference pattern. Since the second reference beam comes from a direction opposite to that of the first reference beam, the reflected beam is the phase conjugate of the object beam. In reality all the processes occur simultaneously. Moreover, two interference patterns are formed in the medium: each reference beam generates a pattern from which the other reference beam is reflected.





DIRECTIVITY OF LASER BEAMS is improved by employing optical phase conjugation. Most lasers designed to generate powerful beams are constructed according to the illustration at the top. A "local oscillator," or primary generator, produces a highly directed beam at the expense of output power. The power is increased by passing the beam through an amplifier. Inhomogeneities in the amplifying medium introduce distortions in the beam, however. The illustration at the bottom shows how to compensate for such distortions. Light from a primary generator is reflected by a semitransparent mirror through an amplifier. The distorted beam emerging from the amplifier is phase-conjugated and sent back through the amplifier. The phase-conjugate beam propagates backward with respect to the original beam through the same inhomogeneities of the laser medium, so that its motion is reversed. The resulting "double-passed" beam is both highly powerful and free of distortions.

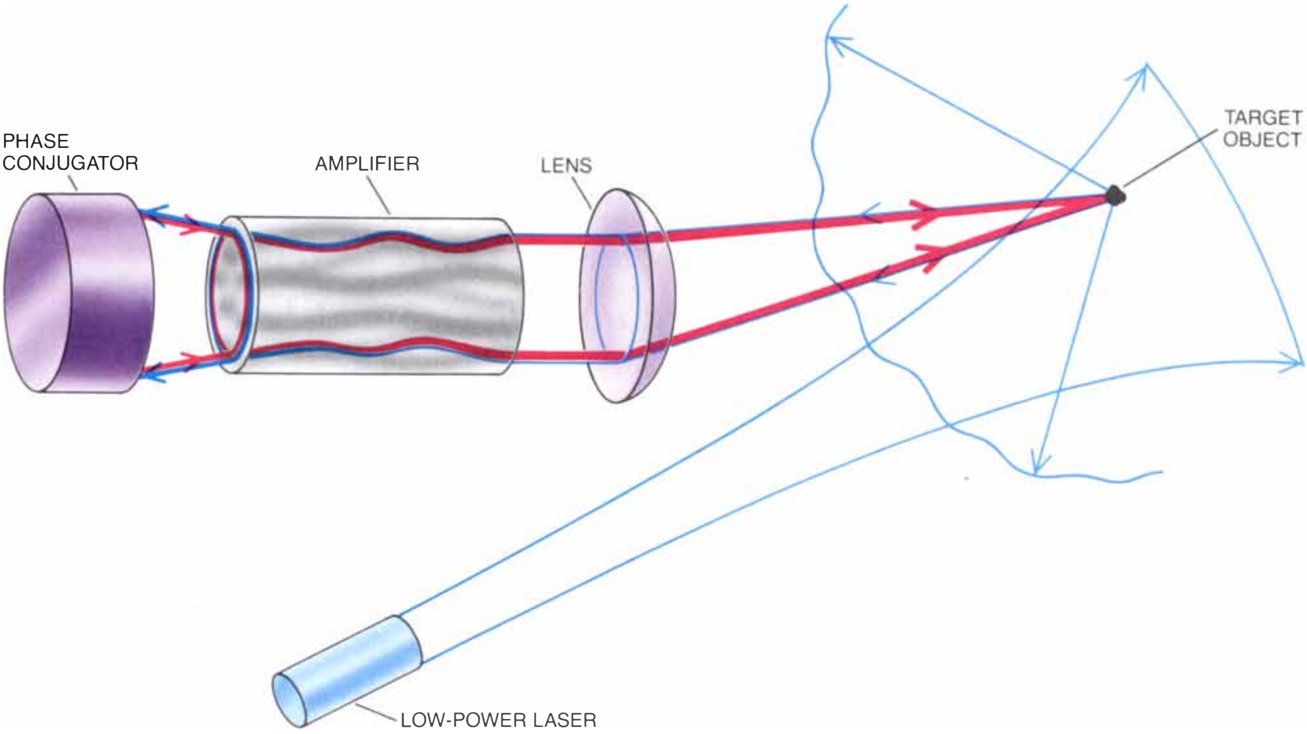
read out with another, or the same, reference beam. Four-wave mixing is an example of dynamic holography because all three processes—recording, developing and reading out—occur simultaneously, the induced variations in the refractive index disappear after the illuminating radiation is switched off, and the hologram is continually changing in response to variations in the object beam.

It is no surprise, therefore, that the idea of four-wave conjugation can be found in the work of the early pioneers of holography such as Dennis Gabor, Yury N. Denisyuk, Emmett N. Leith and Juris Upatnieks. In 1965 Herwig W. Kogelnik of the Bell Telephone Laboratories suggested that the three processes of static holography could be combined in time to yield a dynamic hologram and thereby cause phase conjugation. In 1971 Boris I. Stepanov, Evgeny I. Ivakin and Alexander S. Rubanov of the Institute of Physics in Minsk and J. P. Woerdman of the Philips Research Laboratories in the Netherlands conducted the first tests of dynamic holography employing counterpropagating reference beams. Hellwarth initiated the subsequent intensive study of four-wave mixing. He also successfully described the phenomenon in terms of nonlinear optics. Other investigators who have played an important role in the development of four-wave conjugation include Amnon Yariv and David M. Pepper, then at the California Institute of Technology, and David M. Bloom and Paul F. Liao of AT&T Bell Laboratories.

Many laboratories throughout the world have successfully made use of four-wave conjugation. A particularly attractive feature of the method is that, unlike phase conjugation by stimulated Brillouin scattering, no threshold power of the object wave is required for conjugation to occur. A number of other interesting methods of phase conjugation are also being studied.

The possible applications of optical phase conjugation are many. Two of the early ones are the production of highly directed laser beams and the self-targeting of radiation. Self-targeting provides a way of heating a small object, whose area may be as small as a millionth of the cross-sectional area of the laser beam used in the method, without recourse to a complicated system of lenses, mirrors and other optical elements. In the future the technique could possibly be exploited to heat a dense plasma and thereby initiate thermonuclear fusion.

Almost all laser setups designed to generate powerful laser beams are built in accordance with the following



SELF-TARGETING OF RADIATION is another application of optical phase conjugation. The technique, which provides a way of heating a small object without recourse to a complicated system of lenses and mirrors, could possibly be exploited to heat a dense plasma and thereby initiate thermonuclear fusion. A broad beam of low-powered light is directed to the vicinity of the object to be heated. The object scatters the radiation in all directions. Some of that radiation passes through an optical lens and strikes an amplifier placed near the object. As the radiation travels through the amplifier its power is magnified. A phase conjugator placed at the end of the amplifier creates an "antidistorted" beam and reflects it back through the amplifier, directing a powerful laser beam on the object.

scheme. First, one builds a so-called primary generator ("local oscillator") that produces a highly directed beam at the expense of output power. To obtain the desired power one then passes the beam through an amplifier. The amplifier consists of a solid or gas of highly excited molecules. As the primary beam travels through the amplifier, it stimulates the molecules and causes them to release their energy as radiation.

An ideally homogeneous amplifier would not distort the directivity of the beam. The constancy of the refractive index of such an amplifier would have to be better than one part per million, however. One can hardly hope to obtain such a high degree of homogeneity, particularly under the conditions of intense excitation of the amplifying medium. The refractive index of glass, for instance, changes by one part per million when the temperature changes by a thirtieth of one degree Celsius.

Fortunately distortions in the directivity of the beam introduced by the amplifier can be corrected by a "double pass" scheme utilizing optical phase conjugation. The idea was suggested and experimentally tested in 1972 by Oleg Yu. Nosatch and Ragul'skii and their colleagues at the Lebedev Physical Institute. They generated an ideally directed beam with a primary pulsed laser made of ruby crystal. The beam then passed through a ruby amplifier; the resulting beam, which was powerful but distorted, was subjected to optical phase conjugation. Finally, they sent the phase-conjugate beam back through the amplifier. They found that during the second pass through the amplifier the beam consumes almost all the energy stored in the excited molecules of the ruby.

The investigators also found something truly remarkable, namely that after the phase-conjugate beam passes through the amplifier it becomes ideally directed. The explanation of the phenomenon lies in the fact that the phase-conjugate beam propagates backward with respect to the original beam through the same inhomogeneities of the laser medium and therefore "reverses" its motion. The phase-conjugate beam compensates not only for static inhomogeneities due to optical elements but also for dynamic inhomogeneities. The reason is that the time it takes light to travel through an amplifier several meters long, roughly a hundred-millionth of a second, is much less than the time required for the excitation and relaxation of optical inhomogeneities in a laser.

Optical phase conjugation has also proved useful in the self-targeting of radiation. The idea was first suggested by Kogelnik. There are several possible ways to carry out the method; we shall describe only a single example. One begins by directing a broad beam from a relatively low-power laser in the vicinity of the object one wishes to heat. The object scatters the radiation in all directions; some of that radiation, after passing through optical lenses, falls on an amplifier placed near the object. As the radiation travels through the amplifier its power is magnified. A phase conjugator placed at the end of the amplifier creates an "antidistorted" beam and reflects it back through the amplifier. The result is that a powerful, directed beam is focused on the object. The focusing of the beam on the target is limited only by the wave nature of light (that is, diffraction effects) and is independent of the orientation of a focusing system. The target appears to "attract" the amplified radiation. Nikolay G. Basov and his colleagues at the Lebedev Physical Institute are studying the feasibility of facilitating laser fusion by self-targeting.

Investigators throughout the world are now exploring numerous other uses of optical phase conjugation. At present the number of applications would seem to be limited only by imagination.

Cricket Auditory Communication

The female's ability to recognize the male's calling song and to seek out the source of the song can be used to study how nervous-system activity underlies animal behavior

by Franz Huber and John Thorson

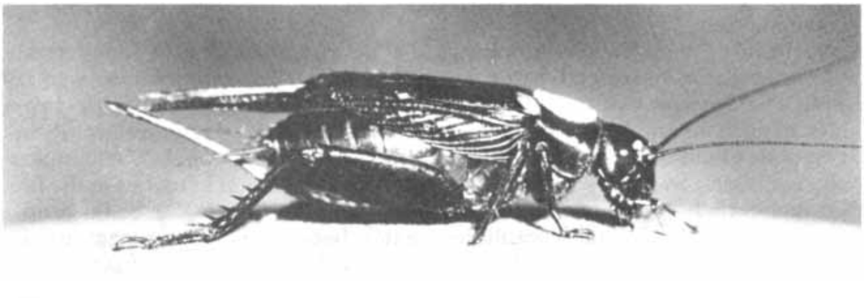
The male cricket sings by scraping his wings, and the female tracks him down. Traces of this interpretation of cricket mating behavior go back to antiquity, but it was not until 1913 that Johann Regan, a high school teacher in Vienna, tested it experimentally. He arranged for a male to sing to a female over the recently developed telephone. The experiment succeeded: when the chirps of the male were broadcast, the female approached the telephone earpiece. Nonauditory stimuli such as chemical signals could therefore be ruled out as cues for her response.

With more modern electronic and physiological methods one's questions can go deeper. What features of the male's song cause the female to seek its source? How do the neurons, or nerve cells, in the female's central nervous system distinguish the song from other songs and sounds? Given the curious arrangement of the cricket's auditory receptors—the cricket's ears are situated below the knees—how does the female determine the direction of the song? More than a dozen laboratories around the world are now attempting to find the answers. Here we shall try to convey the flavor of some of the recent progress.

The answers are well worth the effort. A central aspect of the neurosciences is the attempt to explain the behavior of an animal in terms of the operation and interactions of individual neurons. Often such work is done with invertebrates to take advantage of their stereotyped behavior, the relative paucity of their nerve cells and the fact that many of the largest neurons can now be identified readily in all members of a species, much as a particular tooth is readily identified in person after person. So far the greatest successes have come on the output, or motor, side of neuronal organization. For example, investigators have established that complex bodily movements are orchestrated by particular sets of neurons in leeches, arthropods and mollusks. Neuronal events at high levels on the input, or sensory, side have been more elusive. One such event is a subtle "decision": the neuronal recognition of the cricket calling song and the initiation of the tracking behavior of the cricket in the direction of the source of the song.

It is precisely this "decision" that the work we shall describe here is intended to probe. First we shall summarize the male field cricket's song and what is known about the cricket's auditory organs. Then we shall focus on behavioral and neurophysiological experiments that begin to suggest in detail what happens in the brain of the female when the song is being recognized. Much of the work was done at the Max Planck Institute for Behavioral Physiology in Seewiesen, West Germany. It was done in close collaboration with our colleagues, in particular Theo Weber, Hans-Ulrich Kleindienst, Klaus Schildberger, David W. Wohlers, Dietmar Otto, George Boyan, Eckehart Eibl, Harald Esch and Leslie Williams.

The calling songs we study are those of the chirping crickets, which include European field crickets such as *Gryllus campestris* and *G. bimaculatus*. In males of these species each of the two front wings includes what amount to a file and a scraper, so that when the wings are closed, and hence rub together, the wings are set briefly into oscillation at a frequency of about five kilohertz (5,000 cycles per second). The result is that each time the wings close there is a pulse of almost pure five-kilohertz sound, often called a syllable, lasting for from 15 to 20 milliseconds. Heard close up, it can be painfully loud. The reopening of the wings is silent, but subsequent closings and the syllables generated by each of



SPHERICAL TREADMILL makes it possible for a walking female cricket (*above*) to have a free choice of speed and direction while nonetheless being kept at a fixed distance from loudspeakers emitting various test sounds. The cricket walks on top of a sphere 50 centimeters in diameter mounted in an anechoic chamber (*right*), with a disk of reflective foil stuck to her back. The foil reflects infrared light from an overhead source to photodetectors that sense her position and send corrective signals to motors driving the bearings of the sphere; the motors counterrotate the sphere to keep the cricket near its top. The corrective signals yield continuous records of her speed and intended direction of motion. Loudspeakers broadcast sounds (calling songs of the male cricket or specially altered "songs" synthesized by a computer) from particular directions within the chamber. As she walks, the female cricket expresses her apparent recognition or rejection of the features of such test sounds by tracking or by ignoring the loudspeakers from which the test sounds are emitted. The photographs were made by Theo Weber in the laboratory of one of the authors (Huber) at the Max Planck Institute for Behavioral Physiology in Seewiesen, West Germany.

them follow at intervals of about 35 milliseconds, so that the syllable-production rate is about 30 syllables per second. The commonest kind of chirp emitted by the male cricket is a train of four syllables, which is followed by a brief silence. The chirps are usually repeated at a rate of from two to four chirps per second.

How does the female cricket hear the chirps? In broad outline the auditory apparatus resembles the human arrangement: a sound-transduction system including tympana, or eardrums, excites an array of auditory receptor cells, which in turn excite auditory neurons. In the cricket, however, the ears are below the knee of the foreleg on each side of the body. There a pair of tympana on the surface of each tibia overlie an array of from 55 to 60 auditory receptor cells.

The axons, or nerve fibers, emerging from the cells run up the leg as a bundle, the auditory nerve. Their destination is the prothoracic ganglion, in the central nervous system. Recordings of the electrical activity carried by the nerve (or of the activity of the prothoracic neurons receiving signals from the nerve) establish that the auditory receptor cells respond to airborne sounds ranging from about three kilohertz to what would be, for human hearing, ultrasonic frequencies. Evidently a major population of the receptors responds best at about five kilohertz, the carrier frequency of the male calling song.

Remarkably, the ear on one side of the body is coupled to the ear on the other side by the lower branches of an air-filled tracheal tube that no longer serves only for respiration. Each of the two upper branches ends in a spiracle, a portal to the air surrounding the cricket, on each side of the body. As a result sound pressure reaches each tympanum both directly (at the outside surface of the tympanum) and indirectly (at its inside surface, by way of the air-filled tube). These relations are hard to examine by direct measurement: the small size of the tracheal tube (it ranges from .1 to .3 millimeter in diameter) means one must reason by inference. Still, many studies, in particular the laser-vibrometry measurements made by Axel Michelsen and Ole Naesbye-Larsen at the University of Odense in Denmark, have revealed some of the properties of the arrangement. In such measurements a surface (in this case a tympanum on the foreleg of a cricket) is illuminated by a laser and the velocity of the surface is calculated from the Doppler shift of the frequency of the light that is reflected from the surface.

The results are somewhat surprising. For example, five-kilohertz sound entering, say, the left spiracle proves to be more efficient at deflecting the left tympana (by back pressure from within the tracheal tube) than five-kilohertz sound of the same intensity broadcast just outside the tympana. Apparently the tracheal tube has a resonance near five kilohertz. Of what value is this complex system of tubes? As a female seeks out a singing male, she characteristically follows a zigzag path. It is as if she were obeying a simple rule: turn toward the ear currently receiving the loudest sound.

The problem the cricket faces is that her forelegs, and hence her ears, are only about a centimeter apart, whereas the wavelength of a five-kilohertz tone is about seven centimeters. Under those circumstances the difference in intensity between the sounds arriving from any one source at the outer surface of each ear is at most a few decibels, which is not enough to guide the

female's search for a singing male. The tube system evidently ensures that the sound differences at the two ears are much enhanced by wave phenomena of cancellation and reinforcement due to travel times, resonances and phase shifts in the tube. In sum, it appears most likely that the tube system, with its resonance near the carrier frequency of the male calling song, evolved as an aid to directional hearing.

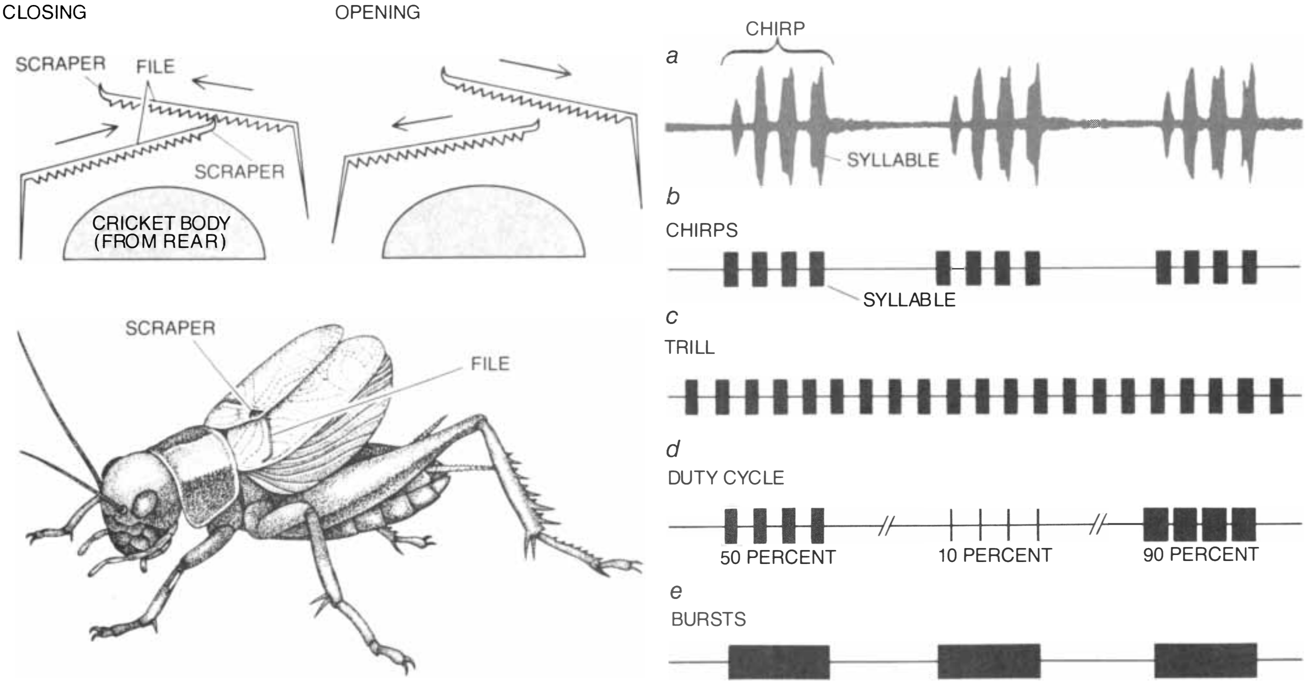
The understanding of the contribution the tracheal tubes make to a cricket's hearing remains incomplete. Until recently, for instance, it was not known which of two distinct phenomena, tympanal deformation or the internal sound pressure in the tracheal tube, actually excites the auditory receptors. (Those quantities can vary grossly from each other in amplitude and phase as the frequency and angle of incidence of the sound change.) The question has now been clarified by Kleindienst and Wohlers at Seewiesen, in collaboration with Naesbye-Larsen at Odense. They separated the two variables by means of sound-cancellation experiments.

A cricket's foreleg is encased in a small, soundproof broadcast chamber called a legphone, so that a sound can be made to impinge exclusively on the tympana of that leg. Within the cricket's body the sound gets pumped through the tracheal tube and thus to the tympana on the opposite leg. Meanwhile sound of adjustable amplitude and phase is broadcast externally to this opposite leg. A laser vibrometer monitors the motion of the posterior tympanum, while a microelectrode monitors the excitation of the auditory receptors by recording the activity of an auditory neuron in the prothoracic ganglion [*see illustration on page 64*].

The idea is to adjust external pulses of five-kilohertz tone so that when matched internal and external pulses are both delivered, the vibrometer detects no tympanal motion. Under those circumstances the findings are unambiguous. When pulses are delivered only internally or only externally, the tympanum moves, the auditory receptors fire volleys of nerve impulses (spikelike electrical signals in the nerve fiber) and the auditory neurons respond by producing several impulses of their own. When, however, the internal and external pulses coincide, so that the tympanum is motionless, the auditory neurons are silent. Yet the internal tracheal sound pressure at such a time can be as great as 80 decibels. Evidently it is tympanal deformation and not local tracheal sound pressure that activates the receptors; direct mechanical effects are presumably implicated in the transduction process.

Behavioral scientists generally study crickets and their auditory abilities by observing what the female does when various sounds are played to her over loudspeakers. In the classical experiment the female is released in a Y-shaped maze or on a flat surface called an arena, and loudspeakers broadcast selected sounds from various directions. One difficulty with such experiments is that the female must be recaptured and repositioned after each walking trial. Moreover, when an animal is to choose between sources of sound, a chance start toward one of the sources makes that source louder simply because it is nearer than it was before the animal moved.

These difficulties are avoided if one arranges to keep the female in place during the tests, on top of a sphere that rotates to compensate for her movements. The cricket can then have a free choice of speed and direction of walking and yet remain more or less stationary with respect to sources of sound. The cricket can be tested continuously, often for hours; extensive sequences of sounds can be presented, including some trials to test the reproducibility of her responses and other

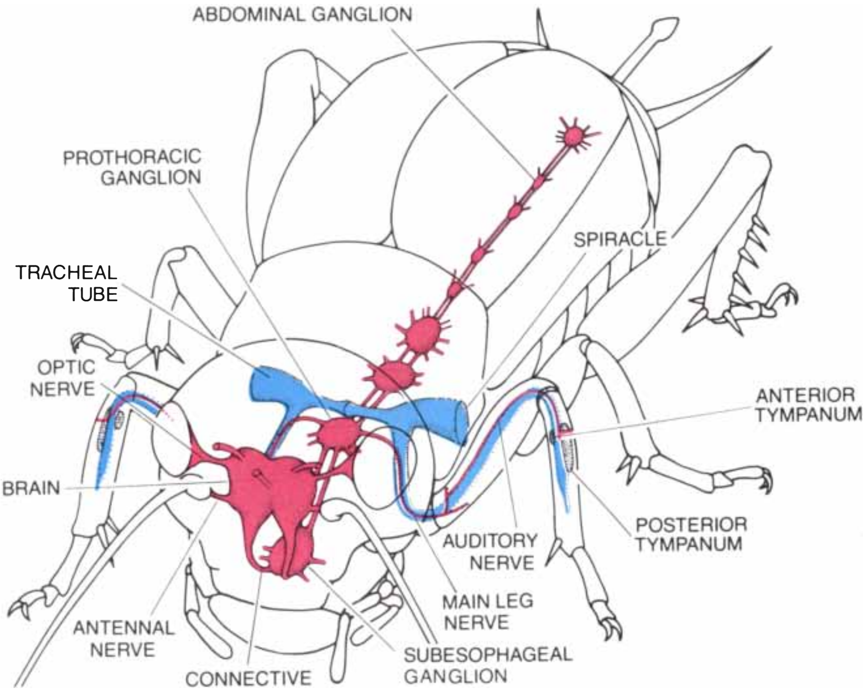


CRICKET SONG is produced by the male when he rubs his wings together. The highly schematic diagrams at the upper left show the sound-making mechanism: each wing incorporates a scraper and a file, so that every time the wings close they are set in brief vibration at a frequency of about five kilohertz (5,000 cycles per second). The diagram at the lower left shows how the wings of the cricket are raised and scraped together when the sound is made. The song (*a*) is therefore made up of "syllables" of five-kilohertz tone. Typically there are four syllables per chirp, and the syllables come at a rate of 30 per second. The temporal patterns of some computer-synthesized songs are schematized below the temporal pattern of the natural song. One synthesized song (*b*) reproduces the principal features of the natural song. Another (*c*) is a "trill": it fails to divide the song into chirps. Others (*d*) preserve the chirps and syllable rate but expand or contract the duration of each syllable, thus varying the duty cycle, or ratio of the sound to the silence within a chirp. Finally, one test pattern (*e*) is made up of "bursts" of five-kilohertz tone of the correct chirp duration but not separated into syllables.

trials to check on whether the order in which sounds are presented is affecting the results. (Methods for studying hearing in tethered flying crickets, developed by Andrew Moiseff, Ronald R. Hoy and Gerald S. Pollack of Cornell University, offer some of these advantages.)

Our device, which is essentially a spherical treadmill, was designed by Ernst Kramer and Peter Heinecke at Seewiesen and is now in use in several laboratories. The cricket walks freely on top of a plastic sphere 50 centimeters in diameter. A small disk of light-reflecting foil is stuck to her back, and infrared light at 963 nanometers, a wavelength invisible to the cricket, arrives from an overhead source. When the cricket moves away from the precise top of the sphere, photodetectors sense her motion by means of the reflected light and motors rotate the sphere so as to compensate for her motion and keep her near the top.

For the cricket, of course, walking on the treadmill is not quite the same as natural walking: the corrective accelerations imparted to the sphere unavoidably alter the inertial forces the cricket may detect by way of sensory receptors within her legs. Still, we can adjust the responses of the motors to the commands they receive in such a way that the cricket gets no farther than about half of her body length from the very top of the sphere, and yet her motions do not look awkward. By keeping track of the position and velocity signals sent from the photodetectors to the motors we gain a continuous record of the direction in which the cricket is trying to walk and also of her walking speed.



TRACHEAL-TUBE ARRANGEMENT (*blue*) dominates the cricket's peripheral auditory system. It connects the two "ears," which are curious structures below the knee of each foreleg. On each side of the body a large, funnel-shaped upper branch of the roughly H-shaped tube leads to an opening called a spiracle at the body surface; the tube's lower branch descends through the foreleg to a position between the ear's two tympana, which are auditory membranes overlying an array of receptor cells. Hence sound reaches each ear not only externally, from the enviroment, but also internally, by way of sound pressure within the tracheal tube. The cricket's central nervous system (*red*) consists of a sequence of ganglia (aggregates of neurons), which are linked by pairs of connectives, or bundles of the fibers sent out by the neurons. The frontmost ganglion is the brain, with its optic and antennal nerves. Next comes the subesophageal ganglion, followed by three thoracic ganglia and several abdominal ones. Sensory signals from the array of receptor cells under the auditory tympana enter the prothoracic ganglion, which is the frontmost thoracic ganglion.

The behavior of female crickets on the spherical treadmill is often quite unambiguous: it routinely gives us yes or no answers to questions about her proclivities. For example, when computer-simulated chirps are played from loudspeakers at various positions around the sphere, the female tends to respond in a characteristic way. In a series of short episodes of walking she tracks the loudspeaker emitting the sound. Her accuracy increases as the intensity of the sound is raised from 50 to 70 decibels. When we switch speakers, she switches direction appropriately. (The experiment is conducted in an anechoic chamber so that the only effective sounds reaching the cricket are the ones coming directly from the speakers.) On the other hand, most females ignore bursts of five-kilohertz tone that last for as long as a chirp would last but are not divided into the sequence of syllables that constitute a natural chirp.

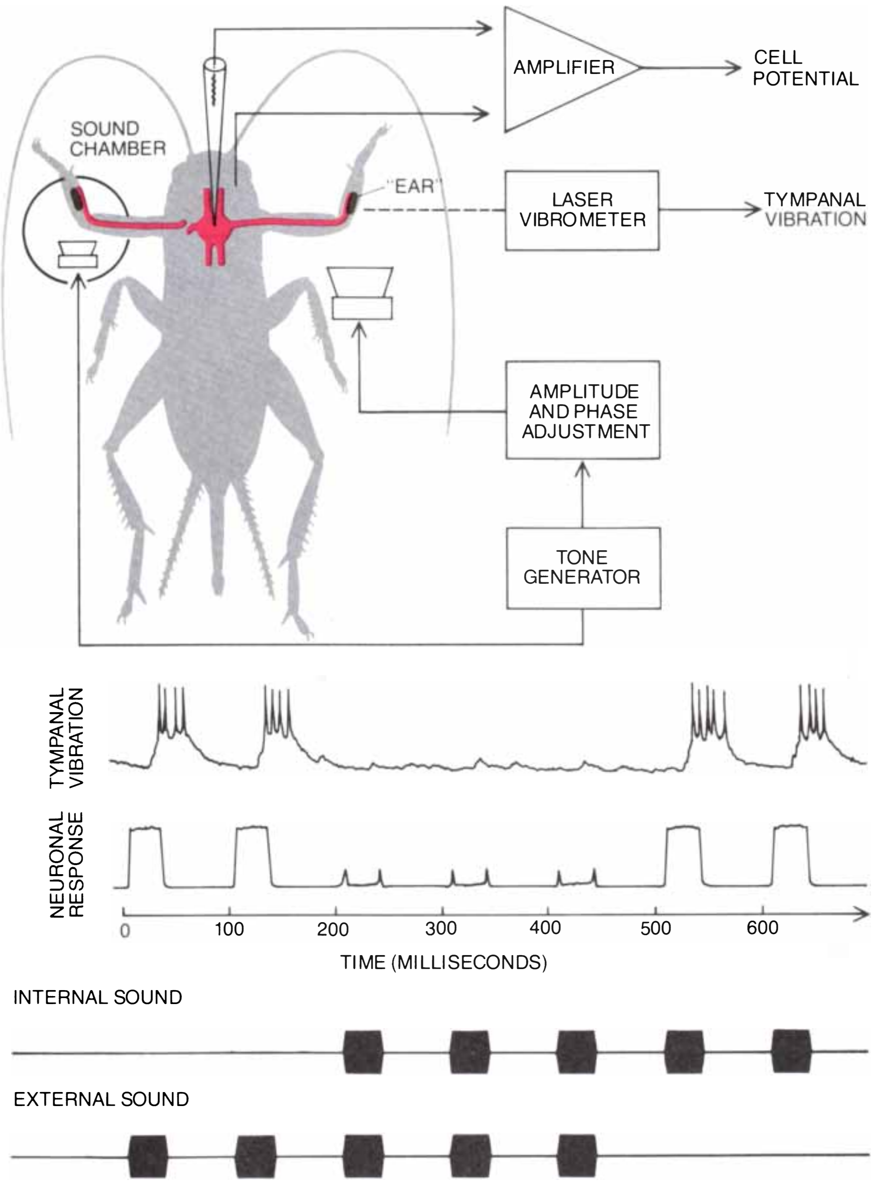
Given this powerful means of assaying the female cricket's responses to sounds, we have asked in detail what features of the temporal pattern of the male calling song appear to be critical for initiating the female's tracking of the song. Candidate features include the chirp rate, the syllable rate and the number of syllables per chirp. Earlier studies had already suggested that the syllable rate is important. Guided in part by those studies, we began with the working hypothesis that the natural syllable rate—approximately 30 per second—is the only feature the female needs in order to recognize the song and begin to seek its source.

In the case of the European field cricket *Gryllus campestris,* the working hypothesis survives a surprising number of tests. For example, the females ignored single-syllable chirps but tracked chirps of three or more syllables when the syllables were presented at the 30-per-second rate. In other words, the syllable rate is crucial; the number of syllables is not.

In a further set of experiments we varied the ratio of tone to silence in each chirp. In a natural chirp the syllables take up about half the time of the chirp, for a "duty cycle" of 50 percent. By synthesizing chirps consisting of four brief pulses of five-kilohertz tone at a 30-per-second rate, we reduced the duty cycle to 10 percent; by synthesizing chirps consisting of four long pulses of tone at a 30-per-second rate, we increased the duty cycle to 90 percent. The female tracked both kinds of synthesized songs. Again, the syllable rate is evidently crucial; the duration of each syllable is not. Test songs with such departures from the natural duty cycle tended, however, to require presentation at greater volume than natural songs require, as expected from the reduction in 30-hertz power that accompanies unnatural duty cycles. Finally, we synthesized continuous "trills" of five-kilohertz syllables at the 30-per-second rate, with no subdivision into chirps. Although earlier reports had suggested that trills are not attractive to female crickets unless the crickets have first been primed by hearing normal chirps, we found that a majority of the females tracked a trill even when it was the first song of the day. The 30-per-second syllable rate appears to be the principal cue.

Our pursuit of the 30-hertz recognition hypothesis has been in the spirit of Alfred North Whitehead's celebrated advice, "Seek simplicity; and distrust it." And indeed there are interesting complications. A minority of the females we tested would not track trills at all but began to track immediately if the trill was interrupted periodically, giving rise to chirps. Moreover, John Doherty, working with our treadmill at Seewiesen, has examined the behavior of female crickets of the species *G. bimaculatus,* which is closely related to *G. campestris.* Although some of the females track trills, a minority will track asyllabic bursts of five-kilohertz sound—a clear refutation of our working hypothesis that the syllable rate of 30 per second is the sole crucial feature of the male calling song.

It is not yet clear whether such findings are occasional glimpses of a complex recognition process or simply reflect an abnormal lack of sensory selectivity within the central nervous system of certain individual crickets. Such differences do, however, highlight an important procedural point: When one "gets to know" individual females during long runs on the treadmill in many experimental conditions, consistent differences among individuals—even in qualitative ability—are found. These are not always discerned in the usual statistical summaries of such experiments.

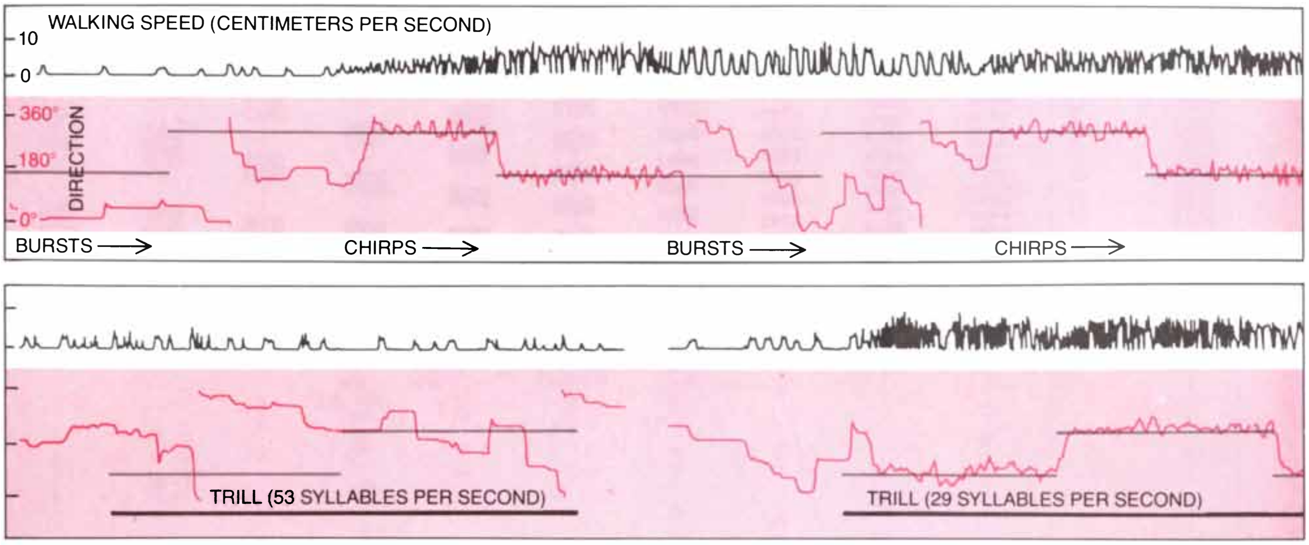


SOUND-CANCELLATION EXPERIMENT developed by Hans-Ulrich Kleindienst at Seewiesen clarified the role of the tympana in sound transduction. One of the cricket's forelegs is enclosed in the small chamber shown at the left in the diagram. Brief five-kilohertz tones are delivered within the chamber; the sound propagates through the tracheal tube and excites receptors on the opposite leg by exerting internal pressure on the tympana. External sound is delivered simultaneously to the leg at the right in the diagram. The motion of the posterior tympanum on the latter leg is monitored by a laser vibrometer. The phase and amplitude of the external sound can be adjusted in such a way that it cancels the effect of the internal sound, whereupon the tympanal motion ceases. Auditory signals from the ear on the leg in the chamber are eliminated by cutting the auditory nerve. An intracellular microelectrode measures the response of an auditory interneuron in the prothoracic ganglion. The results of one experiment are shown. When either internal or external sound is delivered to the ear on the right, the posterior tympanum moves and the neuron is excited. When tympanal motion is canceled by presenting the two sounds together, the neuron is not excited, even though the internal tracheal sound pressure can be substantial. Thus the experiment shows that tympanal deformation is critical for the transduction of sound; the sound pressure inside the trachea does not suffice to activate the receptor cells in the ear.

It is clear, in any event, that the modulation of five-kilohertz sound at the characteristic natural syllable rate is the most reliable way to call a female cricket. In the next stage of our research we undertook, again in collaboration with Weber, to capitalize on this finding by designing a stimulus program for a search for song-recognizer neurons in the central nervous system of the female. The fundamental problem in such a search is that most of the auditory neurons in the central nervous system of the female cricket respond not only to the male calling song but also to sounds we know the female does not track.

One must therefore test each candidate neuron (during the period, sometimes frustratingly brief, when a microelectrode is in place and is recording the activity of the neuron) with a range of songs including some that elicit tracking and others that do not. To this end we employ a sequence of computer-synthesized songs that systematically varies the syllable rate while keeping the duration, energy and repetition rate of the chirps nearly constant. The female's behavior when she is stimulated with this sequence amounts to a band-pass response: for syllable rates near 30 per second she tracks almost flawlessly, whereas for rates of less than 20 per second or more than 40 per second her performance worsens dramatically.

The next question is how the central auditory neurons respond to such sequences of correct and incorrect songs. Here a description of some of the neuroanatomy of the cricket is essential. The auditory nerve runs through the knee joint of the foreleg from the tibia into the femur. There it joins the main leg nerve. Nevertheless, its constituent axons continue to form a distinct, com-



TRACKING OF SYNTHESIZED SONG by a female cricket on the treadmill is demonstrated in typical recordings of speed and direction. For each of two experiments the upper trace (*black*) gives the female's walking speed. When she tracks, she usually follows a zigzag path, sometimes pausing and sometimes walking as fast as six centimeters per second. There are two sets of lower traces (*in colored bands*): horizontal line segments (*gray*) show the direction of the sound source, which is changed by switching from one loudspeaker to another; wiggly curves (*dark color*) show the direction in which the female is walking. In one experiment (*top*) trains of bursts of five-kilohertz tone were alternated at about two-minute intervals with trains of four-syllable chirps simulating natural song. The female ignored the bursts but tracked the source of the chirps avidly, changing direction appropriately when the sound direction changed. In another experiment (*bottom*) two trills were presented, one at a syllable rate of 53 per second and the other at a rate of 29 per second that is found in nature. The female ignored the "wrong," 53-per-second trill but tracked the source of the 29-per-second trill.

pact bundle, distinguishable from the rest of the nerve, until they enter the prothoracic ganglion, a neuronal complex that is (among other things) the first central processing station for auditory input.

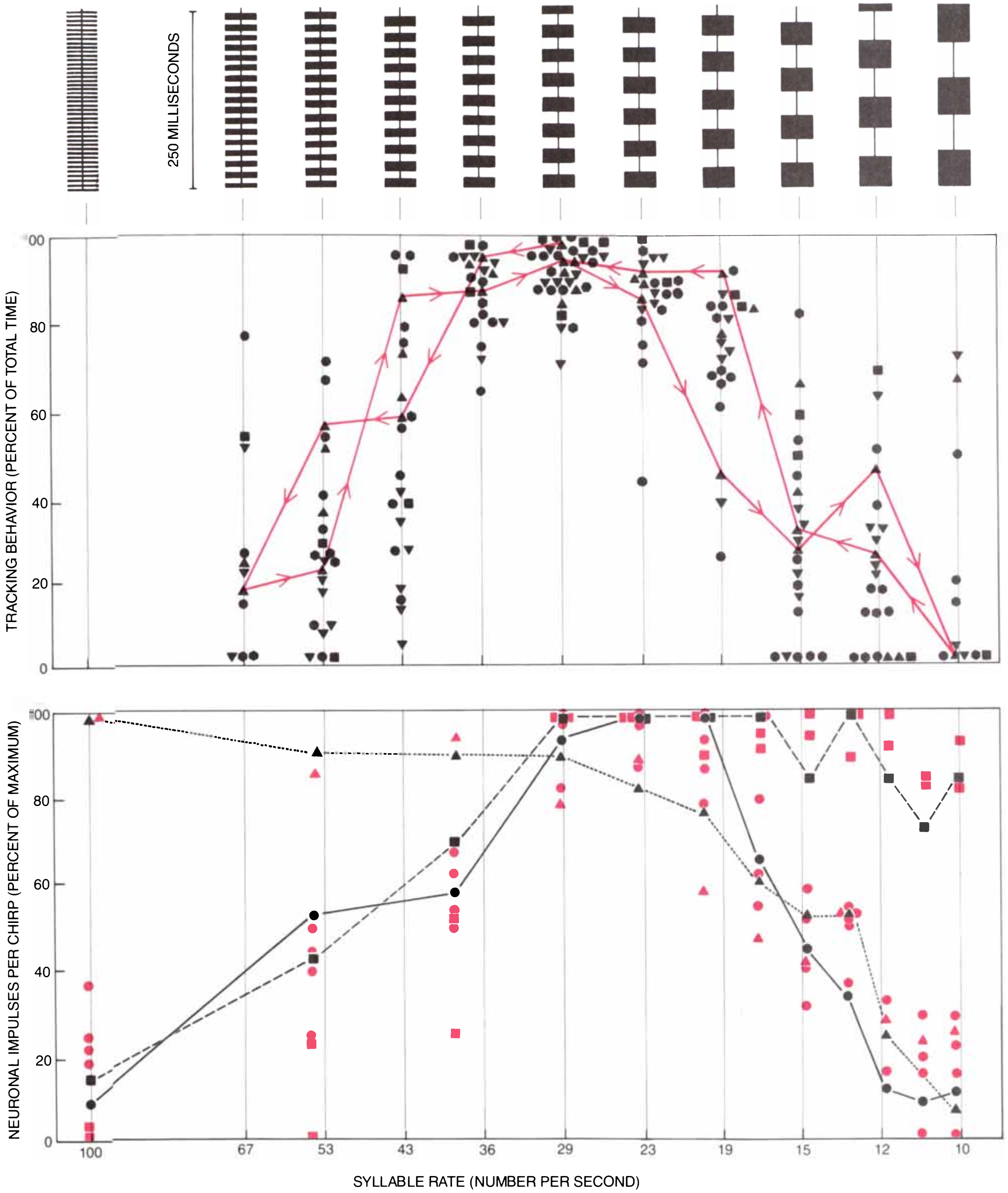
To determine the precise field of distribution of the auditory fibers one cuts the auditory nerve and exposes the cut to a cobalt salt, which travels up the fibers. The precipitation of the salt then reveals that the fibers branch and terminate inside the ganglion on the same side as the leg from which they come, without crossing the midline. The restricted region of the ganglion in which they terminate is known as the auditory neuropil. There the terminal branches of the fibers make synapses with interneurons: nerve cells that process and relay information within the central nervous system. In recent years a number of such cells have been identified. "Identified" is in fact a technical term with a twofold meaning. First, the cells have been shown to be characterized by a particular pattern of electrical activity and a particular anatomical structure. Second, the cells have been shown to be present consistently in many individuals of a species.

Two of the most prominent identified auditory neurons in the prothoracic ganglion, studied by several investigators, are the omega-1 neurons, or ON-1 neurons, named for their resemblance in shape to the Greek letter omega (Ω). They are present as a bilateral pair, each the mirror image of the other. Each ON-1 cell confines its arborizations (the bushlike branchings of the protrusions over which a neuron dispatches or gathers signals) to the two neuropil regions of the prothoracic ganglion. Each ON-1 cell receives excitatory input only from the ear ipsilateral to (on the same side as) its cell body. That is, the ON-1 cell whose cell body occupies the right half of the prothoracic ganglion receives excitatory input only from the ear on the right foreleg. Although it responds over a broad range of frequencies, the ON-1 cell "copies" the male calling song briskly. That is, it generates a train of impulses closely mimicking the temporal pattern of the song. In less than a millisecond such trains are conducted to the other side of the ganglion along the omega-shaped trajectory of the axon of the ON-1 cell.

Curiously, stimulation of the ear contralateral to the cell body of each ON-1 cell inhibits the cell, and the inhibition of each ON-1 turns out to be exerted by its mirror-image partner—as suggested by Wohlers' work and demonstrated recently by Allen I. Selverston of the University of California at San Diego in collaboration with Kleindienst in Seewiesen. Hence the inhibition is reciprocal. In principle, therefore, the inhibition could serve to sharpen differences between auditory signals from the two ears, enhancing the sensitivity to the direction of sounds above and beyond the enhancement the tracheal tube provides. Other neuronal interactions in the prothoracic ganglion are less easy to analyze. For one thing, a second pair of omega neurons, called ON-2 cells, are closely apposed to the ON-1 cells but have a different pattern of sensory activation. Moreover, numbers of smaller neurons remain unstudied.

The output lines emerging from the prothoracic ganglion seem to have a coherent pattern—perhaps deceptively so. In particular several large, identified prothoracic neurons have axons that travel upward to the brain. Among these "ascending neurons" two prominent types, the AN-1 and AN-2 cells, occur as mirror-image pairs. The AN-1 cells are highly sensitive to the five-kilohertz calling-song carrier frequency, whereas the AN-2 cells are sensitive predominantly at higher frequencies. Unlike the omega neurons, the AN-1 cells receive input from the contralateral ears; they show no sign of receiving input from the ipsilateral ones. The dendritic (input-gathering) arborizations of each AN-1 cell overlap with the axonal (output-transmitting) arborizations of an ON-1 cell.

It is interesting that among all these identified prothoracic cells (the omega neurons and the ascending neurons) we find no candidates for neurons that recognize the male calling song. We find only neurons that copy chirps—and other sounds—and convey their temporal pattern to the brain (or, in



BEHAVIOR AND BRAIN-CELL ACTIVITY on the part of female crickets show closely related responses to synthesized songs that have varied syllable rates. Females on the treadmill heard a series of chirps (*top*) that varied the syllable rate but minimized the variation of certain other features of the song. The preference of females for syllable rates near 30 per second emerged from their tracking behavior (*middle*), which showed a band-pass pattern: a sensitivity to a particular range of syllable rates. Several female crickets (*geometric symbols*) were tested in a number of experiments; the data from one of the experiments are connected by colored arrows to show the back-and-forth sequence of the tests. In the brain of the cricket specific neurons (*bottom*) prove to respond to songs (in terms of the number of nerve impulses that are elicited by each chirp) in ways that closely parallel the female's behavioral response. Some cells have a band-pass response (*circles*), that is, they respond best at syllable rates near 30 per second. Other cells have a low-pass response (*squares*), that is, they respond best at or below 30 syllables per second. Still others have a high-pass response (*triangles*). Line segments connect the data for one of the neurons of each type. The behavioral data are from *Gryllus campestris;* the neurons were studied by Klaus Schildberger of Seewiesen in the closely related species *G. bimaculatus,* which shows behavior similar to *G. campestris.*

some cases, in the opposite direction, to lower thoracic ganglia). For these cells the presentation of a range of computer-synthesized songs that systematically varies the syllable rate elicits no preference for the rates near 30 syllables per second: the rates that lead the female to track the source of the male's song.

The search for the song-recognizing neurons moves on, therefore, to the brain. Here the end branches of the axon sent out by each of the AN-1 cells produce dense terminal arborizations (first demonstrated by Boyan and Williams), which overlap the arborizations of a population of brain interneurons (found on each side of the brain) identified by Schildberger. The interneurons are known as BNC-1 (brain neuron, class 1) cells. In turn the arborizations of the BNC-1 cells overlap the arborizations of another bilateral population of brain interneurons, also identified by Schildberger, called BNC-2 cells.

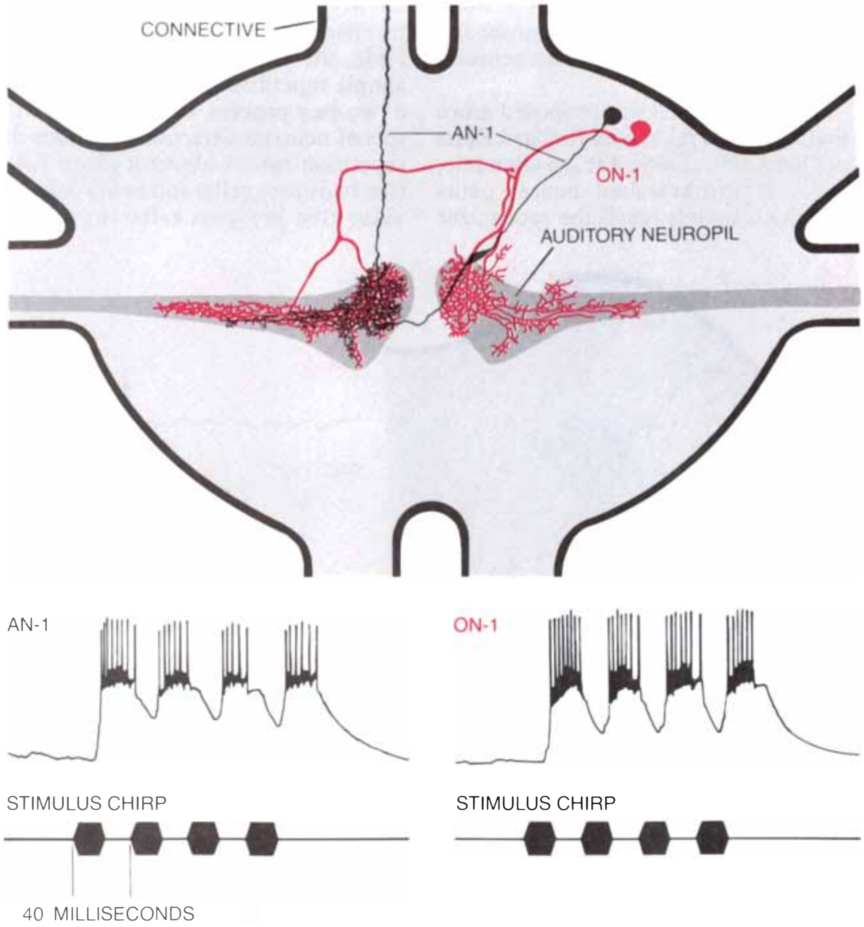
The overlaps suggest that auditory signals ascending to the brain by way of the axon of an AN-1 cell first reach the BNC-1 cells, from which they are relayed to the BNC-2 cells. In this regard it is notable that when the male calling song is presented to the female while the electrical activity of BNC-1 and BNC-2 cells is being monitored, the activity generated by the BNC-1 cells proves to copy the temporal pattern of the chirp less accurately than the neurons of the prothoracic ganglion do. The copying done by the BNC-2 cells is even less accurate: the impulses generated by them are seldom detectably synchronized with the syllables of the chirp.

The BNC cells are small, so that extended recordings of their electrical activity are hard to make. Moreover, the recordings must be followed by the injection of a dye so that the cell can be identified anatomically. Nevertheless, a picture is emerging from Schildberger's work. There is a subpopulation of BNC-2 cells in which each neuron responds (in terms of the number of impulses it generates per chirp as the syllable rate is varied) in a band-pass manner closely reminiscent of the band-pass behavior of the female. That is, the neuron responds preferentially to songs with syllable rates near 30 per second—the rates that most reliably activate the female's tracking. To our knowledge, Schildberger's finding is the first discovery of an unambiguous central neuronal correlate of a temporal pattern-recognition process in an insect.

The BNC-2 population also includes neurons that respond only to signals that have syllable rates in or above the recognition band. In brief, they are high-pass cells. Moreover, the BNC-1 population includes neurons that respond only at syllable rates in or below the recognition band. In brief, they are low-pass cells. The discoveries are a striking parallel to the band-pass, high-pass and low-pass neurons found recently among the auditory neurons in the brain of the frog by Gary Rose and Robert R. Capranica of Cornell.

What the band-pass cells among the BNC-2 population in the brain of the female cricket do when they have recognized the male calling song is still not known. The close relations between their particular sensitivities and the band of syllable rates that elicit tracking behavior encourage the hypothesis that the BNC-2 cells activate rather directly the central mechanisms responsible for tracking. Perhaps the activity of the BNC-2 cells leads in some way to release of a neuropeptide (a short amino acid chain), which serves as a chemical messenger activating the tracking behavior. Or perhaps the BNC-2 cells act by means of descending neurons, which Schildberger has also identified. In any case, Schildberger's discoveries stimulate speculation about the details of the recognition process itself.

How might a neuronal recognizer of simple temporal patterns work? Three quite different ideas have been proposed. One of them emerges from suggestions made over the past two decades by a number of investigators. They include Masakazu Konishi of the California Institute of Technology, who studied bird song; Richard D. Alexander of the University of Michigan and Hoy and his colleagues at



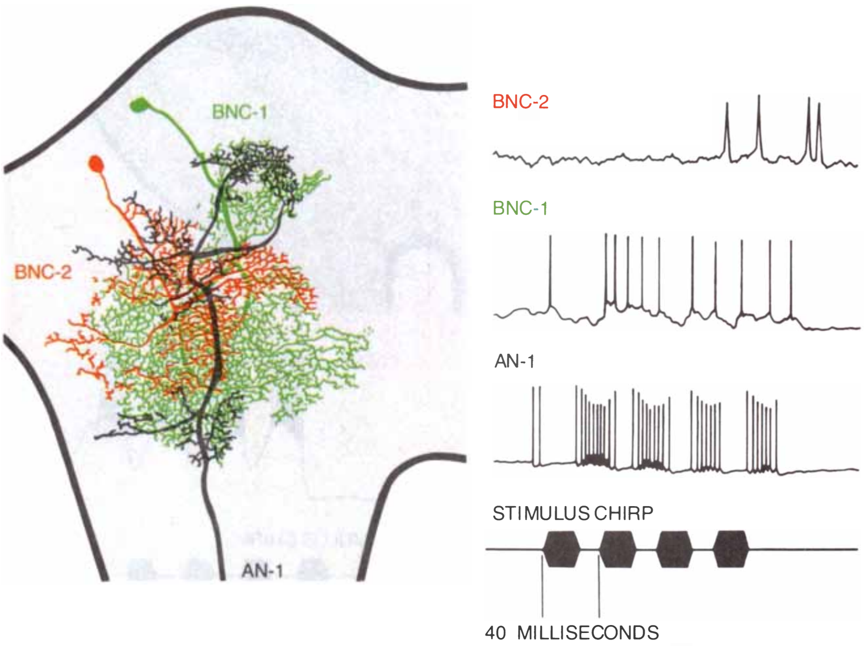
PROTHORACIC GANGLION **includes auditory neurons that constitute the first central processing station for signals arriving from receptors in the cricket ear. Fibers of the auditory nerve (*gray*) distribute their signals in a region of the ganglion called the auditory neuropil. A nerve cell called the omega neuron, or ON-1 cell (*color*), picks up the signals. Its omega-shaped axon, or nerve fiber, crosses to the auditory neuropil on the opposite side of the ganglion. There an ascending neuron, or AN-1 cell (*black*), dispatches its axon upward toward the brain. Recordings of the electrical activity of the two kinds of cell (*bottom*) demonstrate that both the ON-1 cell and the AN-1 cell copy the temporal pattern of the calling song of the male cricket. Mirror-image ON-1 and AN-1 cells (not shown) have similar structures that emanate from cell bodies on the left side of the ganglion. Cells were stained by injecting the dye Lucifer Yellow through the intracellular recording electrode.**

Cornell, who studied cricket song, and H. Carl Gerhardt, Jr., and Doherty of the University of Missouri, who studied frog calls.

In essence the idea is that the female has a built-in pattern generator that produces a template to be compared with the patterns of sensory signals arriving in the central nervous system. In the case of the cricket the idea is particularly compelling. Insects are known to have neuronal "pattern generators" for rhythmic behaviors such as walking, flying and singing. To be sure, the female cricket does not sing. Nevertheless, she shares with the male the relevant neuronal and motor apparatus; sometimes when she becomes aggressive, she moves her wings in the same way the male does when he sings. On the other hand, it is unclear how a template could be compared with an arriving male song, given the latter's arbitrary arrival time with respect to template timing. Apparently one must postulate a triggering mechanism for template generation. The idea remains speculative.

The second idea was proposed more than two decades ago by Richard Reiss of General Precision, Inc., in Glendale, Calif. It invokes dual neural paths by which signals reach the recognizer mechanism. Suppose one path delays signals more than the other path, by a predetermined amount of time. Suppose further that the recognizer requires the temporal coincidence of signals from both paths. The recognizer will then respond only to certain temporal patterns, namely the ones in which the delay matches a periodicity in the pattern. Here too the idea remains speculative. One problem is that in its strict form the recognizer accepts syllable rates that are multiples of the rate it is designed to recognize. Such pseudorecognitions are apparent neither in our behavioral data nor in Schildberger's recordings of cricket auditory neurons.

The third idea depends on the recent discoveries, in both the cricket and the frog, of band-pass neurons in close association with high-pass and low-pass neurons. Perhaps, as Schildberger has pointed out, these cells are trying to tell us something about the recognition mechanism. Perhaps, as a general principle, the nervous system recognizes simple repetitive patterns of sounds in a two-step process. In the first step two sets of neurons determine whether the repetition rate is above a given value (the high-pass cells) and below another value (the low-pass cells). In the second step a set of band-pass, or recognizer, cells become active only if both the low- and the high-pass cells are active. In mathematical logic or computer circuitry the sequence corresponds to the operation called logical AND. The anatomical data are not inconsistent with this idea. The low-pass BNC-1 cells appear to deliver signals to the band-pass BNC-2 cells. Often, moreover, the BNC-1 cells respond to sound stimuli before the BNC-2 cells respond. It would be interesting to know whether comparable relations obtain in the brain of the frog.

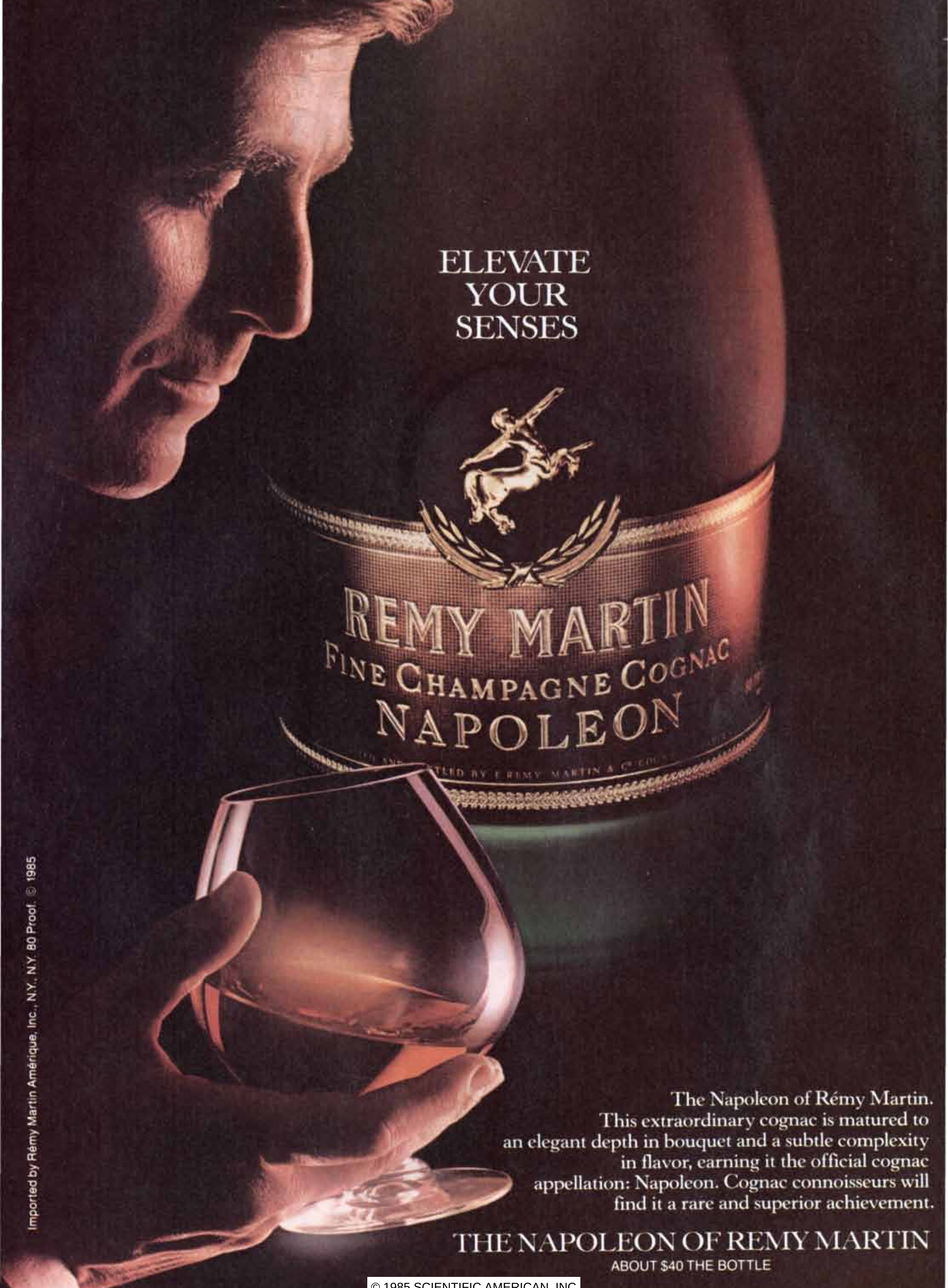


BRAIN of the cricket includes neurons that appear to participate in recognizing (and not simply in copying) the temporal pattern of the male calling song. Here the left side of the brain and two of the neurons within it are drawn as they would be seen from above the head of the cricket. The axon sent out by the AN-1 cell in the prothoracic ganglion (*black*) arrives from below. Its most forward synaptic terminals overlap the forward branchings of a population of neurons designated BNC-1, one of which is shown (*green*). In turn, branchings of the BNC-1 cells overlap those of another neuron type, designated BNC-2 (*red*). Recordings of the electrical activity of the neurons (*right*) show that they do not simply copy the temporal pattern of the male calling song. Instead some BNC-1 cells make low-pass responses to the range of synthesized songs (*see illustration on page 66*); some BNC-2 neurons make band-pass responses and some make high-pass responses. The findings suggest that the neurons take part in recognizing the male calling song and that they activate tracking.

A further step in the quest from the cricket's behavior to the neuronal machinery underlying it would be to determine how the interactions among neurons can produce high- and low-pass responses to sensory signals. Here we face an embarrassment of riches: the known interactions among neurons in fact explain too much. To put it another way, the basic neuronal properties of excitation and inhibition readily yield a wide range of bases for elaborate computations. For example, the well-known phenomenon of temporal summation, in which the rapid arrival of signals excites a neuron whereas slower arrivals do nothing, provides a high-pass filter. Let us say only that the details of the interactions among identified neurons in the brain of the cricket are now being evaluated in light of the computational possibilities. We all know how difficult it is to understand undocumented software written by someone else.

In the meantime studies of cricket behavior on the spherical treadmill continue to produce surprises. For example, Weber and we have now found that if calling songs are synthesized at unnatural carrier frequencies, the female is fooled into tracking an illusory male at an angle to that of the actual source of sound. Moreover, it has emerged that females maturing with only a single ear because of faulty development can often track sound quite accurately. Such findings are fundamentally affecting our ideas about directional hearing in crickets.

Two centuries ago the German artist August Johann Roesel von Rosenhof became so fascinated with insect behavior that he turned from painting human portraits to painting and studying insects. He offered a vivid account of cricket behavior: the male's "lust keeps him singing with scarcely a pause, until finally he gets his wish; if a female is nearby, she goes after the song." The female's performance continues today to offer special challenges to students of acoustics, behavior and neurophysiology.


ELEVATE YOUR SENSES
REMY MARTIN
FINE CHAMPAGNE COGNAC
NAPOLEON
The Napoleon of Rémy Martin. This extraordinary cognac is matured to an elegant depth in bouquet and a subtle complexity in flavor, earning it the official cognac appellation: Napoleon. Cognac connoisseurs will find it a rare and superior achievement.
THE NAPOLEON OF REMY MARTIN
ABOUT $40 THE BOTTLE
Imported by Rémy Martin Amérique, Inc., N.Y., N.Y. 80 Proof. © 1985

Presenting a camera with the human eye than See.

The pictures you see here are actual unretouched photographs, shot simultaneously without any exposure compensation.

High contrast. Shot in Aperture-Priority with the Nikon FA.

Back light. Shot in Programmed with the Nikon FA.

Shot in Aperture-Priority with a leading multi-mode SLR camera.

Shot in Programmed with a leading programmed SLR camera.

The Nikon FA. The biggest advance in automatic photography since automatic exposure.

Until now, the metering system of any automatic camera could do just one thing. Measure light and give you a technically correct exposure.

But as any photographer knows, a technically correct exposure doesn't always give you the best picture.

That's why Nikon developed the FA. The first camera with AMP (Automatic Multi-Pattern) metering. AMP is the only metering system that can automatically give you optimum exposure, not just technically correct exposure, even under extreme lighting conditions.

So what you see in your pictures is a lot more like what you saw with your eyes.

How AMP works.

AMP metering divides your picture into five segments and then individually measures and compares each segment, evaluating such factors as contrast ratios, variations in brightness levels and percentages of light and dark areas.

It then processes this information in its own Nikon microcomputer, comparing the components of your picture with those of nearly 100,000 photographs programmed into its memory, and instantly


IBM

Balancing Act

The effort of the Reagan Administration to clamp down on exports of technology that might be useful to the Soviet bloc has drawn fire from a "think tank" usually regarded as friendly to the Administration and from the presidents of 12 scientific and engineering societies. The Center for Strategic and International Studies of Georgetown University has issued a report saying that the present system of export regulation "is outdated and cost ineffective." The 12 presidents have written to Secretary of Defense Caspar Weinberger to express their "concerns regarding new controls being applied by the Department of Defense to unclassified technical information." They told him "our organizations will not...sponsor closed or restricted-access technical sessions at meetings or conferences conducted under their auspices."

The center's study, which began in 1983, was prompted by evidence that the U.S.S.R. was accelerating its development of high technology for military purposes as well as by the emergence of foreign competition in high-technology markets both at home and abroad. The study panel soon reached the conclusion that U.S. policy on export controls was poorly focused and that the administration of the control program was "overreaching in scope, inconsistent in application and overreliant on unilateral restrictions." Moreover, the innovative capacity of the U.S. was being regarded as a separate issue, when in fact the development of new technologies is crucial to the preparedness of the military, the strength of the economy and the role of the U.S. as the world's leader in technology.

What export policy should focus on, the study group concluded, is "the militarily critical goods and technologies whose acquisition could enable the Soviet Union to overcome its deficiencies or surpass U.S. capabilities in the foreseeable future and whose control is practicable for the period necessary to maintain a U.S. qualitative advantage." To manage such a policy, the group said, the U.S. will have to balance four tasks that are not fully compatible: encouraging new civilian technologies, promoting commercial applications of new technologies that improve the international competitive position of U.S. companies, accelerating the use of advanced civilian and military technologies in deployed weapon systems and, where possible, delaying Soviet access to technologies that are critical to deter or fight a war.

The 12 presidents were motivated by several recent Pentagon moves to restrict the delivery of papers at scientific and technical meetings on the ground that the information might be of technological value to the U.S.S.R. "In science and in engineering research," they wrote to Weinberger, "the open exchange of information ensures that critical peer review is applied to new advances, provides valuable cross-fertilization of ideas and helps avoid duplication of effort. One of the principal missions of our organizations is to encourage and provide opportunities for such exchange and thereby to promote advances in the fields of knowledge which we represent. Since such advances are also important to national security, we feel impelled to advise you of the counterproductive consequences of the current DoD policies."

Learning to Do Well

Racial discrimination in the U.S. seems to have changed its face during the past few decades. There appears to be less open discrimination. At the same time education has become a more important influence on the economic achievement of ethnic groups. The extraordinary educational levels of Asian-Americans have enabled them to compensate for the effects of discrimination and approach economic parity with the white majority. Blacks and Hispanics, whose educational levels are lower than those of the majority, still lag far behind economically. These conclusions emerge from an analysis of the economic effects of racial discrimination published in the *American Journal of Sociology* by Charles Hirschman of Cornell University and Morrison G. Wong of Texas Christian University.

Hirschman and Wong base their findings on the 1960 and 1970 censuses and on a national survey of economic status carried out by the Census Bureau in 1976. (Because of budget cuts at the bureau, comparably detailed economic data from the 1980 census have only recently been made available to social scientists.) Hirschman and Wong selected samples of black, Hispanic, Japanese, Chinese and Filipino men who were from 25 to 64 years old and compared their earnings and economic status with those of their peers who were members of the white majority. To separate the effects of discrimination from indirect effects due to differences in education and other factors Hirschman and Wong employed a statistical method.

The researchers found that although black and Hispanic men made some progress between 1960 and 1976, a large gap remains: black and Hispanic men earned an average of $4,000 less than whites in 1976. (All figures are in constant 1975 dollars.) The three Asian-American groups, on the other hand, now earn about as much as the white majority. Indeed, in some instances they earn more. In 1976 Japanese men made about $1,600 more than comparable whites. Chinese men, however, still lagged behind whites. The most dramatic progress was made by Filipinos: in 1960 they earned $4,500 less than whites; by 1976 they had achieved parity.

For almost all ethnic groups the influence of direct discrimination on earnings decreased over the 16-year period. For example, Hirschman and Wong calculate that in 1960 direct discrimination accounted for $1,600 of the difference in earnings between Hispanics and whites; by 1976 the figure had fallen to $1,000. The effect of lower educational levels among Hispanics, on the other hand, rose from $1,300 to $1,800. Educational overachievement was the most significant factor for Asian-Americans in compensating for the effects of discrimination, which persists as a problem, albeit a diminished one, for them.

"My perception is that there is less old-fashioned direct discrimination than there used to be. Discrimination is now somewhat more subtle and is linked more closely to how society is organized," Hirschman said. He speculated that among the subtle factors that may be operating to sustain discrimination are residential segregation and the exclusion of ethnic minorities from informal social networks. Such exclusion may make it difficult for minority-group members to obtain the contacts that often lead to better employment.

Nobel Prizes

Physiology or Medicine

Cholesterol circulating in the blood in the form of fatty particles called low-density lipoprotein (LDL) can accumulate to form atherosclerotic plaque, a buildup that narrows arteries and precipitates heart attacks and strokes. How does the body regulate the level of cholesterol in the blood? How might cholesterol metabolism be improved so that the blood level can be lowered?

For their answers to these questions the Nobel prize in physiology or medicine went to Michael S. Brown and Joseph L. Goldstein of the University of Texas Health Science Center at Dallas [see "How LDL Receptors Influence Cholesterol and Atherosclerosis," by Michael S. Brown and Joseph L. Goldstein; SCIENTIFIC AMERICAN, November, 1984].

In 1973 Brown and Goldstein discovered that the surface membrane of cells carries molecules of a protein specialized to bind to LDL and remove it from circulation. Genetic deficiencies in the number of receptor molecules, the workers found, are the cause of the extremely high LDL levels found in people suffering from a disorder called familial hypercholesterolemia. They lack either one or both of the genes coding for the receptor; victims of the more severe form have almost no LDL receptors and develop atherosclerosis in childhood.

Most people with atherosclerosis have a lowered number of LDL receptors as a result of subtler genetic factors combined with environmental influences, among them a diet high in cholesterol. Brown and Goldstein found that cells protect themselves from high levels of circulating cholesterol by slowing the synthesis of the LDL receptor molecule. The result is a feedback effect that hampers the removal of LDL from the blood and sustains the high cholesterol level.

Brown and Goldstein's findings led them to try increasing the number of LDL receptors in people suffering from a mild form of familial hypercholesterolemia, in which receptors are synthesized at a reduced rate. The liver removes large amounts of cholesterol from the blood and converts it into bile acids, which aid in the digestion of fats. Ordinarily bile acids are reabsorbed from the intestine and reused. By administering bile-acid-binding resins, which prevent the reuptake of bile acids, and a drug that inhibits the liver's ability to synthesize cholesterol on its own Brown and Goldstein increased patients' need for cholesterol. Their liver cells made more receptors and their LDL levels fell sharply.

The prizewinning work, a combination of genetics, molecular biology and microscopy, makes it possible to conceive of new drugs that would cause cells to make more LDL receptors. In combination with changes in diet such drugs might enable more people to escape the plague of atherosclerosis, which currently causes half of all deaths in this country.

Physics

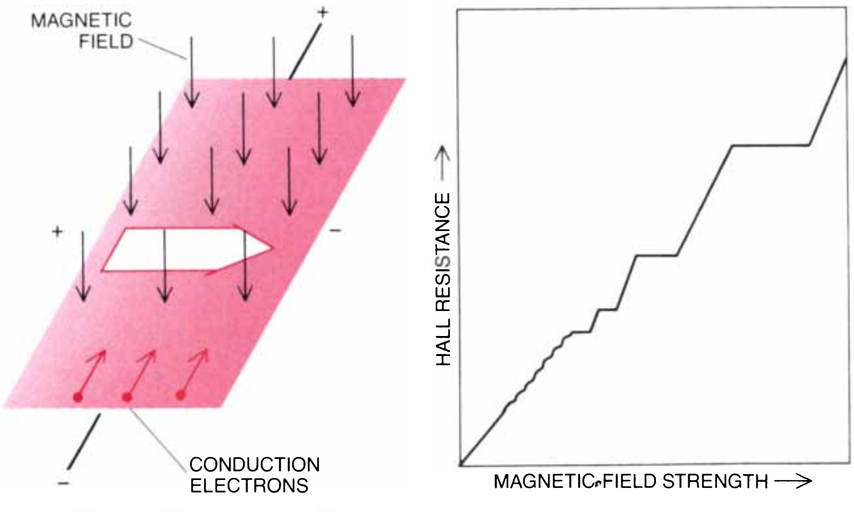
For the discovery of what the Royal Swedish Academy of Sciences called "a new phenomenon in quantum physics," Klaus von Klitzing of the Max Planck Institute for Solid-State Research in Stuttgart was awarded the 1985 Nobel prize in physics. In 1980 von Klitzing found that under specific circumstances the Hall effect, a feature of the interaction of an electric current with a magnetic field observed more than a century ago, becomes quantized. Instead of displaying the continuous variability expected in classical electrodynamics, it changes in discrete steps that are governed by fundamental constants of quantum mechanics.

In 1879 Edwin H. Hall, an American physicist, noted that a conductor carrying an electric current through a perpendicular magnetic field develops a lateral voltage difference. The magnetic field deflects the moving electrons in a direction that is perpendicular to both the current and the lines of force, causing electric charge to pile up on one side of the conductor. The voltage difference across the conductor is called the Hall voltage; dividing the Hall voltage by the current along the conductor gives the Hall resistance.

Ordinarily the Hall resistance increases in proportion to the strength of the magnetic field. Working with Gerhard Dorda of the Siemens Research Laboratory and Michael Pepper of the University of Cambridge, von Klitzing found that in a two-dimensional layer of electrons maintained at a temperature near absolute zero and subjected to strong magnetic fields the Hall resistance exhibits quantized plateaus over certain intervals of magnetic field strength. In the same intervals the resistance to current flow along the conducting layer falls to zero. Between the plateaus the Hall resistance increases as the field grows stronger.

Since the experiment an explanation has emerged. Electrons moving in a two-dimensional layer that is subjected to a strong magnetic field are confined in circular orbits in which they can assume only certain discrete energy states. The allowed states for all the electrons in the conductor cluster in a hierarchy of energy bands. Because of imperfections in the conductor, the states making up each band are not precisely equal in energy. In the presence of impurities some electrons are trapped in "localized states" with a slightly lower or higher energy than the "extended states" at the center of each band, in which electrons are free to move about the conductor.

Raising the magnetic field strength increases the number of states within each energy band. Because the experiment takes place at low temperature, where the kinetic energy of the electrons is small, they assume the lowest unoccupied states. Electrons flow into lower bands and the Fermi level (the energy level of the most energetic electrons) falls. If the Fermi level lies



HALL EFFECT arises when a conductor carries a current through a perpendicular magnetic field (*left*). The moving electrons experience a lateral force, which causes them to accumulate on one side of the conductor. The Hall resistance, one measure of the effect, ordinarily increases in proportion to the strength of the magnetic field. In a layer of electrons subjected to low temperature and intense fields, the Hall resistance becomes quantized: it exhibits plateaus at values that reflect fundamental constants of quantum mechanics (*right*).


Announcing the 1986 Seville.
The new essence of elegance.
You've never seen a Seville like this before. The tailored, smooth-flowing design looks invitingly contemporary, yet its understated elegance remains strictly Seville.
On the road, the 1986 Seville handles magnificently. Trim, taut, and responsive. Without sacrificing the luxury you expect from the ultimate Cadillac.
Behind the real American walnut-trimmed instrument panel lives an electronics system so advanced it was inconceivable a few short years ago. Technology not just for technology's sake, but technology designed to

contribute to your driving comfort.

An especially revealing way to appreciate the 1986 Seville is to study the obvious attention to detail. The fit. The finish. The smooth flow of one design plane into the next. The precision of the electronic instrumentation. The quality of Seville is backed by a 4-year/50,000-mile limited warranty. In some cases, a deductible applies. See your dealer for details.

Destined to become another American classic, the 1986 Seville. The perfect combination of luxury and technology.

1986 SEVILLE
BEST OF ALL...IT'S A CADILLAC.

within a region of localized states at the upper or lower edge of a band, the number of electrons occupying extended states and therefore capable of contributing to the Hall effect does not change. Thus the Hall resistance exhibits a plateau.

When the Fermi level descends to the center of the band and electrons begin to vacate the extended states, the number of electrons available for conduction decreases and the Hall resistance begins to climb again. The Hall resistance at the plateau is determined by an integer and two fundamental constants of quantum mechanics: the charge of the electron and Planck's constant.

At the plateaus the most energetic electrons are trapped by impurities; as a result the flow of current can take place only among the electrons in the filled extended states of lower energy bands. Under such circumstances current flows without dissipating energy, and the resistance along the conductor drops to zero.

The quantized Hall effect is independent of the crystal geometry of the conductor and the number or kind of impurities. The values of the Hall resistance at the plateaus deviate from perfect quantization by less than one part in 10 million. The high precision of the effect may make it useful as a standard of electrical resistance and a means of assigning more precise values to quantum-mechanical constants.

Chemistry

Twenty years ago arriving at the structure of a simple biochemical molecule from X-ray crystallographic data could take years; now such structures can be solved in a few days. Underlying the gains are mathematical techniques known as direct methods. Citing "the great importance of this methodology for chemical research," the Royal Swedish Academy of Sciences awarded this year's Nobel prize in chemistry to the developers of the direct methods, Herbert A. Hauptman of the Medical Foundation of Buffalo and Jerome Karle of the U.S. Naval Research Laboratory.

X rays beamed through a crystal of a substance pick up clues about structure. The planes of atoms or molecules in the crystal scatter the waves, which interfere with one another to produce a diffraction pattern reflecting the large-scale crystal geometry. Information on the structure of the individual crystal units is contained in the relative intensities and phases of the waves scattered to each point in the pattern.

The intensity of the radiation can be measured directly, but when Hauptman and Karle developed their methods in the 1950's, information on phase was thought to be inaccessible. To solve the structure of even a simple molecule workers of the day "decorated" the molecule with atoms of a heavy element such as gold; heavy atoms scatter X rays more strongly than the lighter atoms of biological molecules do. Comparison of data from several samples that had been imbued with different heavy elements made it possible to infer the phases of waves scattered from the other atoms.

Building on earlier proposals that the relative intensities of the spots in the diffraction pattern contain information about phase, Hauptman and Karle developed a mathematical means of extracting the information. A fundamental proposition of their direct methods is that if three intense spots in the diffraction pattern have positions whose coordinates add up to zero, their relative phases will cancel as well. Computations done with many trios of spots yield probable phases for a significant number of diffracted waves. Further mathematical analysis leads to a likely solution for the structure of the molecule as a whole. Controversial at first, the direct methods are now the standard procedure for determining the structure of small molecules such as hormones, vitamins and drugs.

The methods are not yet adequate for analyzing proteins and other large molecules, for which the heavy-atom method is still a mainstay. But improvements in X-ray sources, in the collecting of data and in computing power may soon make it possible to apply the techniques devised by Hauptman and Karle to complex biological molecules. "There is very great promise," says William L. Duax of the Medical Foundation of Buffalo, "that these techniques will underlie a revolution in the study of protein structures."

Bad Blood?

Have the screening tests for AIDS-virus antibodies made the U.S. blood supply safe? At least one investigator believes the answer is negative, and although his evidence is far from conclusive his colleagues take his point seriously. Myron Essex of the Harvard School of Public Health first stated his case in a paper published a year ago in *The Lancet,* which he wrote in collaboration with several other workers, including Robert C. Gallo of the National Cancer Institute, one of the discoverers of the AIDS virus. Essex has repeated his conclusion in recent meetings at the Memorial Sloan-Kettering Cancer Center and at Harvard. According to Essex, at least 5 percent of the blood samples from people carrying the AIDS virus escape detection by the routine ELISA (enzyme-linked immunosorbent assay) test for AIDS and even from the more definitive Western-blot confirmatory test. Both tests detect the presence of antibody to the AIDS virus in the blood.

In Essex' original study virus-carrying blood samples from three symptom-free sexual partners of known AIDS patients tested negatively on ELISA and Western-blot tests, an indication that these people did not produce antibody to the AIDS virus. This result would make their blood potentially acceptable to blood banks.

More recently Essex analyzed an ELISA screening of 107,000 blood samples. On the first run 719 specimens had been positive, that is, they carried antibodies to the AIDS virus. When the same test was repeated twice more, only 218 specimens were found to be positive by all three tests. Western blotting yielded only 35 positives out of the group of 218. What do the results mean? Essex says the antibody tests are inconclusive and studies to determine how many of the positive blood samples actually harbor virus have not been done, but that clearly the results are not consistent with the "99.8 percent accurate" rating the U.S. Food and Drug Administration has claimed for AIDS blood testing. Essex has no plans to do a conclusive study, but he thinks his work reveals a serious situation. "The FDA gave an unrealistic picture of the existing ELISA tests," he says. "My purpose is to show that these first-generation antibody tests have to be improved before they are reliable."

Essex suggests three reasons the antibody tests give confusing results. First, the tests are done with a crude fraction of the virus, and it is conceivable that the most antigenic parts of the virus were not incorporated. In his laboratory Essex is currently pursuing methods that sharply define particular AIDS antigens. Second, there are some people who may be "in transition." This means that the virus has already infected them but they have not yet started to make antibodies against it. Third, there may be some infected people who will never make antibodies to the AIDS virus for reasons as yet not known.

Jeffrey Laurence of the Cornell Medical Center would like to see Essex' results tested in a large-scale study. Nevertheless, Laurence says, "I believe in the work—the evidence is very disturbing." He thinks a new blood test is required that detects the viral particles themselves (on the order of the hepatitis-*B* test). Until then many "false negative" individuals will continue to be potential blood donors.

Networking

A human protein known as angiogenin that promotes the growth of blood vessels has been isolated and characterized by Bert L. Vallee and his colleagues at the Harvard Medical School. Their work represents the first direct confirmation of the existence of a specific protein that stimulates angiogenesis, the development of vasculature. The discovery yields insights about both normal and cancerous growth and could lead to therapies for a number of serious diseases.

The existence of such a regulatory protein was first proposed in 1961 by Judah Folkman, then at the Naval Medical Research Institute. Folkman suggested a tumor might require its own system of blood vessels in order to grow beyond a millimeter or two in diameter. He hypothesized that a so-called tumor angiogenesis factor stimulates the development of such growth. For the past 20 years, working at the Harvard Medical School, Folkman has searched for the factor.

A $23-million, 12-year grant from the Monsanto Company has supported both Folkman's and Vallee's work. It was agreed that Folkman would focus on studying the biology of vessel growth and that Vallee would concentrate on searching for the angiogenesis factor. The first major success came last year, when Folkman succeeded in extracting from rat carcinoma cells an active but still crude angiogenesis factor that promotes vascularization.

Initially Vallee and his colleagues also worked with rat carcinoma cells. Four or five years ago they switched to human tumor cells. James F. Riordan of the Vallee group notes that they made the change hoping the successful isolation of a human angiogenesis molecule might have an immediate impact on the therapy of diseases.

Working with cells derived from a human colon cancer, Vallee's group isolated about a milligram of pure human angiogenin from roughly 2,000 liters of fluid in which the cells were cultivated. The key factor in their success was that they were able to grow and maintain the tumor cells under serum-free conditions. The large number of proteins in serum would have complicated the search for angiogenin.

To winnow out the factor from other proteins produced in the cells, the Vallee group resorted to a conventional purification technique known as liquid chromatography. They determined whether or not an isolated protein or proteins stimulated blood-vessel growth in the living tissue of such animals as chicks, rabbits and mice. The workers confirmed the effect of angiogenin when they observed that very small amounts of it stimulated such growth. In a series of three articles in *Biochemistry* the investigators note that they have determined the amino acid sequence of angiogenin and, together with Kotoku Kurachi and Earl W. Davie of the University of Washington, have cloned its gene.

Angiogenin may someday be used to increase the blood supply in patients who have had a stroke or suffer from heart disease or fetal blood insufficiency. It could also promote the healing of wounds. Obstruction of angiogenin's effect could be exploited to decrease the flow of blood to solid tumors; rheumatoid arthritis, diabetic symptoms and certain skin disorders might also respond.

Quantum Cats

Does the theory of quantum mechanics apply to macroscopic objects? The question has paradoxical overtones. A fundamental tenet of quantum mechanics is that a particle such as an electron does not exist in a definite physical state until it has been observed. The unobserved particle has an endless number of options; it is said to exist in a superposition of states, each one described by a wave function that defines the probability of finding the particle in that state. The act of observation "collapses" the wave superposition, fixing the particle in a single position in space and time. Hence on the atomic level, at least, human observers create reality. Can this bizarre denial of objectivity extend to the everyday world?

Fifty years ago Erwin Schrödinger, who wrote the quantum-mechanical wave equation, dramatized the question with a thought experiment. Imagine a sealed box, he proposed, containing a cat, a vial of poison gas, a piece of radioactive material and a radioactive-particle detector. The device is rigged so that the detection of a radioactive particle triggers the release of the poison and the death of the cat. Although it is impossible to predict when a radioactive atom will decay and emit a particle, one can calculate the probability that a decay will occur in a given time period. The detector is switched on just long enough for it to have a 50 percent chance of registering a particle, and so the cat has a 50 percent chance of surviving. According to the dominant interpretation of quantum mechanics (the Copenhagen interpretation), it is meaningless to say, before the box has been opened, that the cat must be either alive or dead. Rather, it exists in a superposition of states: it is both alive and dead.

The quantum weirdness of Schrödinger's cat is unobservable; if someone were to look at the cat, one of its two superposed states would instantly crystallize into reality. For many physicists this makes it unnecessary to think about the uncomfortable implications of the paradox. In recent years, however, a few workers with a taste for fundamental questions have searched for macroscopic variables other than the aliveness of a cat that might exhibit experimentally verifiable quantum-mechanical behavior.

Many of these efforts have focused on an electronic device called a Josephson junction, which consists of a pair of superconductors separated by a thin layer of insulating material. According to classical physics, no current should flow through such a junction. Yet in quantum mechanics the electrons in the superconductors are represented as probability waves, and their position is indeterminate; there is a finite probability that an electron from one superconductor will pop up on the other side of the insulator. If the insulator is thin enough, many electrons can "tunnel" through the insulator and a current can flow. Indeed, the current can flow without resistance (as it does through a superconductor). In that case no voltage spans the junction. The observation of tunneling currents through Josephson junctions is itself testimony of the validity of quantum mechanics, but only at the atomic scale.

The electrons in a superconductor, however, also display a macroscopic order: their wave functions share the same wavelength and the waves are all in phase. The electrons on one side of a Josephson junction have a different phase from those on the other side; the magnitude of the phase difference depends on the intensity of the tunneling current. Since the phase difference is a property of an immense number of electrons, it can be considered a macroscopic variable. Is its behavior classical or quantum-mechanical? John M. Martinis, Michel H. Devoret and John Clarke of the University of California at Berkeley have examined the question. They report their results in *Physical Review Letters*.

The experiment carried out by the Berkeley group with a Josephson junction in the zero-voltage state is best understood in terms of the system's classical mechanical analogue: a particle moving in a potential energy well. The particle corresponds to the phase difference. The escape of the particle from the well corresponds to the appearance of a voltage across the junction. According to classical physics, the particle can escape from the well only by acquiring enough energy to go over the top, and as it oscillates back and forth it may be found at any energy in the well. According to quantum mechanics, however, the particle can escape the well by tunneling through the energy barrier. Moreover, while it is in the well it can occupy only certain quantized energy levels whose values can be calculated from its Schrödinger wave function.

Martinis and his colleagues observed both quantum effects by irradiating their circuit with microwaves—the equivalent, in the classical analogy, of applying a force to the particle. The microwaves raised the phase difference to a higher "energy" and thereby increased its "escape rate" (determined by the speed with which a voltage appeared across the junction). More important, the escape rate showed sharp peaks at certain microwave frequencies. Since frequency is proportional to energy, this implied the energy levels in the well were quantized; the workers found that the peaks came at precisely those frequencies corresponding to the energy transitions allowed by the wave equation. Furthermore, the phase difference escaped at a significant rate even when its energy was far below the top of the well, indicating it was tunneling through the barrier.

The Berkeley experiment proves that representing the phase difference across a Josephson junction as a particle is inadequate, just as representing an electron as a particle is inadequate: both must be described as a wave function, as a superposition of states. Of course, it is still a long way from an electronic phase difference to the world of everyday objects. Nevertheless, the fact that quantum mechanics applies to at least one macroscopic variable suggests one should get used to the idea of living dead cats.

Tiny Transistor

In order to build faster and more powerful computers investigators have sought to increase the speed of the building blocks of computational devices: integrated-circuit chips consisting of groups of transistors. In this domain of electronics faster usually means smaller: a transistor that can be switched on and off very rapidly must be very small. Workers at the Massachusetts Institute of Technology have now produced what could be the smallest transistor ever imprinted on a silicon chip.

The basic component of a conventional transistor is a piece of semiconducting material such as silicon. The substrate consists of three regions known as the gate, the source and the drain. By applying a relatively small voltage to the gate one can regulate a large flow of current between the source and the drain. In other words, the gate acts as a switch: a small signal applied to the gate results in the passage of a large signal from the source to the drain. The amplified signal travels through an electronic "channel" of the transistor. The shorter the channel is, the faster the transistor responds.

The M.I.T. team, consisting of Stephen Chou, Henry I. Smith and Dimitri A. Antoniadis, fabricated their transistor by means of X-ray lithography, a process in which a mask pattern containing an electronic circuit is imprinted on a chip. The channel length of the transistor is about 17 times shorter than that of transistors now in commercial use, which is roughly a micrometer, or 40 millionths of an inch. Moreover, at temperatures of 4.2 degrees Kelvin (269 degrees below zero Celsius) the average electron velocity in the tiny transistor channel is 1.8 times greater than the "saturation," or maximum, velocity observed in some kinds of transistors. The M.I.T. workers hope to reduce the level of impurities in the transistors so that the electrons can travel even faster and at high temperatures as well.

An Uncertain Beacon

The celestial object designated GX5-1 emits X rays whose intensity varies irregularly between 20 and 40 times a second. The discovery of these quasi-periodic oscillations is puzzling because an X-ray binary (the kind of object GX5-1 is presumed to be) typically emits radiation at highly predictable intervals. What causes GX5-1 to deviate from strict periodicity?

An X-ray binary is made up of two proximate stars: a "normal" star and a massive, extremely dense star. Matter from the normal star is drawn toward its small but superdense companion by the latter's strong gravitational pull. The matter tends to accrete around the superdense star, forming a shrouding disk of gas. Matter from the inner region of the disk streams continuously toward the star's surface, accelerating to speeds in excess of 80,000 kilometers per second.

At such extremely high speeds the gas heats up to temperatures in the tens of millions of degrees and begins to radiate, primarily in the X-ray region of the electromagnetic spectrum. X rays traveling in a direction more or less perpendicular to the accretion disk can then escape into space. Such stellar configurations would produce periodic fluctuations in X-ray intensity each time the donor star (which, although it is less massive, is considerably more voluminous) eclipses the X-ray-emitting star. Only two objects could exert enough gravitational pull on a nearby companion to "drain" its matter and accelerate the matter to X-ray-emitting velocities: a black hole or a neutron star.

A neutron star in such a binary system can display a second, more rapid kind of periodic fluctuation. Because the "falling" matter is electrically charged, it is forced to follow the neutron star's magnetic field lines, which terminate at the magnetic poles. A sufficiently strong magnetic field would constrain the X-ray emissions to two sites on the neutron star: its north and south poles. If the magnetic poles are not aligned with the star's axis of rotation, then, for every revolution the star makes, a circle would be swept out by an X-ray "beam" emanating from each pole. To an observer on the earth such a star would manifest itself much as a lighthouse does. The star would emit brief and regular "flashes" of X rays each time a pole lined up with the observer.

Because a neutron star's rate of rotation is stable, the X-ray pulsations normally occur at very precise intervals. Nevertheless, such a cosmic "clock" tends to run fast. The matter impinging on the surface of the neutron star imparts a tiny but constant angular momentum to the rotating star. This causes the star to gradually rotate faster while "ingesting" the matter from its companion star. Moreover, it can be predicted that the radius of the neutron star would actually shrink as mass accumulates—an effect that further increases the rate of rotation.

Consequently an X-ray clock of this kind should speed up in the course of millions of years. Although many relatively young X-ray binaries (those a million or so years old) with periods of roughly half a second have been observed, X-ray astronomers had yet to find an old one (about a billion years old) with a period measured in milliseconds.

It was the hope of proving GX5-1 to be precisely such an object that brought together six investigators: Michiel van der Klis of the European Space Agency's European Space Research and Technology Centre in Noordwijk, Fred Jansen of the Laboratory for Space Research in Leiden, Jan van Paradijs and Edward P. J. van den Heuvel of the University of Amsterdam, Joachim E. Trümper and Mirek Sztajno of the Max Planck Institute for Extraterrestrial Physics in West Germany and Walter H. G. Lewin of the Massachusetts Institute of Technology. As they report in *Nature,* they found GX5-1 to be a surprisingly wobbly X-ray source.

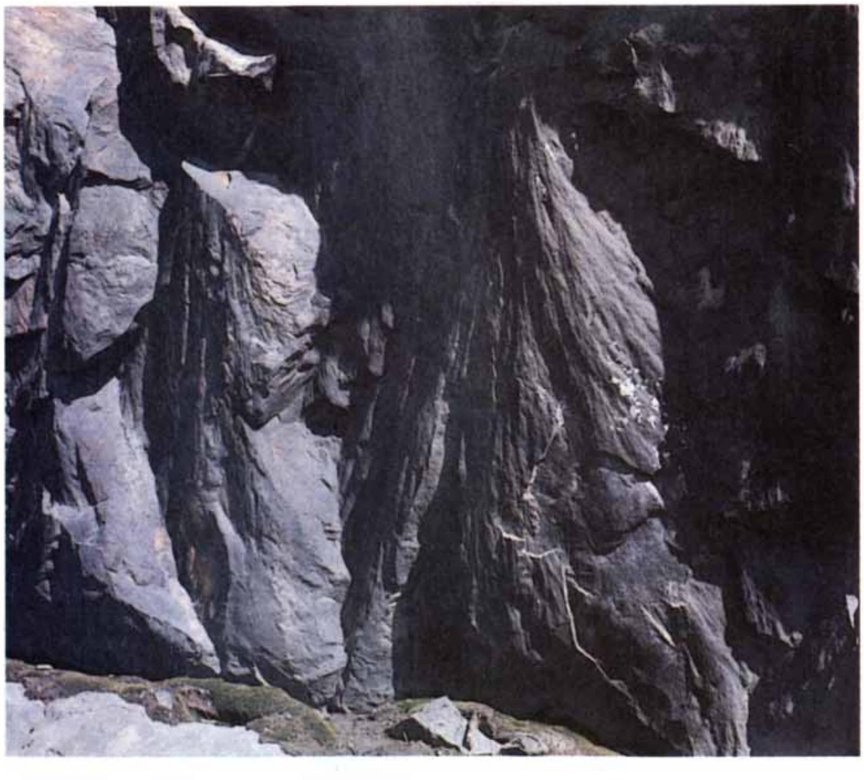
The theories that best fit their observations of GX5-1 share a common premise, namely that the accretion disk is broken up into "blobs" by the neutron star's strong, rotating magnetic field. These blobs of matter give off X rays as they plunge to the surface of the star but are opaque to X rays while they are in orbit. X-radiation can "leak out" between the orbiting blobs, and because the blobs occupy a distribution of orbital periods, a distribution of X-ray fluctuations is consequently detected on the earth.

Almost all these theories also share a common deficiency: they require the presence of a fairly strong magnetic field. Yet a strong magnetic field would produce the characteristic X-ray beacons, which appear to be absent in the case of GX5-1. In addition the supposed strength of the neutron star's magnetic field would seem to be inconsistent with the advanced age attributed to GX5-1. As a neutron star ages its magnetic field is supposed to decay. The hypothesized magnetic field strengths for GX5-1 are more characteristic of younger neutron stars.

Nickels from Heaven

The impact of a giant meteorite, supposedly the most disastrous of natural disasters, may in the fullness of time have a small beneficial effect. That is one conclusion to be drawn from a recent study of an elliptical basin of igneous rock, 60 kilometers long by 27 kilometers wide, near Sudbury, Ontario. Known as the Sudbury Igneous Complex, the basin supplies about a fifth of the world's nickel and is also a significant source of copper and platinum. New geochemical evidence suggests it was formed by a meteorite impact some 1.84 billion years ago.

The standard explanation of igneous formations—that they were created by the cooling of molten rock (magma) rising from the earth's mantle—has never quite fit the Sudbury structure. For example, Sudbury rocks are richer in silica and in certain rare-earth elements than typical mantle-derived



SHATTER CONES in rocks around the Sudbury Igneous Complex suggest the rocks were fractured by shock waves from a meteorite impact. This cone is about three meters high.

rocks. Furthermore, rocks around the igneous complex display signs of having been shattered by a tremendous shock, and the basin itself is overlain by igneous breccias (fragmental rock) that incorporate shocked fragments of the surrounding country rock. On the basis of such evidence Robert S. Dietz first proposed in 1964 that Sudbury was the site of an ancient meteorite impact. According to this hypothesis, heat released in the impact melted the crust, and the melt cooled to form the igneous complex.

Igneous rocks thus derived from ancient continental crust should be chemically distinguishable from a more recent intrusion of magma from the mantle, because the process of chemical differentiation that has separated the crust from the mantle has progressively concentrated certain elements in the crust. The larger the ionic radius of an element is, the more likely it is to migrate out of the dense crystal structure of the mantle into the melt that later becomes crust. The rare-earth element neodymium, for instance, is slightly larger than the closely related element samarium, and so over billions of years the crust has become enriched in neodymium. In contrast, the uppermost part of the mantle contains relatively more samarium. By analyzing the neodymium and samarium content of an igneous rock one can therefore determine its source.

In practice the most unambiguous signature is the ratio of two neodymium isotopes. Neodymium 143 is produced by the radioactive decay of samarium 147, whereas neodymium 144 is neither produced nor consumed by radioactive decay. Over time the ratio of neodymium 143 to neodymium 144 increases, but the increase is much slower in crustal rocks, which contain less samarium to begin with.

Asish R. Basu and Billy E. Faggart, Jr., of the University of Rochester and Mitsunobu Tatsumoto of the U.S. Geological Survey have analyzed 16 rock samples from the Sudbury complex. The workers report in *Science* that the neodymium isotope ratios of all the samples are unequivocally crustal and are radically different from those observed at other igneous formations. From the known rate at which samarium 147 decays to neodymium 143, Basu and his colleagues calculate that the Sudbury rocks crystallized about 1.84 billion years ago. At that time, they suggest, a meteorite impact melted existing layers of basalt, granite and sediments. The heavy and immiscible sulfides of nickel and copper separated out of the magma, eventually becoming concentrated in rich ores.

The Immune System in AIDS

The AIDS virus alters the growth and function of T4 lymphocytes, a class of white blood cells that is crucial to the immune system. New knowledge of how the virus does so may lead to treatments and perhaps a vaccine

by Jeffrey Laurence

In 1981 the Centers for Disease Control in Atlanta recognized the first cases of a fatal new disorder that came to be known as acquired immune deficiency syndrome (AIDS). Its victims died of a variety of rare infections and malignancies, among them a pneumonia caused by the protozoan *Pneumocystis carinii* and Kaposi's sarcoma, a cancer of the lining of blood vessels. They also suffered from other "opportunistic" infections, caused by microorganisms that are ubiquitous but ordinarily not able to cause disease. Indeed, the infections and cancers seen in AIDS patients were previously known only in people born with certain defects in their immune system and in patients whose immunity had been impaired by cancer chemotherapy or the immunosuppressive drugs given for organ transplantation. AIDS, it appeared, killed its victims by destroying their immune system.

Since the disease was first recognized the number of cases has risen swiftly; in the U.S. alone the figure reached 14,000 by late 1985. Knowledge of AIDS has increased proportionately. Françoise Barré-Sinoussi, Jean-Claude Chermann and Luc Montagnier at the Pasteur Institute in Paris and a group led by Robert C. Gallo at the National Cancer Institute independently identified the causative agent, a virus of the retrovirus family, in 1983 and 1984 respectively; the French group called it LAV (lymphadenopathy-associated virus) and the American workers HTLV-III (human *T*-lymphotropic virus type III).

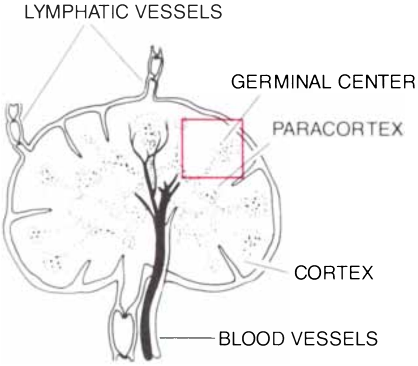
The groups at highest risk for infection have become increasingly well defined; they include homosexual and bisexual men, abusers of injected drugs, the sexual partners of people in AIDS risk groups, and children born to mothers at risk. Recipients of blood transfusions and blood products have also contracted AIDS, but the screening of donated blood for evidence of infection has drastically reduced their risk. The fact that the disease shows no sign of spreading beyond those groups, except to predictable targets such as women who are artificially inseminated with sperm from infected donors, indicates that the virus is ordinarily transmitted only through the blood or through sexual intercourse. All epidemiologic evidence indicates that food, water, insects and casual contact do not spread AIDS.

An understanding of the mechanisms by which the disease cripples the immune system is also emerging. The knowledge may make it possible to lessen the effects of the disease, and eventually to prevent and cure it. It also underscores the intricacy of the immune system. It now appears that the total collapse of the immune defenses in AIDS victims stems largely from a single defect: a reduction in the number and a change in the function of the *T*4 lymphocytes, one of the many distinct kinds of cells that make up the immune system.

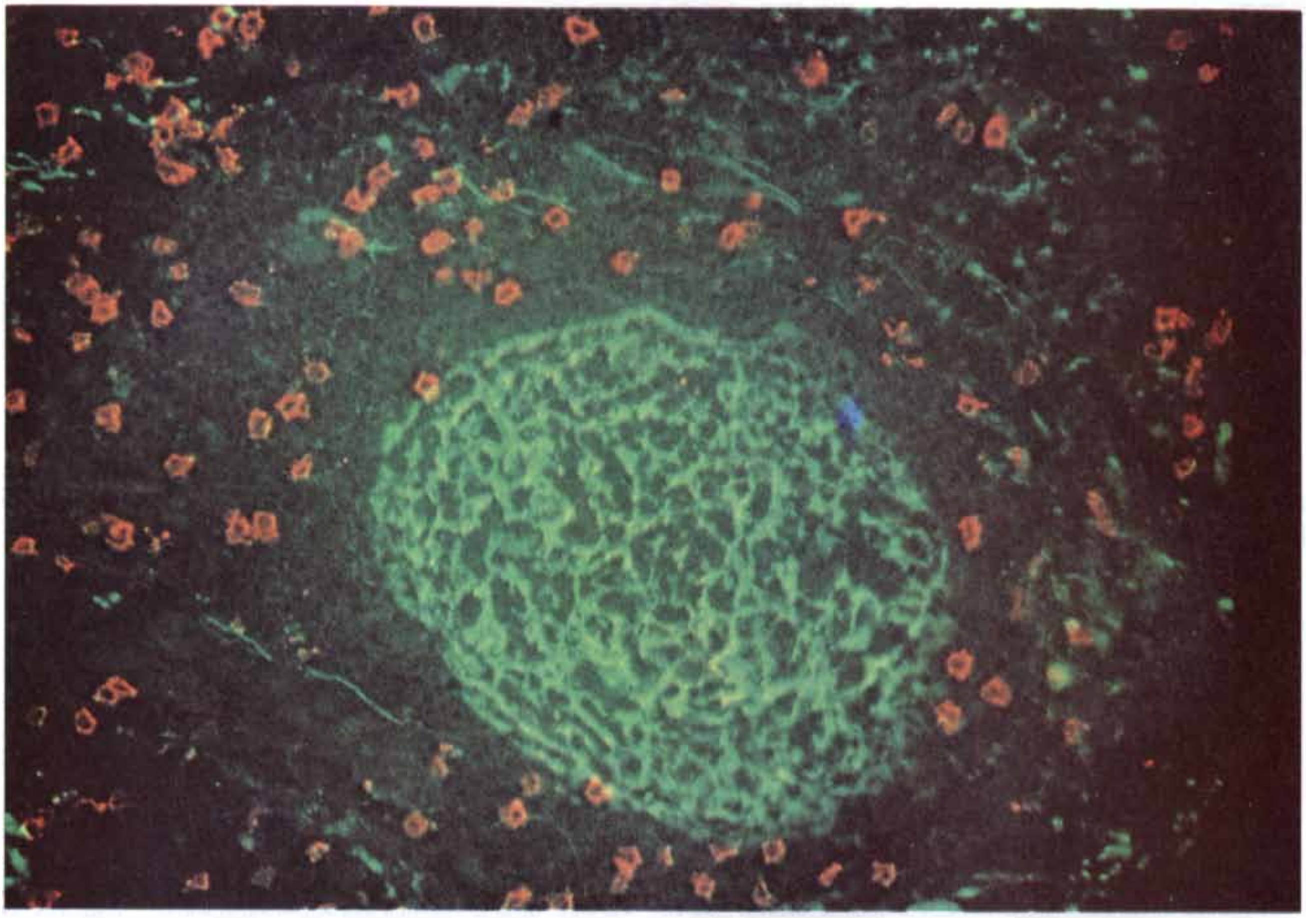
The Immune System

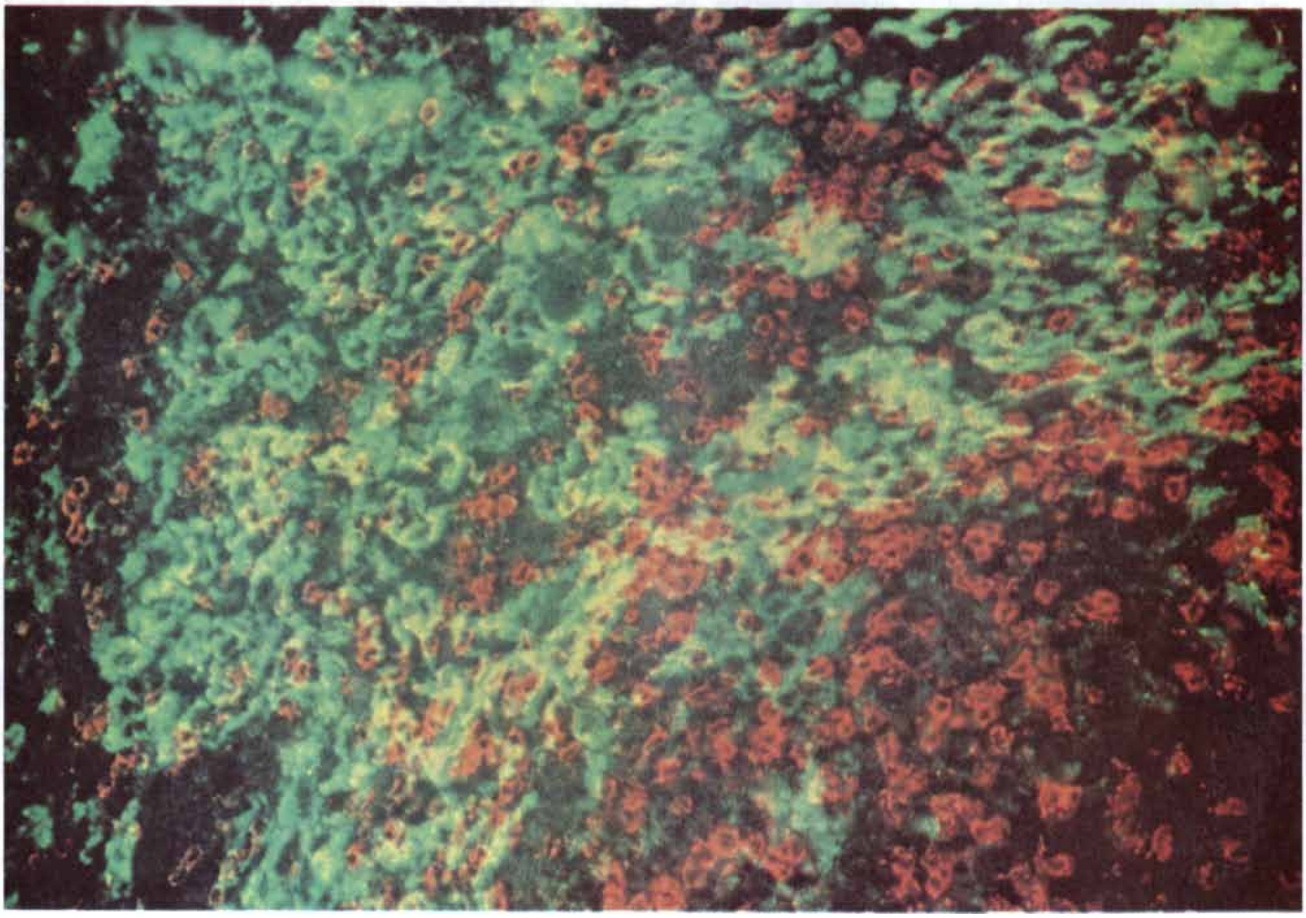
How is it that by damaging a single link the AIDS virus causes the immune system as a whole to unravel? The answer lies in the complex web of interactions among the different classes of blood cells that take part in immunity. The immune system is a flexible but highly specific defense mechanism that kills microorganisms and the cells they infect, destroys malignant cells and removes debris. It distinguishes such threats from normal tissue by recognizing antigens, or foreign molecules, and mounting a response that varies with the nature of the antigen.

Over a lifetime many thousands of different antigens challenge the body. The cells of the immune system can recognize and respond to virtually every antigen because they are divided into millions of clones, each of them made up of one or more cells specialized to recognize one or a few distinct antigens. In addition to this differentiation by antigen specificity, the cells also diverge into several types as they mature from common precursor cells. The cell types can be distinguished by the contexts in which they recognize antigen, by the molecular characteristics of their surface membranes and



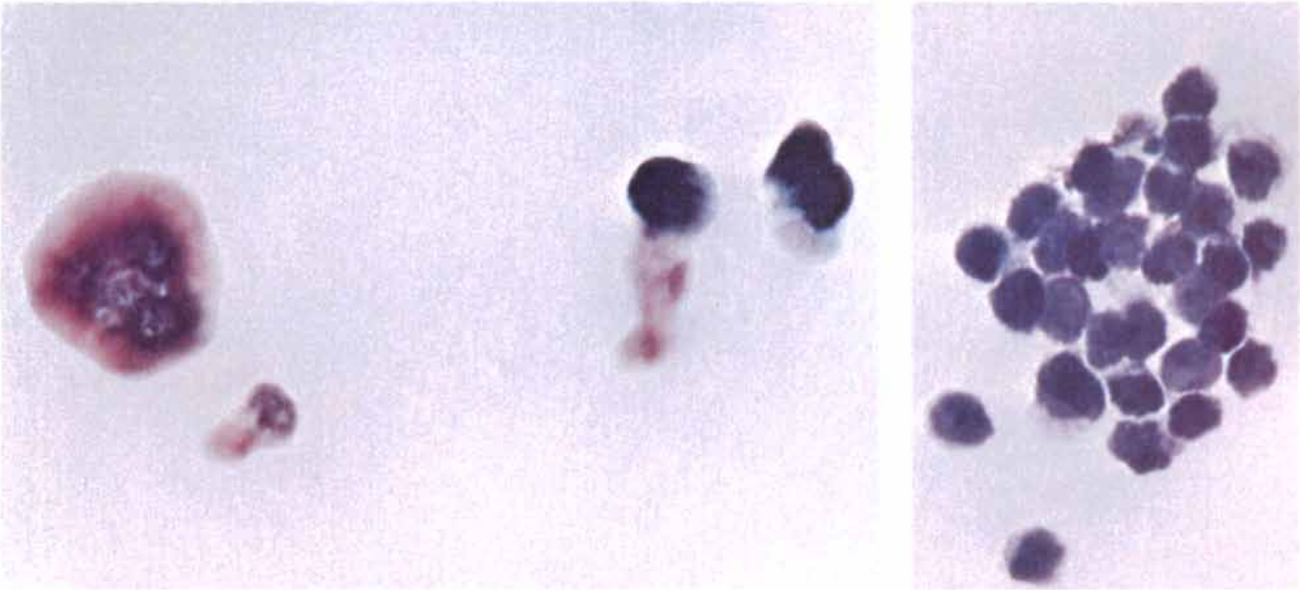
LYMPH-NODE DISRUPTION is characteristic of AIDS and the collection of symptoms, called AIDS-related complex (ARC), that is one of its precursors. Fluorescence micrographs of sections of lymph node show the structures outlined in the diagram (*above*): a germinal center and some of the surrounding cortex and paracortex. In a normal node (*top*) the white blood cells known as *T*8 lymphocytes (*orange*) populate the paracortex but are not present in the germinal center, in which *B* lymphocytes and the structural cells of the node, the dendritic reticular cells (*green*), form a regular network. A section of a swollen node from a patient with ARC (*bottom*) contains many more *T*8 cells; they invade the germinal center, whose regular structure is disrupted. It is not known how the AIDS virus causes these changes in node architecture, but they may have a relation to the immune deficiency: the surface membranes of normal dendritic reticular cells are thought to trap and retain foreign proteins that trigger an immune response. Lymph-node disruption may also serve as an early clue to infection with AIDS virus. George Janossy of the Royal Free Hospital in Hampstead, England, made the images.



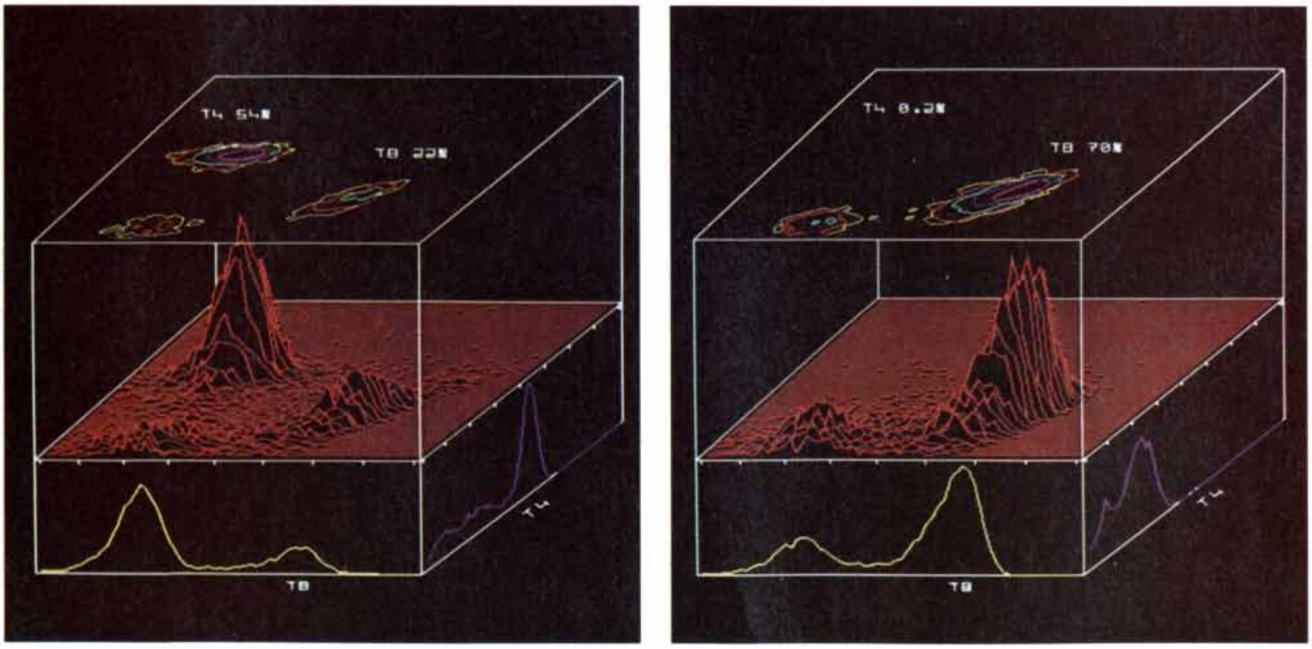


by their role in the immune response.

In broadest outline the white blood cells that are central to the immune response, the lymphocytes, fall into two classes: the *B* cells, so named because they develop in the bone marrow, and the *T* cells, which originate in the bone marrow but complete their development in the thymus gland. *B* cells are the source of antibodies: proteins that bind to antigens and aid in their removal or destruction. When a *B* cell recognizes an antigen, which may be circulating in the blood or the lymph or displayed on the surface of an infected cell, it becomes activated. It divides to produce an enlarged clone of *B* cells, all bearing antibody molecules on their membrane that act as specific receptors for the antigen. Some of the cells, known as plasma cells, actively secrete antibody. Other, longer-lived *B* cells are one root of the immunity that forestalls recurrence of



INFECTED *T* CELLS (*left*) contrast with normal *T* cells (*right*) because the AIDS virus causes the membranes of infected cells to fuse, yielding multinucleated complexes known as syncytia. The virus primarily infects *T*4 cells, which look like other *T* cells but differ from them in function and biochemical markers. Fusion of infected cells with normal ones might spread virus to other tissues.



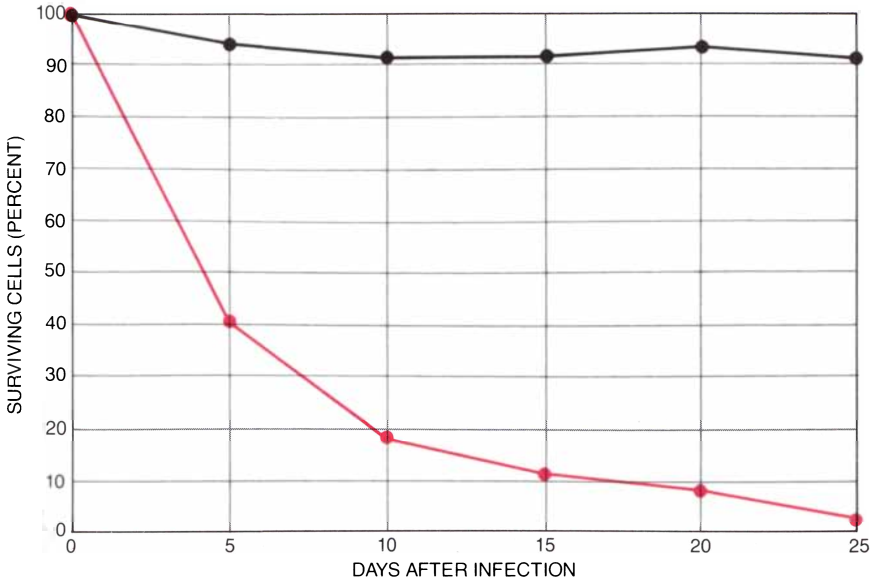
RATIO OF *T*-CELL TYPES is inverted in AIDS. In normal blood the ratio of *T*4 cells (*blue*), which induce the proliferation of lymphocytes and help other cells to mount an immune defense, and *T*8 cells (*yellow*), which destroy infected tissue and shut down the immune response, is about two to one (*left*). In the blood of AIDS patients a drastic decrease in the number of *T*4 cells leaves *T*8 cells as the predominant type (*right*). The computer-generated graphs show results from fluorescence-activated cell sorting, a technique in which antibodies specific for cell-surface proteins are used to label the *T*4 and *T*8 molecules with dyes that fluoresce in different colors. The cells then pass through automated equipment that counts and separates them. In the lower graphs the vertical axis corresponds to the number of cells and the horizontal axis to fluorescence intensity. The middle representations combine the graphs, plotting cell number against fluorescence intensity for both dyes at once. Unlabeled peaks represent cells bearing neither *T*4 nor *T*8 markers.

many infections. These memory cells remain in circulation for years, ready to mount a swift response to the antigen if it challenges the body again.

The *T* cells are more complex in classification and function. Most *T* cells cannot recognize free antigen circulating in the blood or the lymph, and they can respond to antigen on a cell surface only under particular conditions. The foreign substance must be displayed in conjunction with one of the host cell's own proteins: molecules coded for by the segments of DNA making up the Major Histocompatibility Complex (MHC). For an immune response to ensue the antigen receptor on the surface of a *T* cell must simultaneously recognize the antigen and the MHC protein.

Once a *T* lymphocyte has recognized antigen it performs a function that depends on its subclass. Only one kind of *T* cell actively defends the body: the cytotoxic *T* cells, which destroy infected, foreign or malignant cells by lysing them (disrupting their cell membrane). The other kinds of *T* cells modulate the immune response by secreting messenger proteins or by direct contact with the participating cells.

Inducer *T* cells trigger the maturation of *T* lymphocytes from precursor forms into functionally distinct cells. Helper *T* cells are a precondition for the action of other *T* cells and most *B* cells; having recognized a specific antigen, they enable cytotoxic *T* lymphocytes to destroy cells bearing the antigen and *B* lymphocytes to secrete appropriate antibody. The fourth kind of *T* cells, suppressor *T* cells, dampen the immune response of *B* and *T* cells, in effect shutting down the immune defenses several weeks after an infection activates them. Helper and suppressor *T* cells interact in complex ways: they have opposite effects on cytotoxic cells, and in shutting down the immune response suppressor cells also turn off helper cells.



SURVIVAL OF *T* CELLS in culture after exposure to the AIDS virus differs by cell type. The virus has little effect on the number of *T*8 cells (*black*). It causes the number of *T*4 cells, its preferred host, to decline dramatically (*color*). The virus affects the replication of infected cells. It also prevents infected cells from displaying the *T*4 marker protein on their surface membrane, and a protein from the viral envelope binds to and masks the *T*4 marker when it is displayed. The last two effects compound the measured decline of *T*4 cells.

*T*4 and *T*8

The *T* cells thus make up four subsets on the basis of function, but on the basis of biochemical markers on their surface there are two main kinds: *T*4 cells, which have helper and inducer roles, and *T*8 cells, with suppressor and cytotoxic functions. *T*4 and *T*8 cells are also set apart by the kind of MHC protein that must be associated with an antigen if they are to recognize it. *T*8 cells recognize antigen in the context of Class I MHC proteins, a kind of molecule that is present on the surface of all nucleated cells. Hence a cytotoxic *T*8 cell ordinarily can kill any infected cell that carries an antigen for which the *T*8 cell is specific. *T*4 cells, in contrast, respond to antigen that is associated with Class II MHC proteins, which are found primarily on the surface of specialized cells known as antigen-presenting cells.

Antigen-presenting cells and natural killer cells are the other major actors in the immune response. Chief among the former are the macrophages (scavenger cells that develop from monocytes, a kind of white blood cell) found in the skin and other tissues. The Langerhans cells of the skin and the dendritic cells of the blood, lymph nodes and spleen also present antigen. Macrophages function by engulfing a virus or other intruder, enzymatically breaking down its proteins in a highly specific way and displaying the antigenic protein fragments on the cell membrane together with Class II MHC proteins. Macrophages thereby prepare the antigen for recognition by *T*4 cells. Natural killer cells, a peripheral arm of the immune system, kill virus-infected cells and tumor cells spontaneously, without directly interacting with lymphocytes or recognizing antigens.

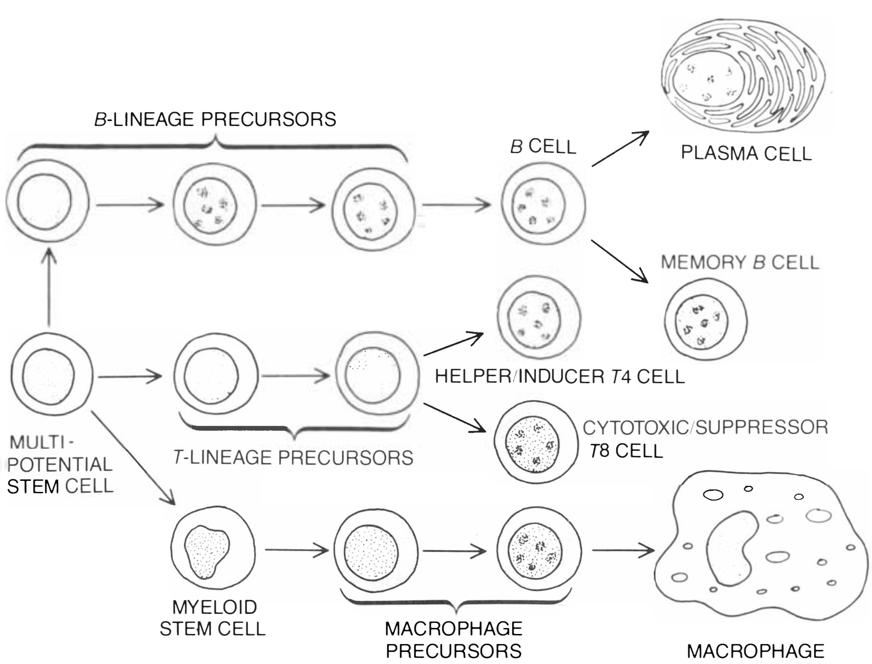
Responses to Infection

An invading virus ordinarily evokes a complex interplay of these cellular elements. To appreciate the intricacy of the phenomenon, it is worth tracing the immune response to an ordinary viral infection such as measles. Virus-infected cells secrete proteins known as interferons, which stimulate a first line of defense against the measles infection: natural killer cells. Their activity peaks within a day or two after infection. The next combatants are the macrophages, which engulf and degrade the virus. They thereby set in motion the *T*-cell response.

The macrophages stimulate the immune response not only by processing the viral antigens into a form that helper and inducer *T* cells can recognize but also by secreting soluble proteins known as monokines. One of the monokines secreted by certain macrophages is gamma-interferon; another is interleukin-1. Interleukin-1 has the ability to activate *T* cells that have recognized the viral antigen on a cell surface, preparing them to differentiate and divide. Together the interleukin-1 and gamma-interferon cause the fever and malaise that accompany measles and most other viral infections.

Activated *T* cells themselves produce soluble factors, which are known as lymphokines. It is a lymphokine called interleukin-2, secreted by *T*4 and *T*8 cells, that mediates the next step in the immune response. Under the influence of interleukin-2, *T* cells that have been stimulated by antigen and interleukin-1 proliferate into enlarged clones of mature cells: cytotoxic, suppressor and helper *T* cells. In a response that peaks about a week after infection, the cytotoxic *T* cells lyse measles-infected cells; the elimination of infected cells at this stage can bring the infection to a rapid halt.

The suppressor cells, whose number



CELLULAR PARTICIPANTS in the immune defense diverge from common precursors: the multipotential stem cells found in bone marrow. *B* cells develop through various stages in the bone marrow and then are released into the bloodstream or lymph. Contact with antigen stimulates them to differentiate into plasma cells, which secrete antibody, and memory cells, which are responsible for lasting immunity. The precursors of *T* cells migrate to the thymus gland to mature. After their release into the blood or lymph *T* cells take on distinct biochemical and functional identities as *T*4 cells, with helper and inducer roles, and *T*8 cells, with cytotoxic and suppressor roles. Macrophages begin developing in the bone marrow as myeloid stem cells, which also give rise to several other classes of white blood cells (not shown). Macrophage precursors circulate in the blood; mature macrophages migrate to the skin, the spleen and the lymph nodes, where they are most likely to encounter antigen.

increases more slowly than that of cytotoxic cells, help to shut down the *T*-cell response after several weeks. Following suppression a population of memory *T* cells persists, probably for life. Memory *T* cells, like memory *B* cells, mediate recall reactions: accelerated responses to subsequent encounters with the antigen.

Central to this *T*-cell activity are the helper and inducer *T*4 cells. Without their influence, exerted through lymphokines or through direct contact, neither the cytotoxic nor the suppressor cells could function. *T*4 cells also influence other aspects of the immune defense. The interleukin-2 they secrete bolsters the natural killer cells, and they produce gamma-interferon (considered a lymphokine when it is secreted by *T* cells), which stimulates macrophages in their role of engulfing virus and presenting antigen.

*T*4 cells are also vital to the production of antibody, the second and less important phase of the body's initial defense against a viral infection such as measles. (Antibody plays a larger role in the phenomenon of lasting immunity.) To respond to the measles antigen the *B* cells require a signal from the helper *T* cells, in the form either of lymphokines or of direct contact. The antigen-specific *B* cells then multiply into an enlarged clone of antibody-secreting plasma cells and a population of memory *B* cells. The plasma cells secrete two main classes of antibodies to the measles virus; their levels peak respectively one week and three weeks after infection, before suppressor *T*8 cells (also aided by *T*4 cells) dampen the *B*-cell response.

Foiling Immunity

Just as surely as these mechanisms defeat measles, they would block infection with the AIDS virus if it did not have a counterstrategy. The AIDS virus is not unique in its ability to avoid destruction by the immune system. Certain viruses, such as the agent of caprine arthritis encephalitis (CAE), a degenerative disease of goats, do so by triggering an inappropriate immune response, for example by eliciting the production of antibodies that do not block the infection.

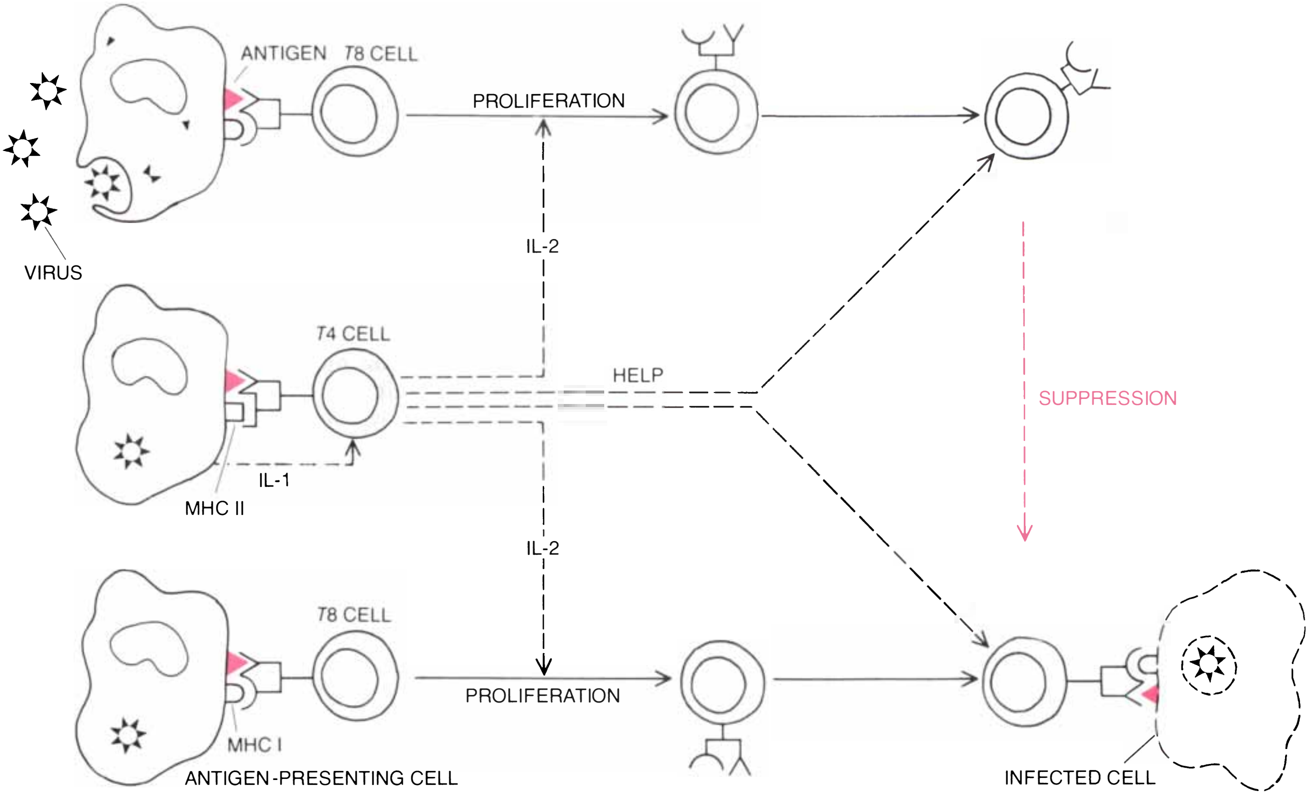
A virus could also evade the immune system if its proteins changed frequently because of mutations, causing any antigen-specific immune response to miss its mark. Sheep with a disease known as visna at first produce antibodies that can counteract the viral agent, but the virus mutates in the course of an infection, rendering the antibody ineffective. The AIDS virus, which resembles the visna virus in some genetic material, may profit from such variability, which is called antigenic drift. The virus is known to undergo steady genetic change, perhaps because of its very rapid and relatively inaccurate replication. Antigenic drift may account in part for the finding that the antibodies AIDS victims are able to produce have little neutralizing effect when they are tested in vitro against the virus.

The AIDS virus has little need for evasion, however; it avoids destruction by preemptively destroying the immune system. The phenomenon has parallels in other viral infections. Even measles temporarily weakens immunity, an effect noted in 1908 by Clemens von Pirquet. In people who have been exposed to tuberculosis and have acquired resistance to the disease the injection of a small amount of tubercle proteins under the skin produces an observable reaction. Von Pirquet noted that sensitive individuals frequently lost their skin-test reactivity during a bout of measles, although the reactivity returned later.

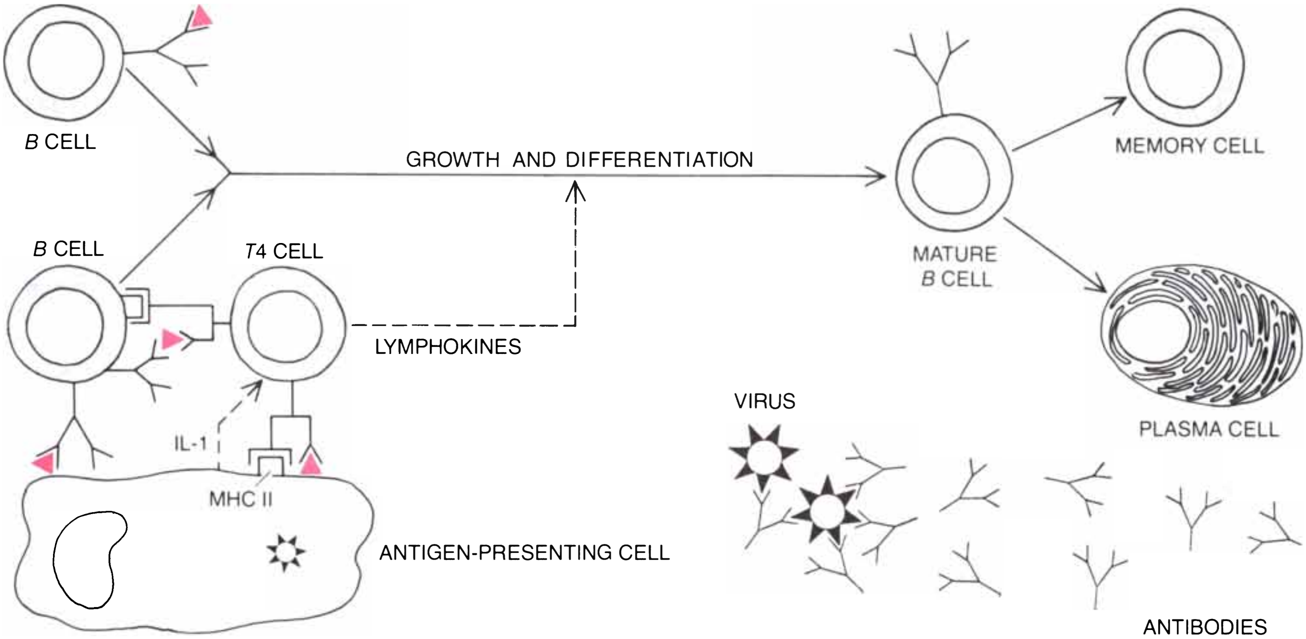
In contrast to this temporary loss of reactivity, certain viruses that infect animals produce a lasting depression of immune response. Many are retroviruses, a group that takes its name from a unique feature of its members' life cycle. In the usual biological sequence genetic information, carried on DNA, is transcribed into RNA, which is then translated into the proteins necessary for life. As retroviruses infect a cell they reverse the sequence: their genetic code, carried on RNA, is transcribed "backward" into DNA. In some cases the DNA is integrated into the host cell's chromosomes in the form of a sequence known as a provirus. Later the host cell transcribes the viral genes and synthesizes the proteins they encode, which are assembled into new viruses. Retroviruses can have a profound effect on the genetic makeup and therefore the properties of the cells they infect; many are known to cause cancers. Knowledge of the lasting immunosuppression produced by some animal retroviruses led to the suspicion, confirmed by the discovery of LAV and HTLV-III, that the cause of AIDS is a retrovirus.

*T*4-Cell Depletion

In the case of the AIDS retrovirus, immunosuppression results from viral infection of *T*4 lymphocytes, which in their role as inducer and helper cells orchestrate much of the immune re-



***T*-CELL COLLABORATION** leads to the destruction of a virus-infected cell. An antigen-presenting cell such as a macrophage ingests virus, breaks it down and displays the antigenic viral proteins (*color*) on the cell membrane together with a Class II MHC molecule (*MHC II*), one of the macrophage's own proteins. A *T*4 cell becomes activated when it binds simultaneously to the antigen and to the MHC protein; interleukin-1 (*IL-1*), a protein secreted by the macrophage, also plays a role in activating the *T*4 cell. The *T*4 cell then secretes interleukin-2 (*IL-2*). Interleukin-2 induces *T*8 cells that have also recognized antigen (together with a Class I MHC protein) to proliferate. Some of the *T*8 cells go on to kill infected cells displaying the viral antigen. Later other *T*8 cells suppress this cytotoxic response (*colored arrow*), thereby turning off the immune defense when its job is done. To fulfill their cytotoxic and suppressor roles the *T*8 cells require help from the *T*4 cells, delivered either through soluble proteins or through direct contact with the *T*8 cells.



***T*4 CELLS HELP *B* CELLS** to secrete antibody against a viral antigen. A *T*4 cell is activated by interleukin-1 (*IL-1*) after recognizing antigen together with a Class II MHC protein (*MHC II*) on a macrophage or other antigen-presenting cell. The *T*4 cell then binds to a *B* cell that has also recognized antigen on an antigen-presenting cell. Contact with the *T*4 cell stimulates the *B* cell to mature, multiply and differentiate into a clone of memory cells and a clone of plasma cells, which secrete antibody; antibody molecules bind to the virus, surrounding and inactivating it. Lymphokines secreted by the *T*4 cell aid maturation. *B* cells can also recognize free antigen that circulates in the blood or the lymph (*as shown at top left*), but they still require *T*-cell help to mature, grow and differentiate.

sponse. The loss of such cells from the blood, lymph nodes, spleen and other tissues in which they are normally concentrated is one of the most striking and consistent findings in AIDS patients. Ordinarily *T*4 cells make up from 60 to 80 percent of the circulating *T*-cell population; in AIDS they can become too rare to be detected. Many viruses kill the cells they infect, usually by rupturing the cell membrane. In culture the AIDS virus instead appears to alter and ultimately slow the growth of infected *T*4 cells, while other kinds of *T* cells continue to multiply normally. In time the *T*4 cells are selectively depleted, although a small number may remain, harboring the virus in a latent state. Recent work by Robin Weiss of the Institute of Cancer Research in London shows that in surviving cells the AIDS virus can mask the *T*4 marker on the cell surface or prevent its display. As a result the *T*4-cell decline appears to be even more dramatic than it is.

The reduction of the *T*4-cell population has consequences that reflect the cell's central place in the immune system. Lacking *T*4-cell help, *B* cells are unable to produce adequate quantities of specific antibody to the AIDS virus or to any other infection. The cytotoxic *T*-cell response is similarly hampered. Suppressor *T* cells cannot fulfill their role either. The *B* cells of AIDS patients, for example, continuously secrete large amounts of nonspecific immunoglobulin (the class of proteins to which antibodies belong); they never receive the *T*-cell signal that ordinarily would shut them down.

With the loss of *T*4 cells the level of interleukin-2 falls, slowing the clonal expansion of mature *T* cells, which is normally induced by the lymphokine. The reduced production of interleukin-2 and gamma-interferon depresses the activity of natural killer cells and macrophages, which these proteins normally stimulate.

The propensity of the AIDS virus for infecting a single kind of cell sets it apart from other retroviruses, which tend to affect a range of cells. Recent work accounts for the specificity of the AIDS virus. Weiss and an independent group led by David Klatzmann at the Pasteur Institute showed that a region of the cell membrane associated with the *T*4 marker, the protein that distinguishes *T*4 cells from other lymphocytes, acts as a receptor for the virus. The region serves as an initial attachment point for the virus as it infects the lymphocyte.

The virus's preference for the *T*4 cell is not absolute, however; it is likely that macrophages, blood platelets and *B* cells serve as reservoirs of the virus. Infection of *B* cells, for example, may explain their continuous secretion of immunoglobulin. Other cells outside the blood may also serve as reservoirs: the endothelial cells lining the blood and lymphatic vessels, the cells of epithelium (skin and related tissues), the glial cells of the nervous system and nerve cells themselves. The ability of the AIDS virus to infect the central nervous system may account for the psychosis and brain atrophy that is common in patients. Cells outside the blood may lack the surface proteins that would enable the virus to invade them directly, but they may become infected when diseased *T*4 cells or macrophages fuse with them.

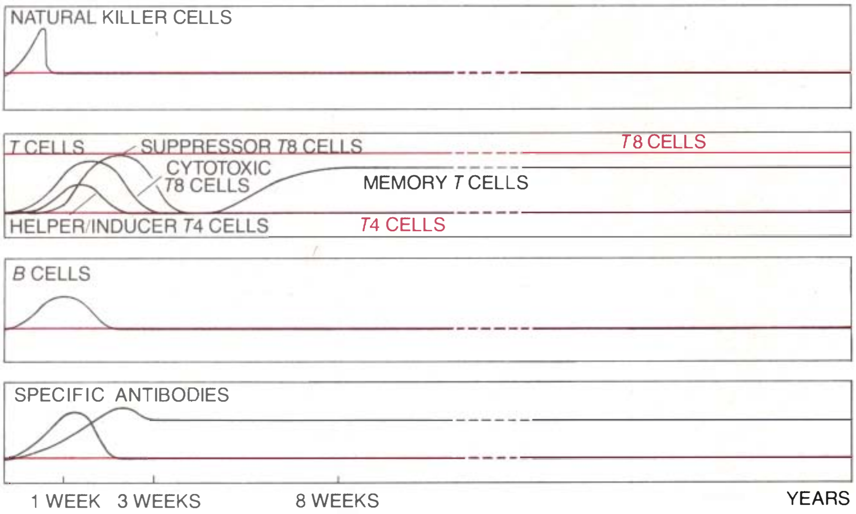
*T*4 cells may be most susceptible to infection when they have been stimulated and their numbers increased by chronic parasitic or viral infections. Infection with hepatitis *B*, Epstein-Barr virus or cytomegalovirus is common among several of the groups at risk for AIDS.

Soluble Suppressor Factor

A decrease in the number of *T*4 cells cannot account for the full extent of the immune defects seen in AIDS patients. In the early stages of the disease, for example, patients may still have a normal number of *T*4 cells, and yet their immune defenses are already severely weakened. Some workers have proposed that the virus elicits the production of anti-*T*4-cell antibodies, which besides killing *T*4 cells would also inhibit surviving *T*4 cells. Others suggest that cells previously infected by hepatitis *B* and carrying the viral genes in their DNA not only would be more susceptible to infection by the AIDS virus but also would respond differently to the infection.

The parallels between the AIDS virus and other immunosuppressive retroviruses suggested that a further change in *T*4-cell function occurs. Cats infected with feline leukemia virus (FeLV), a retrovirus, often die not from the leukemia but from other diseases, including some of the same opportunistic infections seen in AIDS. In 1978 Lawrence E. Mathes and Richard G. Olsen of the Ohio State University College of Veterinary Medicine found one basis of the immune suppression that leaves the cats open to such diseases.

Mathes and Olsen reported that a part of the FeLV protein envelope inhibits the response of cat *T* cells in culture to a substance that ordinarily causes the cells to proliferate, as they would in a normal immune response. The workers showed that the envelope molecule, which protrudes through the surface of an FeLV-infected cell, impairs the immune response of a living



IMMUNE SUPPRESSION is evident in a comparison of the responses to a measles infection typical of an AIDS patient (*color*) and of an individual with an intact immune system (*black*). The curves plot the activity of cells or antibodies over time. In a normal individual natural killer cells constitute the earliest response; their activity peaks within one or two days. The next response to intensify is that of *T* cells; their varieties (helper/inducer *T*4, cytotoxic *T*8 and suppressor *T*8), proliferate at different times, with the suppressor *T*8 cells being the last to peak. After suppression of the *T*-cell response a population of memory *T* cells persists, conferring immunity against a recurrence of the infection. Meanwhile *B* cells multiply and secrete antibodies. The various kinds of antibody reach their highest levels between one week and three weeks after infection. Suppression dampens the *B*-cell activity, leaving a small population of memory *B* cells, which maintain high levels of antibody for years. In AIDS patients *T*8 cells chronically outnumber *T*4 cells. Because of the reduced number and altered function of *T*4 cells, specific immune responses to infection do not ensue.

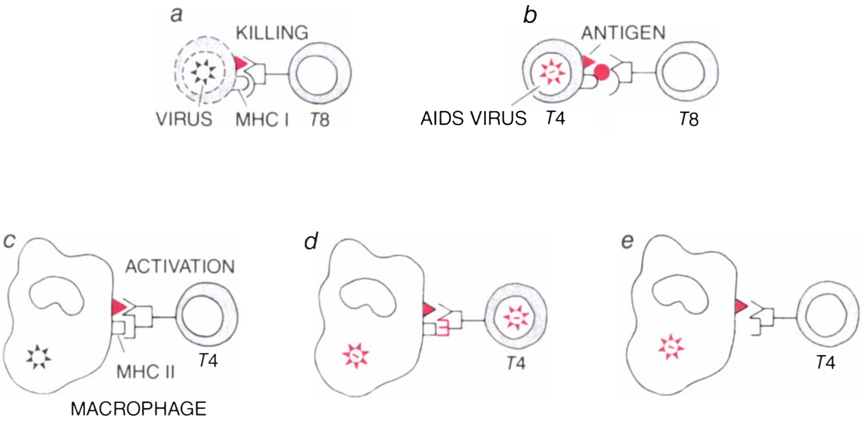
animal. They injected the purified protein, called p15E, into cats and found that the treated animals were more likely than normal cats to develop cancer after exposure to feline sarcoma virus, a tumor virus. Since then evidence has accumulated from other laboratories suggesting that many other mammalian retroviruses have protein components that produce immune deficiencies.

Such retroviral immune deficiencies share many of the features of AIDS, including a breakdown of all the responses that depend on *T*4 cells: antibody production, cytotoxic *T*-cell activity, *T*-cell proliferation induced by interleukin-2 and macrophage stimulation by gamma-interferon. Henry G. Kunkel and I, working at Rockefeller University, hypothesized that a protein factor might also play a part in AIDS. In 1982 we showed that blood cells cultured from patients suffering from AIDS release a factor that can inhibit certain immune responses. The substance, which we called soluble suppressor factor, blocks *T*-cell-dependent immune responses such as the production of specific antibody and *T*-cell proliferation both in vitro and when the factor is injected into mice.

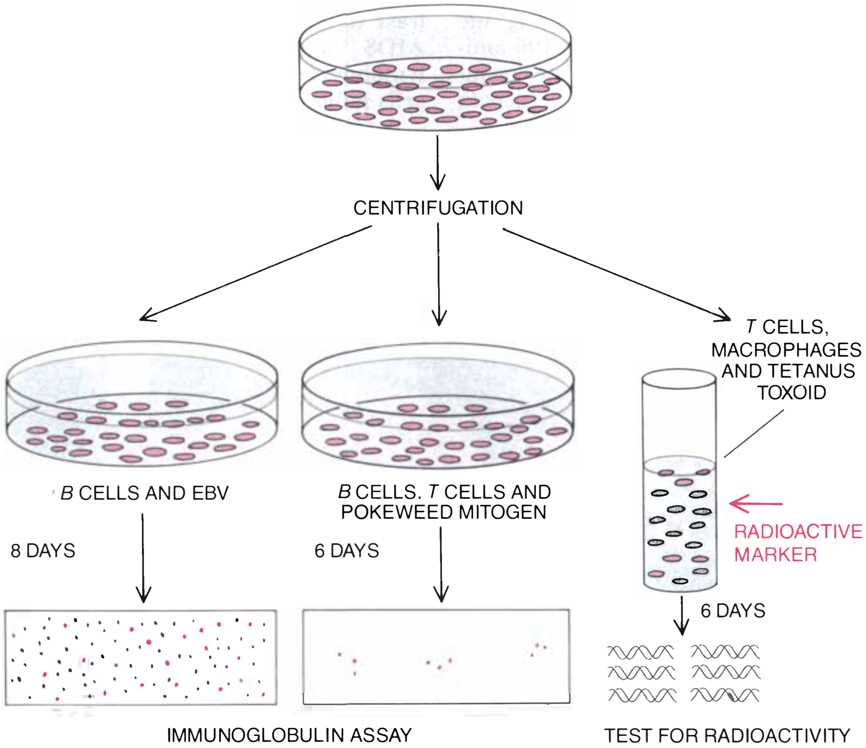
We went on to demonstrate that the source of soluble suppressor factor was infected *T*4 cells. We took *T*4 cells from a patient who was producing large quantities of the factor and fused them with normal *T* cells to immortalize them. The resulting clones of hybrid cells, we found, produced the suppressor factor in abundance. Later, working with Phillip D. Markham in Gallo's laboratory, we exposed normal *T*4 cells to the AIDS virus and showed that infection triggers secretion of a suppressor factor.

In an effort to determine whether our soluble suppressor factor is related to p15E, George J. Cianciolo and Ralph Snyderman of the Duke University Medical Center compared the sequences of amino acids making up the envelope proteins of the AIDS virus with those of p15E. They also included in the comparison another retrovirus with an affinity for human *T* cells: HTLV-I. A 26-amino-acid stretch of a protein in the HTLV-I envelope matched p15E in 73 percent of its amino acids. An envelope protein of the AIDS virus also bore a resemblance, looser but still close, to a region of p15E. One hypothesis, then, is that soluble suppressor factor originates in the envelope of the AIDS virus.

Malcolm A. Martin and his coworkers at the National Institute of Allergy and Infectious Diseases suggest another possible origin for soluble suppressor factor. They found that the



DISRUPTION OF ANTIGEN RECOGNITION may be one way the AIDS virus confounds the immune system. *T* cells must recognize antigen in conjunction with an MHC protein. Under normal circumstances a cytotoxic *T*8 cell destroys a virus-infected cell after simultaneously recognizing the viral antigen and Class I MHC molecule on the cell's surface (*a*). Hypothetically the AIDS virus could disrupt this process, protecting infected cells (mostly *T*4 cells) from recognition and destruction, by coding for an altered Class I MHC protein (*b*). The virus may also make an infected *T*4 cell itself unable to recognize antigen. *T*4 cells must bind both to antigen and to a Class II MHC molecule, displayed together on a macrophage (*c*), before they can induce *T*-cell proliferation and help the immune response. The virus may disrupt the *T*4 cell's receptor for MHC protein (*d*), causing recognition to fail. It could also incapacitate *T*4 cells by infecting macrophages and causing them to display reduced amounts of Class II MHC proteins, so that a *T*4 cell could not bind to antigen (*e*).



SOLUBLE SUPPRESSOR FACTOR produced by infected *T*4 lymphocytes from patients with AIDS or AIDS-related complex inhibits immune responses that depend on *T* cells. In the author's laboratory white blood cells from AIDS patients were grown in culture (*top*) and then spun in a centrifuge. The supernatant, or cell-free liquid, was added to three cultures of healthy cells. One contained *B* cells and Epstein-Barr virus (*EBV*), which ordinarily induces *B* cells to proliferate and secrete immunoglobulin without *T*-cell help. The response of the *B* cells to the EBV was not affected. The second culture contained *B* cells, *T* cells and pokeweed mitogen, a stimulus to which *B* cells can secrete antibody only with *T*-cell help. The *B* cells secreted little antibody. The third culture contained *T* cells, macrophages (needed to present antigen to the *T* cells) and tetanus toxoid. Under normal circumstances previously sensitized *T* cells proliferate in response to the tetanus antigen. Addition of a radioactively labeled nucleotide (a subunit of DNA) to the culture, followed by an assay of the amount of radioactivity incorporated into new DNA, indicated that the *T* cells had not multiplied. A factor in the supernatant evidently suppressed responses involving *T* cells.

region of the FeLV genome coding for p15E and the region of the HTLV-I genome coding for the matching envelope protein share a resemblance to a segment of normal human DNA. The segment is thought to represent the genetic code of an endogenous retrovirus: a retrovirus that long ago became part of the human genetic makeup. It is possible that the AIDS virus causes the host cell to express the endogenous sequence (that is, to make the protein it encodes). Rather than encoding the soluble suppressor factor as part of its envelope gene, the AIDS virus would simply induce the synthesis of the factor by the host *T*4 cell. The same mechanism could also underlie the immunosuppression seen in other infections.

Either account of soluble suppressor factor is plausible, Cianciolo and Snyderman have shown. They synthesized a peptide (a short amino acid chain) that includes a series of amino acids common to p15E, the envelope sequences of HTLV-I and the AIDS virus, and the protein encoded by the human retroviral sequence. The 17-amino-acid peptide, the workers found, is capable of inhibiting the same responses as the intact, 196-amino-acid p15E molecule. The result strongly suggests that all the sequences have immunosuppressive effects.

Besides reducing the number of *T*4 cells and causing the release of a soluble suppressor factor from those that remain, the AIDS virus also makes the surviving *T*4 cells incapable of the crucial first step in the immune response: recognition of antigen. Anthony S. Fauci and H. Clifford Lane of the National Institute of Allergy and Infectious Diseases have shown, for example, that *T* cells from AIDS patients do not respond to a common bacterial antigen, tetanus toxoid. When the workers exposed *T* cells to the antigen in the presence of macrophages (needed to process and present the antigen), the cells did not proliferate, as normal *T* cells would.

A possible explanation is that the virus somehow impairs the receptor for antigen on the surface of *T*4 cells—the molecular lock into which a key consisting of antigen and a Class II MHC protein must fit to trigger the *T*-cell response. The virus might encode a protein that is expressed on the surface of the infected cell, intruding into and disrupting the receptor mechanism. Certain retroviruses that infect mice are known to disrupt receptors on *T* lymphocytes; as a result the *T* cells cannot single out antigen. In contrast to the incapacitated *T* cells in AIDS victims, however, the mouse lymphocytes kill infected and healthy cells indiscriminately.

A related mechanism could protect the AIDS-infected cells themselves from recognition by the immune system. A cancer-causing virus called Ad-12 blocks the transcription of certain genetic sequences in infected cells, causing the cells to express reduced amounts of Class I MHC proteins on their surface. Because cytotoxic *T*8 cells can only bind to antigen together with the host cell's own Class I MHC protein, the effect hinders the recognition and destruction of infected cells. There is some evidence that the AIDS virus acts similarly, thereby protecting the *T*4 cells and the other cells it infects. Such an effect would help the virus to elude any vestiges of a functioning immune system.

AIDS VIRUS PARTICLES, distinguished by their dark core of RNA, bud from the surface of infected *T* cells. The electron micrograph, made at the Pasteur Institute in Paris in 1983, is the first image of the earliest viral isolate, LAV; the following year the same virus was identified at the National Cancer Institute under another name, HTLV-III. The AIDS virus is a retrovirus: after entering a *T*4 cell it "reverse transcribes" its genetic code, carried on RNA, into the host cell's DNA. It subsequently induces the host cell to transcribe the retroviral genes and make RNA and proteins. The viral components form new viruses, which are then released from the cell. The micrograph has a magnification of 95,000 diameters.

Treatment

Where, in this tangle of pathology, might one intervene to cure the disease or lessen its effects? The question is an urgent one, not only because of the number of cases already diagnosed but also because AIDS as it is now defined probably includes only a very small proportion of the people actually infected with the virus. In recent years persistent, generalized enlargement of lymph nodes has become increasingly common among otherwise healthy members of the groups at highest risk for AIDS. Microscopic examination of the nodes following their surgical removal often reveals the depletion of *T*4 cells and the disruption of cellular architecture that is also found in full-blown cases of AIDS. It is thought many of these people will go on to develop first the AIDS-related complex (ARC), which is marked by unexplained fever, night sweats, and weight loss or chronic cough or diarrhea, and then AIDS itself.

Results of blood tests suggest an even larger number of people may carry the infection in spite of showing no sign of disease. The presence of antibody to the AIDS virus in the blood indicates prior exposure to the virus, and in many cases it also reveals infection: most homosexual men with antibody also have infectious virus in their blood, and in some cases also in their semen, saliva and tears. The prevalence of antibody among people at risk for the disease has led to estimates that between one and two million people in the U.S. alone are infected with the AIDS virus. Some of them may never show symptoms, even though retroviral infections persist for life. For others the incubation period may vary from several months to decades. By one estimate, which is probably conservative, 7 percent of the currently infected but still healthy individuals will develop AIDS each year.

Equally troubling is the fact that these people represent a vast reservoir of carriers capable of spreading the disease. Because a small fraction of AIDS carriers do not produce detectable amounts of antibody to the virus, current procedures for screening blood donors may not entirely eliminate the risk of infection from blood transfusions.

In treating active cases of AIDS physicians usually concentrate on the clinical manifestations of the disease: the infections and cancers that develop because of the immune deficiency. Long-term success is scant: almost every AIDS patient diagnosed as having an opportunistic infection dies within four years. Other treatment strategies have attempted to restore a measure of immune function. Bone-marrow transplants and injections of white cells have been tried in an effort to replenish the immune system with healthy cells. Patients have also been given interleukin-2 and interferons to stimulate their immune system. These efforts too have met with little success so far.

Another line of endeavor is a search for a weapon against the virus itself. Workers in Gallo's laboratory have found that the drug suramin, now used to treat protozoal infections, is capable in vitro of preventing the AIDS virus from infecting and damaging *T* cells. The needed concentration of the drug could be attained in the human body. Suramin acts by inhibiting the retroviral enzyme reverse transcriptase, which mediates the transcription of the viral RNA into host DNA (the first step in viral growth). Other drugs that counteract reverse transcriptase or interfere with later steps in the viral life cycle, the production of proteins from the viral genes and the assembly of new viruses, may also prove useful. Among the substances that have been proposed or tried are ribavirin, previously used against cold and flu viruses; the novel compounds HPA-23 and phosphonoformate; 3′-azido-3′-deoxythymidine, one of a family of anticancer agents; ansamycin, which is related to a drug given for tuberculosis, and alpha-interferon. Unfortunately many drugs that inhibit reverse transcription or viral replication also hinder the growth of the host's own cells including, ironically, those of the immune system.

In my laboratory at Cornell my colleagues and I are investigating antiviral drugs that have the added effect of stimulating the host's immune system. The compounds are related to guanosine, one of the building blocks of DNA and RNA; one basis for our interest in such substances is the fact that one of them, acyclovir, has proved successful against another viral infection, herpes simplex. In vitro the substances not only inhibit the replication of the AIDS virus but also increase the response of lymphocytes from AIDS patients to compounds that induce *T*-cell proliferation. The drugs, which seem to work by inhibiting enzymes that are crucial to the synthesis of viral RNA and DNA, are promising in themselves, and they emphasize the need to investigate interactions between antiviral drugs and the immune system.

	PROTEIN NAME	AMINO ACID SEQUENCE
AIDS VIRUS	gp41	L Q A R I L A V E R Y L K D Q Q L
HTLV-I	gp21	Q N R R G L D L L F W E Q G G L
FeLV	p15E	Q N R R G L D I L F L Q E G G L
ENDOGENOUS HUMAN RETROVIRUS 4–1	p15E	Q N R L A L D Y L L A A E G G V
SYNTHETIC PEPTIDE	CKS-17	L Q N R R G L D L L F L K E G G L

AMINO ACID SEQUENCES of retroviral proteins that may suppress the immune system show close similarities. Each letter stands for one of the 20 amino acids that make up proteins; the shades of gray and color group amino acids with similar properties. Using a known immunosuppressive envelope protein from feline leukemia virus (FeLV), George J. Cianciolo and Ralph Snyderman of the Duke University Medical Center identified similar sequences in envelope proteins of the AIDS virus and HTLV-I (a virus that infects human *T* cells) and in a protein encoded by a segment of the human genome. The segment is thought to be an endogenous retrovirus. The investigators later synthesized a peptide containing amino acids shared by the sequences. The synthetic peptide suppressed immune responses as effectively as the FeLV protein, indicating that the other proteins are probably immunosuppressive also. Thus AIDS virus could produce a suppressor protein as part of its envelope or induce the cell to synthesize the protein encoded by the endogenous sequence.

An AIDS Vaccine?

For the growing ranks of the infected an effective antiviral drug is the best hope. The need for an AIDS vaccine is equally pressing. The genetic variability of the virus will hamper the search for a vaccine; samples of virus isolated from separate patients can differ by more than 30 percent in the RNA sequences encoding proteins that are thought to be key to recognition by *T* cells and antibody. A vaccine stimulates the immune system with an antigen, eliciting the production of antibody and the proliferation of memory cells. The variability of the AIDS virus means that subsequent exposure to the virus might not awaken the immunologic memory created by a vaccine. Nevertheless, it may be possible to identify invariant regions of the viral envelope to which antibody can bind effectively, and to use those regions as the basis of a vaccine. The recent development of the first effective vaccine against a mammalian retrovirus, FeLV, by Olsen and his associates has also encouraged investigators.

Research by William Haseltine and Joseph G. Sodroski at the Harvard School of Public Health suggests one approach to an AIDS vaccine. They showed that *T* cells infected with HTLV-I or the related virus HTLV-II secrete a regulatory factor that increases the transcription of the virus itself and of sequences in the host cell's DNA. The genome of the AIDS virus, they found, contains a sequence known as *tat*, which encodes a similar regulatory protein. In the case of the AIDS virus the protein might stimulate the transcription of viral genes (and perhaps the viral or host gene for soluble suppressor factor) while either inhibiting genes that stimulate replication of the *T*4 host cell or activating genes that turn off cell division. In addition to finding the region of RNA encoding the regulatory protein, Haseltine and Sodroski also identified the RNA sequence with which the protein interacts to stimulate transcription.

If the AIDS virus could be modified genetically by deletion of *tat* or the sequence with which the *tat* protein interacts, it could serve as a safe vaccine. It might elicit an immune response that would block a later infection with the unmodified virus, but the modified virus itself would not give rise to a widespread infection or deplete *T*4 cells. Alternatively, a drug that inhibited the synthesis of the regulatory protein encoded by *tat* could provide a chemical defense, as opposed to an immunologic one, against infection with AIDS. Either strategy for protection against the virus would be a boon to the tens of millions of people in this country alone who now are numbered among the AIDS risk groups.

Polar Wandering on Mars

Regions at the planet's equator seem once to have been near a pole; possibly the entire lithosphere has shifted in relation to the axis of spin. This theory explains many puzzling features and processes

by Peter H. Schultz

The Martian surface seems to present a host of contradictions. In certain areas it has eroded at a very low rate (less than one millimeter in a million years). Yet elsewhere at the same latitudes there are regions that have been heavily stripped and etched by the wind. Ancient networks of narrow valleys, once cut into the surface by flowing water and mud, suggest a warm climate, and yet such networks are found within 10 degrees of the southern polar ice cap. The polar caps themselves appear to be much younger than other Martian surfaces.

These paradoxes and many others can be resolved by one simplifying hypothesis, namely that the orientation in space of the Martian crust has not always remained the same throughout geologic time—that the crust has shifted with respect to the planet's axis of spin. If such is the case, the north and south poles (the points where the spin axis intersects the surface) would appear to have wandered over the planet's crust; certain regions of the crust that are now far from the poles would at some time in the past have been within the polar regions. If Mars has indeed undergone polar wandering, then Martian geology may have to be viewed in the context both of a dynamically changing planet like the earth and of a stable, rigid body like the moon. In this sense the Martian equivalent of plate tectonics might simply be the movement of the entire lithosphere, the solid outer portion of the planet, as one plate.

Polar wandering does take place on other planets. For example, the earth's crust shifts periodically with respect to its spin axis. This motion is not the same as the motions that change the orientation of the spin axis in space, which are called either precession or nutation, depending on whether they take place over a long or a short period, and are due to gravitational perturbations caused by other planets. In polar wandering the orientation of the spin axis in space remains the same; it is the geographic location of the poles, their location on the planet's crust, that changes.

Polar wandering is caused by instabilities due to the redistribution of mass within a planet or on its surface. The stablest alignment for a spinning object is the one in which the most massive parts of the object are farthest from its axis of rotation. On a spinning planet the distance from the spin axis to the surface is greatest at the equator. When geologic forces place a large mass at or near a planet's surface at a point far from the equator, the spinning motion of the planet becomes unstable, like the unstable motion of an improperly balanced bearing or top. If the planet is not entirely rigid, the crust shifts to move the anomalous mass toward the equator. As the crust of the planet shifts, the spin axis retains its alignment in space and the poles wander over the planet's surface.

If the lithosphere of the planet is very flexible, polar wandering does not take place, because any excess mass sinks into the surface before the crust has had time to realign itself. There can be no polar wandering on a completely rigid planet either, because the lithosphere cannot realign itself at all.

The lithosphere of Mars displays both rigid and flexible features. Roger J. Phillips, now at Southern Methodist University, and R. Stephen Saunders of the Jet Propulsion Laboratory have shown that Mars has ancient lava-filled impact basins that are partly isostatically compensated (partly absorbed by local movements, such as the rising or falling of local regions of the crust) as well as volcanic regions and tectonic features (features formed during structural deformation of the planet's crust) in which sudden changes in the mass distribution of the crust have not been compensated. The change in mass distribution that is necessary to produce polar wandering could have been provided by convection in the mantle (the rising of hot sections and the sinking of colder ones), by a redistribution of mass due to the impact of large objects or by the filling with lava of the resulting impact basins.

Mars seems, then, to be a planet that is likely to have undergone polar wandering in the course of its history. Because the surface remains relatively well preserved—there are nearly undisturbed features in some areas of the Martian surface that are more than four billion years old—and because it has not been masked by plate tectonics, some evidence that polar wandering has taken place may still exist. What evidence should an observer expect to find if polar wandering has indeed occurred on Mars? Does such evidence in fact exist? What are the implications for Martian geology if the Martian poles have wandered?

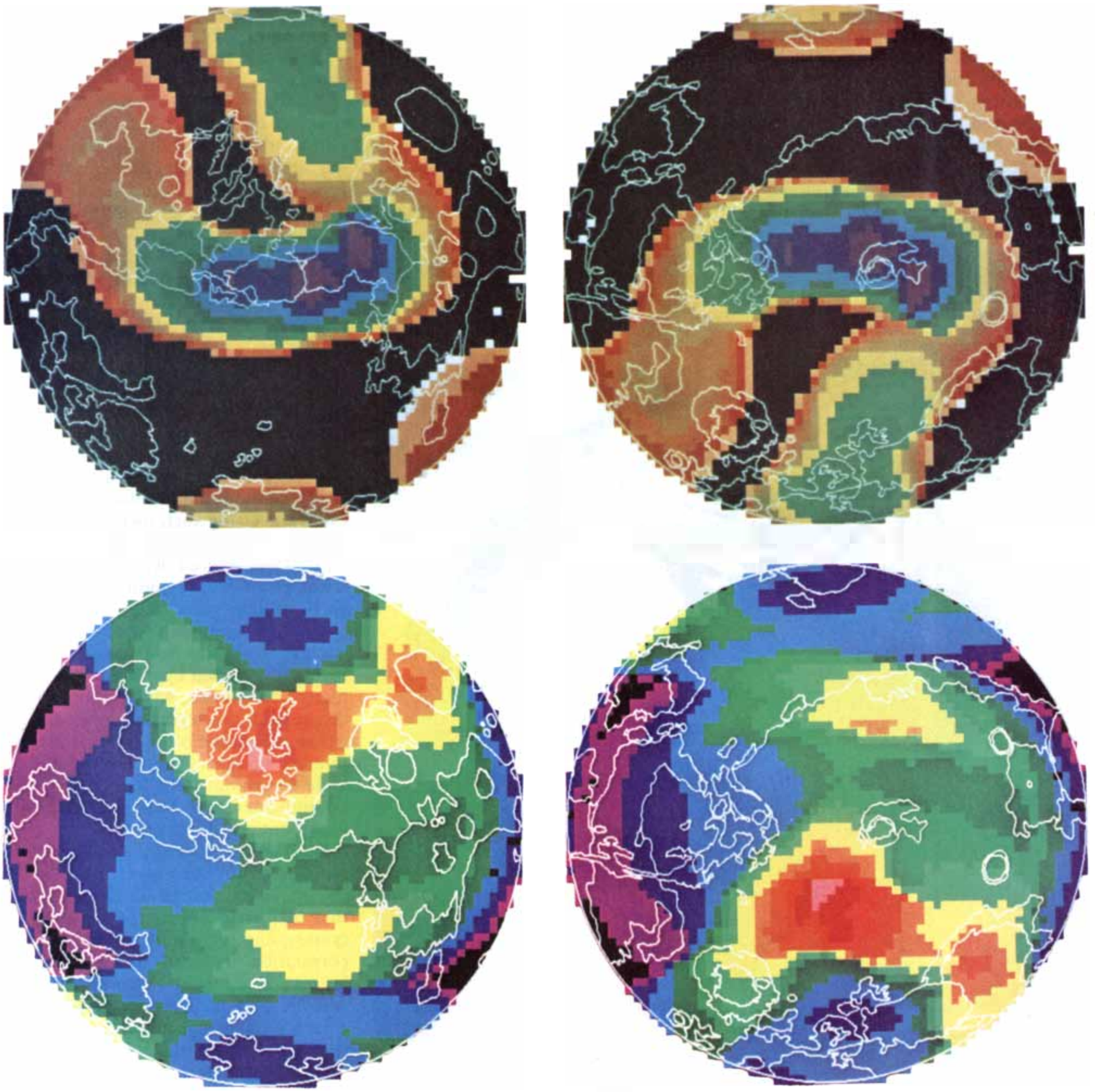
When searching for evidence of polar wandering on Mars, the investigator does not have the luxury of taking a field trip; he is faced with a geological question but is denied the fundamental tests and tools of the trade. Nevertheless, the problem can be approached by examining features and processes remotely, by means of planetary probes, much as an astronomer tries to discern the composition and history of a star without the benefit of an actual sample of stellar material.

If there has been polar wandering on Mars, the polar regions should contain the traces of processes that normally take place in warmer, nonpolar regions. Features that are normally characteristic of the poles should also be visible outside the polar regions.

The most obvious features of the Martian poles are the white polar caps. Planetary missions have revealed that the polar caps consist of a veneer of frost that expands in the fall and recedes in the spring. The frost is made up largely of condensed carbon dioxide, but in the summer the carbon dioxide layer escapes as gas into the atmosphere, leaving a small residual cap composed mainly of water ice.

The regions within 10 degrees of the poles, as well as those directly under the polar caps, are also characterized by thick layered deposits. The deposits near the south pole can be as thick as one or two kilometers and those in the north may range in thickness from four to six kilometers. The deposits' dark color indicates that they are composed largely of dust, but about 25 percent of their mass may consist of water ice. The dust is believed to derive from the large dust storms that periodically envelop the planet. Winds in the upper atmosphere carry small dust particles to the poles, where volatile molecules such as carbon dioxide and water freeze around them and they settle to the surface.

There is evidence that the deposits

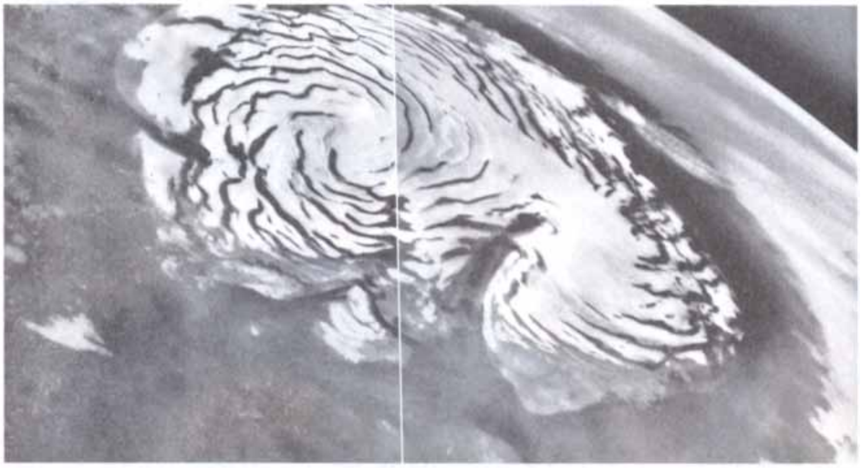


POSITION OF THE MARTIAN POLES on the planet's crust may have changed repeatedly over geologic time: as the poles retained their orientation in space the lithosphere may have shifted because of geologic disturbances, making the poles seem to wander over the planet's surface. The appearance near the equator of features that are normally associated with polar processes has made it possible to define the path that the planet's poles may have followed (*top*). The path probably followed by the planet's north pole, plotted on the basis of evidence found in the western hemisphere (*top left*), was used to infer what path the south pole might have followed in the eastern hemisphere (*top right*). The present-day north pole is at the top of both images. Bluer colors indicate regions in which the poles remained longer, and the white lines represent boundaries between distinct geologic terrains that were formed at different times in Martian history. The images at the bottom, which were based on the path depicted in the upper images, indicate areas that should contain remnants of a periglacial climate (a climate characteristic of the regions between 25 and 50 degrees of the pole). Because the poles' path curved back on itself, certain regions (*redder colors*) were within 50 degrees of the pole for much longer than other regions.

are continually undergoing processes of deposition and erosion. Old layers have been truncated and then covered over by the accumulation of newer ones. These layers of ice and dust cover preexisting topography like a gentle blanket, masking craters as large as 100 kilometers in diameter. Some of the deposits are undergoing complete destruction by wind scouring. Often, as a layer is eroded by the wind, delicate geologic features under it are revealed. These features show little evidence that they have been disturbed by the deposition of dust layers or by wind erosion, demonstrating both that the process of deposition is very gentle and that the deposits themselves may be eroded by relatively weak winds.

Farther from the poles than the layered deposits are there is an assortment of other features that are thought to be related to certain polar processes. One such polar outlier is the pedestal crater: an old crater from which the surrounding terrains have eroded, leaving a flat plateau with etched borders [*see illustration on page 98*].

Pedestal craters are most likely to form in a near-polar region, where surface ices can exist. When an impact is energetic enough to form a crater that penetrates the veneer of icy dust, the debris ejected from below (which may contain less ice and frozen carbon dioxide than the surface) surrounds the impact crater, covering and protecting the icy layers below. If there is a change in the local climate (either a temporary change, such as might occur periodically because of polar precession or the planet's changing distance from the sun, or a permanent change), the icy deposits surrounding the crater evaporate. The ice in the area under the ejected debris is shielded, however, and does not evaporate. This area is also shielded from wind erosion by the layer of debris. As a result the ice and dust surrounding the impact crater disappear, leaving the crater seemingly perched on a pedestal formed by deposits directly under it.



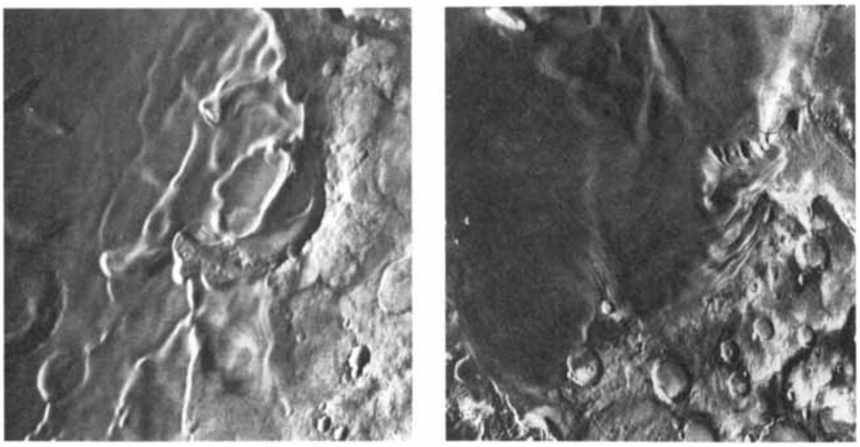
NORTH POLE of Mars lies under a white cap of ice and frozen carbon dioxide (*top*), which extends as far as 25 to 40 degrees from the pole in winter and recedes to between five and 10 degrees from the pole in summer. Wind-etched canyons (*bottom*) reveal thick layered deposits of ice and dust that lie under the polar cap. Layered deposits similar to those under the poles have been found near the equator; they may mark regions that were once polar.

The relative ages of Martian surface features, such as the polar deposits, can be gauged by comparing the number of large impact craters on them. Cratering serves as a clock: the older a surface feature is, the more time there will have been for an impact by a large object. Features that have about the same number of large craters per unit of area are assumed to be about the same age.

To approximate the absolute, or actual, age of a feature, a task more difficult and less reliable than determining the relative ages of two features, one must make some assumptions about the number of impacts that can be expected to occur within a given time. Most models for generating such assumptions are based on comparisons between Mars and the moon. The absolute age of some lunar features has been determined by examining samples returned from the Apollo missions; if one could determine the difference between the rates at which craters are formed on Mars and on the moon, it would be possible to use the relatively precise lunar dates as a scale with which to calibrate the dates of events on Mars. Using such methods, several groups of investigators have estimated that the layered deposits in the polar and near-polar regions are as young as 10 million years old.

Another approach is to count the number of layers in a deposit. If the layering is caused by periodic climatological conditions, which may in turn be caused by such periodic phenomena as the precession of the poles, then the number of layers in a deposit should yield an estimate of its age. This method indicates that the polar deposits are about a million years old.

Both the method of counting craters and that of counting layers have flaws: the deposits change so dynamically that they may destroy or mask older craters, and there may not be a direct correlation between the number of layers visible from a distant orbiting satellite and the various periodically driven climatological cycles. Another way to estimate the age of a polar region is to determine the density of large pedestal craters. The number of pedestal craters larger than one kilometer in diameter near the present poles indicates that the polar deposits could be as old as 100 million years. Three very different analyses, then, yield the conclusion that the present poles are young on a geologic time scale. Their youth could indicate that



POLAR AND EQUATORIAL LAYERED DEPOSITS bear striking similarities. The dusty deposits shown at the left lie at 215 degrees west, 73 degrees south, in the region of the present-day south pole. They cover old craters and mask most preexisting features. Equatorial layered deposits (*right*) show the same kind of blanketing, although they are more heavily eroded. The equatorial deposits shown lie at 156 degrees west, four degrees south.

polar caps appeared only late in Martian geology. Alternatively, it could indicate that these deposits have moved across the surface of Mars during epochs of polar wandering.

If the Martian pole has wandered, there should be evidence elsewhere on the planet of deposits similar to those at the present poles. Such evidence does indeed exist. Thick deposits of dusty material, and relics of such deposits, can be found in two broad zones outside the polar regions. These zones are antipodal: they are on opposite faces of the planet. The deposits show many of the processes and characteristics of today's poles, but they lie near the present-day equator. Like the polar deposits, the equatorial deposits blanket preexisting craters. The equatorial deposits are also layered, and they are easily eroded by the wind. As the wind strips away the dusty layers it reveals buried craters and channels, which do not seem to have been altered heavily either by the process of burial or by that of exhumation.

The composition of the deposits can be determined in part by data from the Infrared Thermal Mapper (IRTM) that traveled on the Viking Orbiter, which measured the temperature of the Martian surface just before sunrise. Looser, smaller grains radiate heat faster than larger grains or rocks because smaller grains have a greater ratio of surface area to mass. Terrain consisting of small-grained dust therefore tends at night to lose the heat it has absorbed during the daytime faster than terrain consisting of larger fragments. The IRTM data, as analyzed by James Zimbleman, now at the Lunar and Planetary Institute in Houston, and Hugh E. Kieffer of the U.S. Geological Survey in Flagstaff, show that the deposits at the equator, like the polar ones, are composed mainly of small dust grains.

Surrounding the equatorial deposits are features that are rarely found except near the present pole. For example, the deposits near the equator are surrounded by pedestal craters that are larger and more numerous than those around today's poles. The equatorial pedestal craters, like the polar ones, typically occur in association with other etched plateaus that have similar relief. Large craters in the regions where there are equatorial deposits typically contain mounds of eroded layered material, much as the craters around the present poles do. Perhaps most important, the equatorial layered deposits are found in locations that are directly antipodal to each other and that otherwise have very different geologic features. It was this observation that first suggested to Anne B. Lutz, now at the Computer Sciences Corporation, and me that these deposits might indicate ancient pole positions.

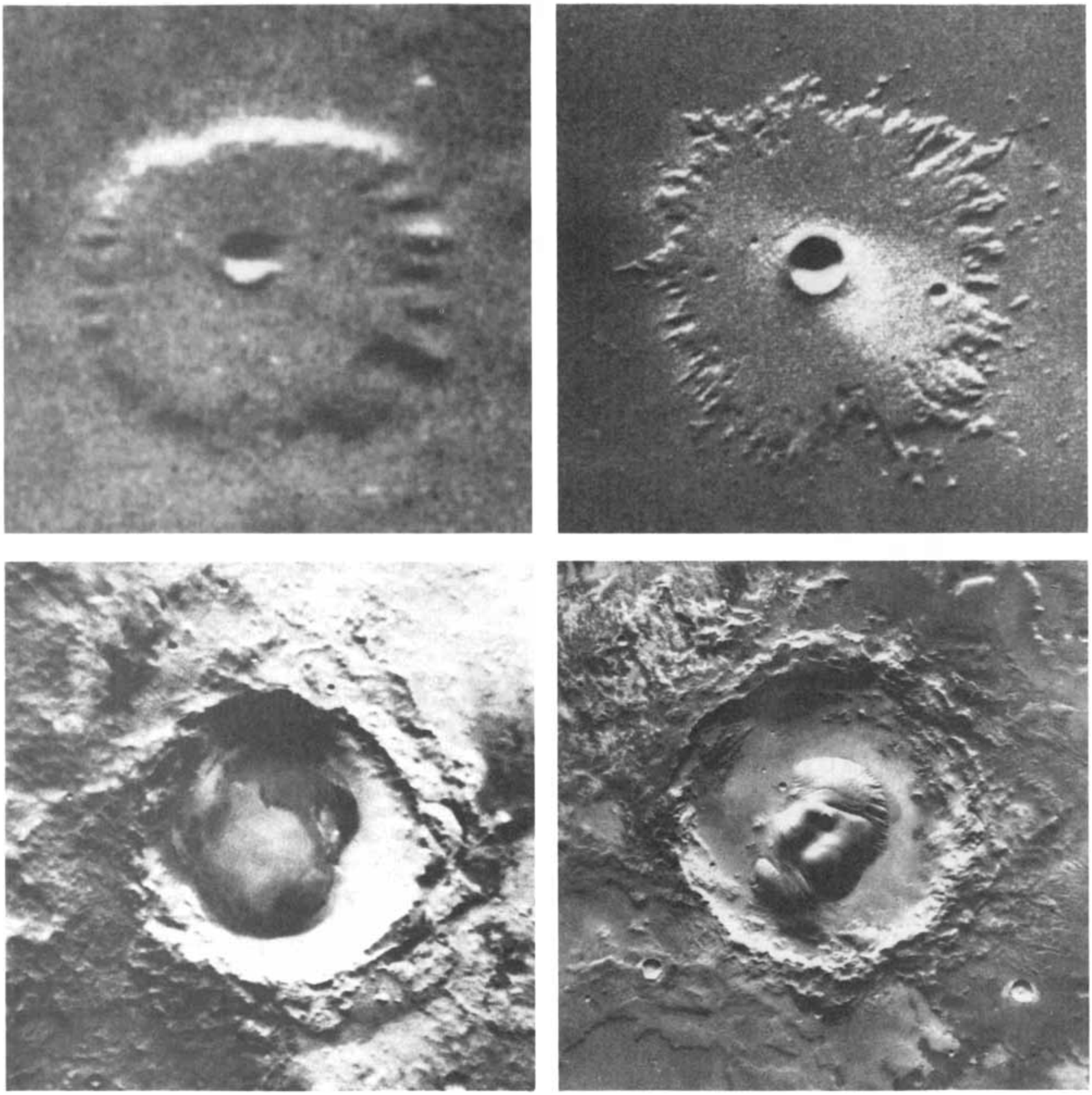
The equatorial layered deposits in the western hemisphere occur in three distinct concentrations within 10 or 15 degrees of the equator and between 145 and 215 degrees west longitude. The westernmost of these accumulations is the oldest and most eroded and the easternmost is the youngest. The best-preserved and youngest deposits resemble present-day polar terrains in both area and thickness. They lie between 145 and 165 degrees west longitude and cover nearly one million square kilometers; their maximum thickness is from about two to about five kilometers. These deposits have the distinctive undulating smooth surfaces, stripped terrains, parallel valleys and wind-scoured canyons that are characteristic of the polar deposits, but they have been much more heavily scoured by the wind. On the opposite side of the planet similar deposits blanket old cratered uplands.

The age of the equatorial deposits can be approximated by the number and size of pedestal craters and later impact craters. The deposits are all very ancient. The youngest equatorial deposits appear to be no younger than between two and three billion years old, implying they were created no earlier than the time when large volcanoes in an area known as the Tharsis region, which lies on the equator at 110 degrees west longitude, began to erupt. The adjacent, older deposits appear to extend back to a time when extensive volcanic plains were being emplaced over much of the planet's surface, which was before and during the period from about three to 3.5 billion years ago. The oldest deposits cannot be dated with certainty, but their age is probably not much greater than that of the intensive network of narrow Martian valleys, which were formed more than 3.5 billion years ago.

From a very different and unexpected source Lutz and I found additional support for the hypothesis that Mars's poles have wandered. In 1976 the Viking Orbiter relayed images of the Martian surface that showed an unexpectedly large number of elliptical craters with distinctive blankets of ejected debris. The ejecta of these craters do not surround the craters uniformly; rather, they form a pattern that resembles the wings of a butterfly [*see illustration on page 87*]. In a classic series of experiments Donald E. Gault and John A. Wedekind, then both at the Ames Research Center of the National Aeronautics and Space Administration, showed that because of the high velocity with which objects strike Mars (the impact velocities are nearly always greater than five kilometers per second), such craters form only when the angle of impact is very low; the angle between the object's path and the surface must be less than five degrees.

Objects whose orbits are centered on the sun, such as comets, would rarely strike the planet at such a low angle of incidence. Grazing craters caused by these objects should account for only about .7 percent of the Martian crater population. Such a small percentage would be consistent with the number of grazing impacts that are now visible in the lunar maria. On Mars, however, the number of grazing impacts is from five to 10 times greater than what has been predicted on the basis of theory and what is observed on the moon. Moreover, the direction of impact, which is parallel to the long axis of the crater, seems to have changed systematically with time: younger, well-preserved craters generally have trajectories aligned along paths that run east and west, whereas older craters are aligned along north-south paths.

These two observations—the large number of grazing impacts and the systematic change in the craters' orientation—indicate that the grazing impacts were not caused by randomly arriving objects that had been in orbit around the sun. The observations can be understood, however, if one assumes the orbits of ancient Martian satellites decayed gradually until the satellites struck the planet's surface. If these satellites had orbited Mars in the equatorial plane, as the satellites Phobos and Deimos do today, the geographic location and direction of impact of a crater should establish the

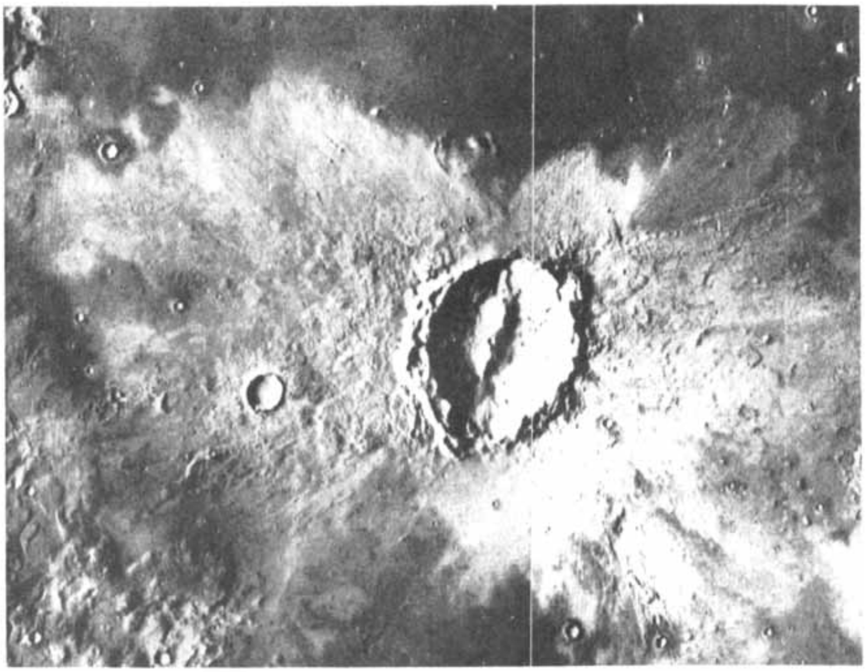


PEDESTAL CRATERS AND TRAPPED DEPOSITS, features that typically form in polar or near-polar conditions, are found not only near the present-day poles but also near the equatorial layered deposits. Pedestal craters (*top*) are old craters that seem to be perched on pedestals above the surrounding terrain. They are thought to result from impacts energetic enough to form craters that penetrate through a layer of ice and dust to the material below. The result is a layer of ejected debris that contains relatively little ice, which insulates the icy deposits under it. When the surrounding deposits melt, the ejected debris around and on top of the crater protects the material below and remains elevated. The pedestal shown at the left is 16 kilometers wide and is within 30 degrees of the north pole. The one at the right, which is about 25 kilometers across, is near the equator. Mounds of layered deposits trapped within large craters (*bottom*) are thought to form after polar layered terrains have buried preexisting craters. As the layered terrains shift in location (perhaps because of a local change in climate) thick mounds remain on the crater floors. The mound shown at the left is in one of many craters surrounding the polar deposits. The mound at the right is in a crater near the periphery of an equatorial deposit.

location of the poles at the time the satellite struck. (Several groups of researchers calculate that Phobos itself will collide with Mars in about 20 million years, forming a new east-west grazing-impact crater.)

The polar points that can be inferred from the youngest grazing-impact craters do indeed match the locations of the present-day poles. Older impacts indicate polar points that cluster far away from the present poles and quite close to the ancient layered deposits found near the equator.

The approximate age of a grazer impact can be estimated by observing how well preserved the ejecta around the crater are and by comparing the ejecta's state of preservation with that of craters of similar size lying on a surface whose age has been reliably established. On this basis it has been determined that the ages of the various polar points indicated by the grazer impacts are in close agreement with the ages deduced independently from the number and size of craters on equatorial deposits. The grazing-impact craters, then, provide an independent piece of evidence that the poles of Mars have shifted, much as frozen magnetic field directions in terrestrial rocks provide evidence for the movement of tectonic plates on the earth.



GRAZING IMPACTS produce craters that have a distinctive, elongated shape and a pattern of ejected debris that resembles the wings of a butterfly. On Mars many of these craters probably result from impacts of satellites whose orbits have decayed. Such satellites would have orbited in the plane of the Martian equator, and so the location and direction of a grazing-impact crater should provide a good indication of where the Martian equator and poles were when the satellite struck. Recent craters tend to have trajectories oriented in the east-west direction, whereas ancient craters tend to be oriented in a north-south direction. The polar points that can be inferred from grazing-impact craters are in close agreement with those inferred from other data, such as the buildup of thick, layered deposits.

The grazing impacts, the ages and locations of the layered equatorial deposits and the distribution of the features that seem to indicate ancient polar or near-polar locations make it possible to trace the path the poles may have followed as they wandered across the Martian surface. They seem not to have wandered steadily but to have changed place in rapid spurts followed by long pauses.

The north pole's most recent stable position (before its present one) seems to have been at 160 degrees west longitude and 45 degrees north latitude, an area to the northwest of Olympus Mons, the planet's highest mountain. This orientation would place the south pole near the southwestern part of an enormous impact crater called the Hellas basin, in a region called Noachis. Indeed, the Noachis region does contain numerous pedestal craters as well as many other polar outliers.

About two to three billion years farther back in time the north pole seems to have paused at 150 degrees west longitude and five degrees latitude, a location to the south of Olympus Mons and near the Tharsis volcanoes. Dates based on comparisons of crater statistics indicate that the volcanoes may have begun their massive eruptions just before the pole shifted to the north. Such volcanic activity may well have contributed to the pole's shift to the later position northwest of Olympus Mons; subsequently processes associated with the construction of Olympus Mons, coupled with continued activity in the Tharsis region, probably caused the pole to shift to its present position. Two thick layered deposits just to the west of the site near 150 degrees west longitude suggest that before the pole moved close to the Tharsis region it had probably remained relatively stable, shifting only slightly over a period of between 500 million and one billion years.

These pole positions can all be defined with some accuracy because the rate of wind erosion has been very low on Mars for the past 3.8 billion years. Consequently a reasonable record of the polar deposits and outliers has remained. Before 3.8 billion years ago, however, the erosion rate was much higher. Perhaps Mars had a much thicker and warmer atmosphere. Such a hypothesis could account for the myriad narrow valleys that seem to have been formed by flowing water and mud. Alternatively, the valleys could have been caused by an increase in the temperature of the Martian interior: if the planet's interior had warmed, trapped ice would have melted and formed runoff valleys even if the Martian atmosphere was no warmer or thicker than it is today.

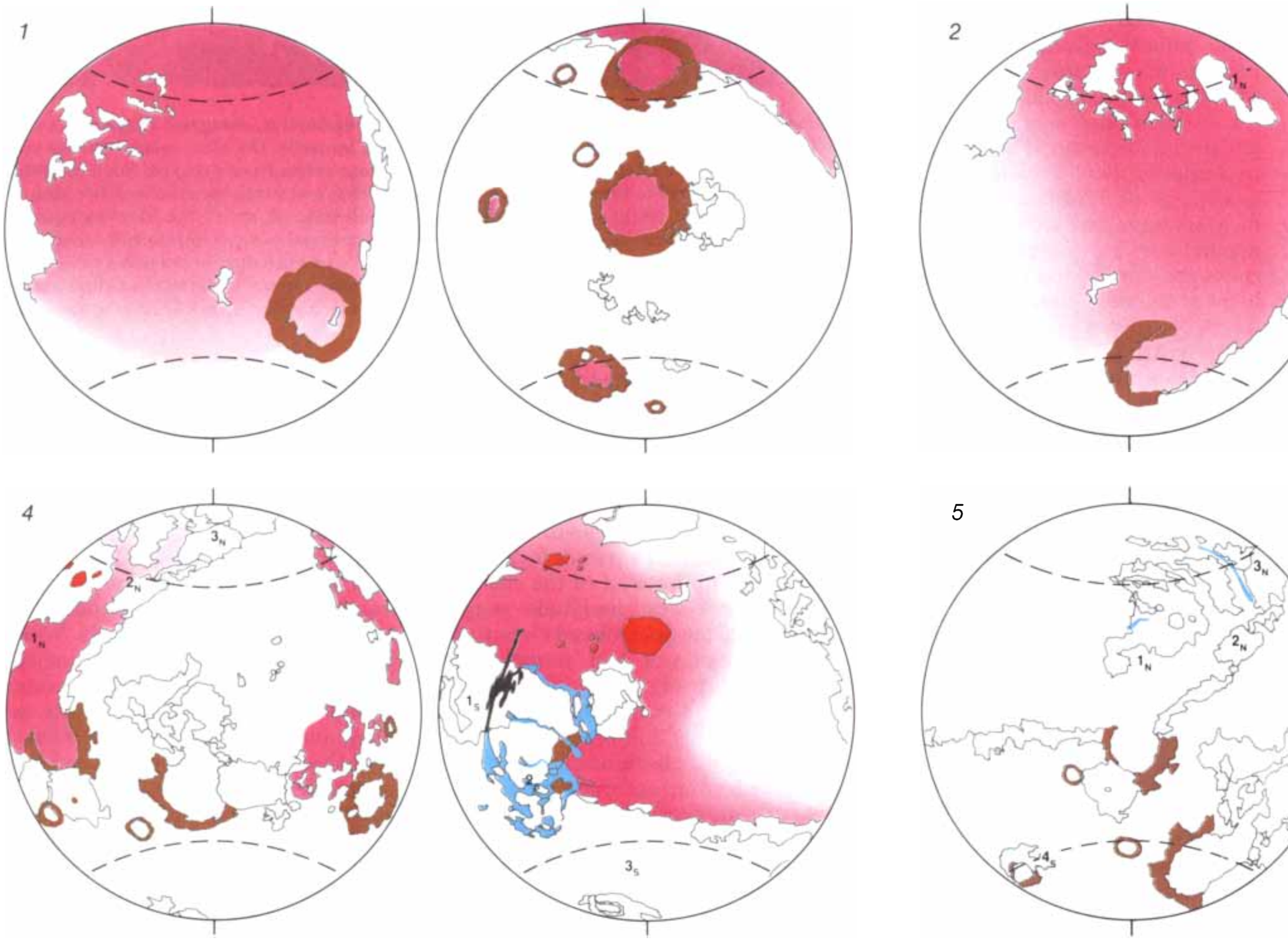
In either case—whether the early Martian atmosphere was warmer or whether the internal temperature of the planet increased—the remnant polar deposits would not have been preserved in a state resembling today's thick polar sequences (the deposits near 215 degrees west longitude have indeed been heavily reworked). When the poles shifted from these ancient positions, some of the trapped water ice and carbon dioxide would have escaped into the atmosphere and the rest would have melted and percolated into the terrains below. Pole locations that date back farther than 3.8 billion years should therefore bear features indicative of large reservoirs of trapped water and carbon dioxide. Such features might include springs, chaotic deterioration of the ground, extensive runoff systems and terrains particularly susceptible to wind erosion.

As a matter of fact, the ancient polar positions do show such features. For example, one of the oldest of the layered terrains, which lies near the equator between 200 and 225 degrees west longitude, is modified by heavy channeling. The region on the opposite side of the Martian globe is the site of extensive ground collapse, and it has large outflow channels that are characteristic of the voluminous release of water-rich sediments.

The particular route the poles seem to have taken helps to explain many anomalous features of the Martian landscape. Like the earth, Mars has a number of climate zones that are largely dependent on latitude. For example, within 50 degrees of the poles water ice on the surface is in equilibrium with water vapor in the atmosphere; such regions tend to be characterized by an assortment of features produced or enhanced by the freezing and thawing of ice. Some investigators have found that these so-called periglacial features are localized not only at certain latitudes but also within certain longitudinally defined zones. The polar-wandering hypothesis may help to explain these zones. The poles did not follow a straight path; in one region the path curves back on itself [*see illustration on page 102*]. As a result certain regions—those around which the path of the poles curved—have remained within 50 degrees of a pole throughout most of geologic time. It is to be expected that such regions would show periglacial features regardless of their present latitude.

The polar-wandering hypothesis can also explain enigmatic deposits that cover the northwest flanks of the Tharsis volcanoes. Within about 30 degrees of today's poles, frosts and dust deposits tend to accumulate on surfaces that face away from the equator and are therefore not exposed to as much sun as other surfaces. When the Tharsis volcanoes erupted, they would have been within about 30 to 40 degrees of the north pole, and the flanks that now face northwest would have been facing directly north. The greatest accumulations of frost and dust should have occurred on the oldest volcanic constructs, which are closer to the old pole position. In fact the deposits that are found on what are now the northwest flanks of the Tharsis volcanoes do become smaller with decreasing age and increasing distance from the old pole positions.

Another feature explained by polar wandering is a high scarp, or cliff, surrounding the base of Olympus Mons. Carol Ann Hodges and Henry J.



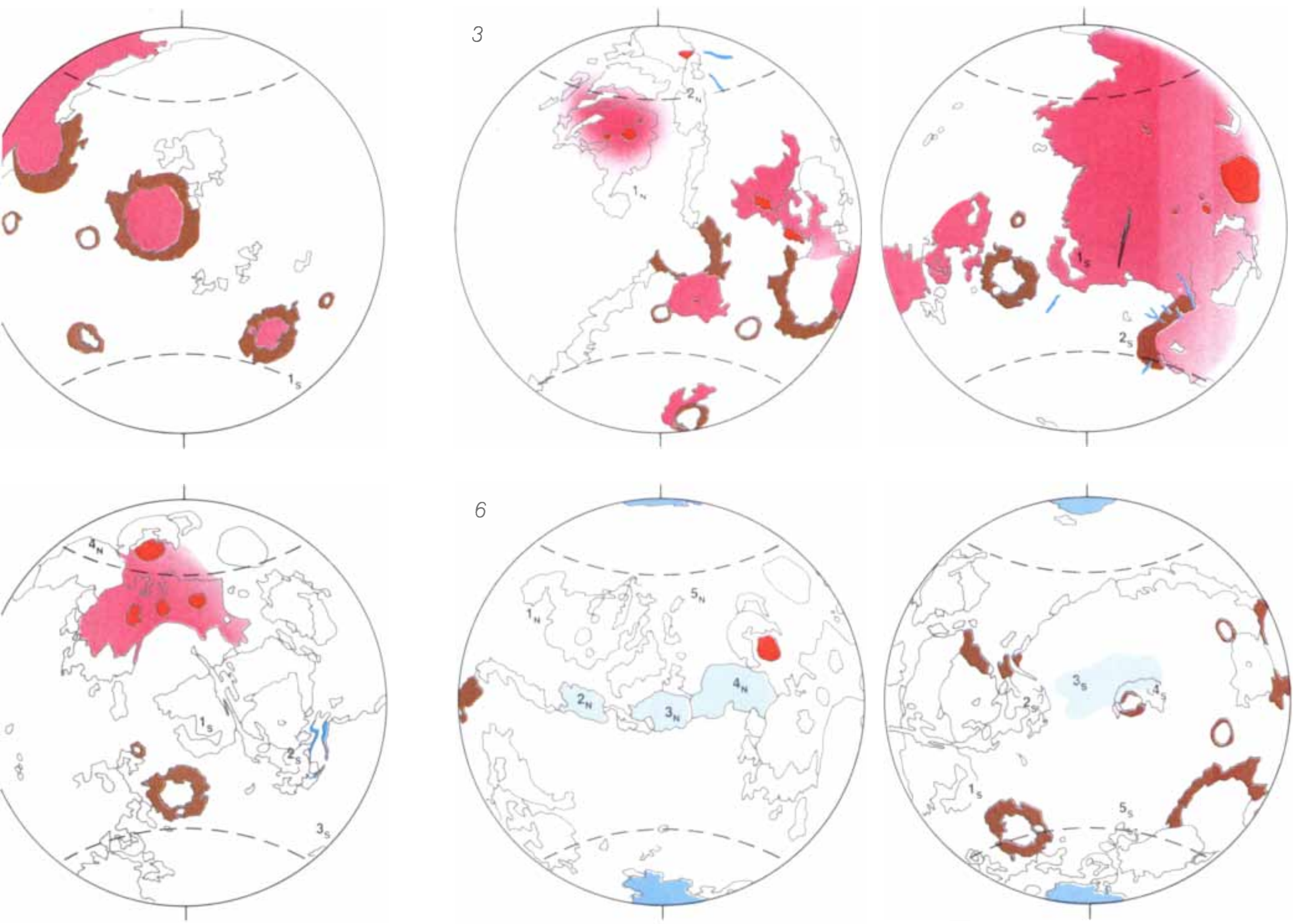
CRUST OF MARS has developed and changed in character, as well as shifting with respect to the planet's spin axis. About 3.8 billion years B.P. (before the present) (*1*) extensive lava plains (*pink*) were forming in one hemisphere (*left*), while few formed in the other. The mountainous rims of impact basins are shown in brown, and broken lines represent the boundaries of the planet's periglacial regions. During the period between 3.8 and 3.5 billion years B.P. (*2*) the crust of the planet shifted with respect to the axis of spin, probably because of the continuing creation of lava plains. Positions once occupied by the poles are represented by numbers. When the crust shifted, much of the ice that had occupied the former polar regions melted from below and seeped into the crust, where it was trapped in large subterranean pools. The period between 3.5 and 2.5 billion years B.P. (*3*) marked the beginning of a time of extensive volcanic mountain building. Volcanic mountains are shown in red. Lava plains continued to develop, in some cases filling old craters. Cutting the major lava plain in the eastern hemisphere (*right*) is the precursor of the Valles Marineris (*black*), a system of grabens, or major fractures. The grabens are believed to have formed partly because of tectonic stresses produced when the rigid crust was stretched as it passed over the planet's equatorial bulge and partly because of volcanic activity that was later to produce what are now known as the Tharsis volcanoes. During this period some of the water and ice trapped under the crust in regions that had once been polar began

Moore of the U.S. Geological Survey in Menlo Park, Calif., have noted that this scarp closely resembles the ones surrounding terrestrial volcanoes in Iceland; these appear when eruptions take place first under and then on top of ice sheets. For the analogy to be perfect, Olympus Mons, which is now near the Martian equator, would have to have been in a region of thick, ice-rich deposits when it formed. If the polar-wandering hypothesis is correct, Olympus Mons would in fact have been very close to the north pole during its initial development.

Certain tectonic features are also explained by polar wandering. Mars is not a perfect sphere: it bulges at the equator. As the planet's lithosphere rotated over the equatorial bulge, the change in the radius of curvature presumably would have produced great stresses. Polar regions approaching the equator would have been pulled apart enough to form grabens (valleys created when one block of crust drops down between two other blocks) and to promote volcanic activity. Conversely, equatorial regions approaching the pole would have undergone compressional stresses, which might have produced ridges. (Whether or not ridges are actually generated under such conditions depends on a number of factors, such as the elasticity of the lithosphere, the presence of nearby volcanic centers and the degree to which the entire globe is stressed.)

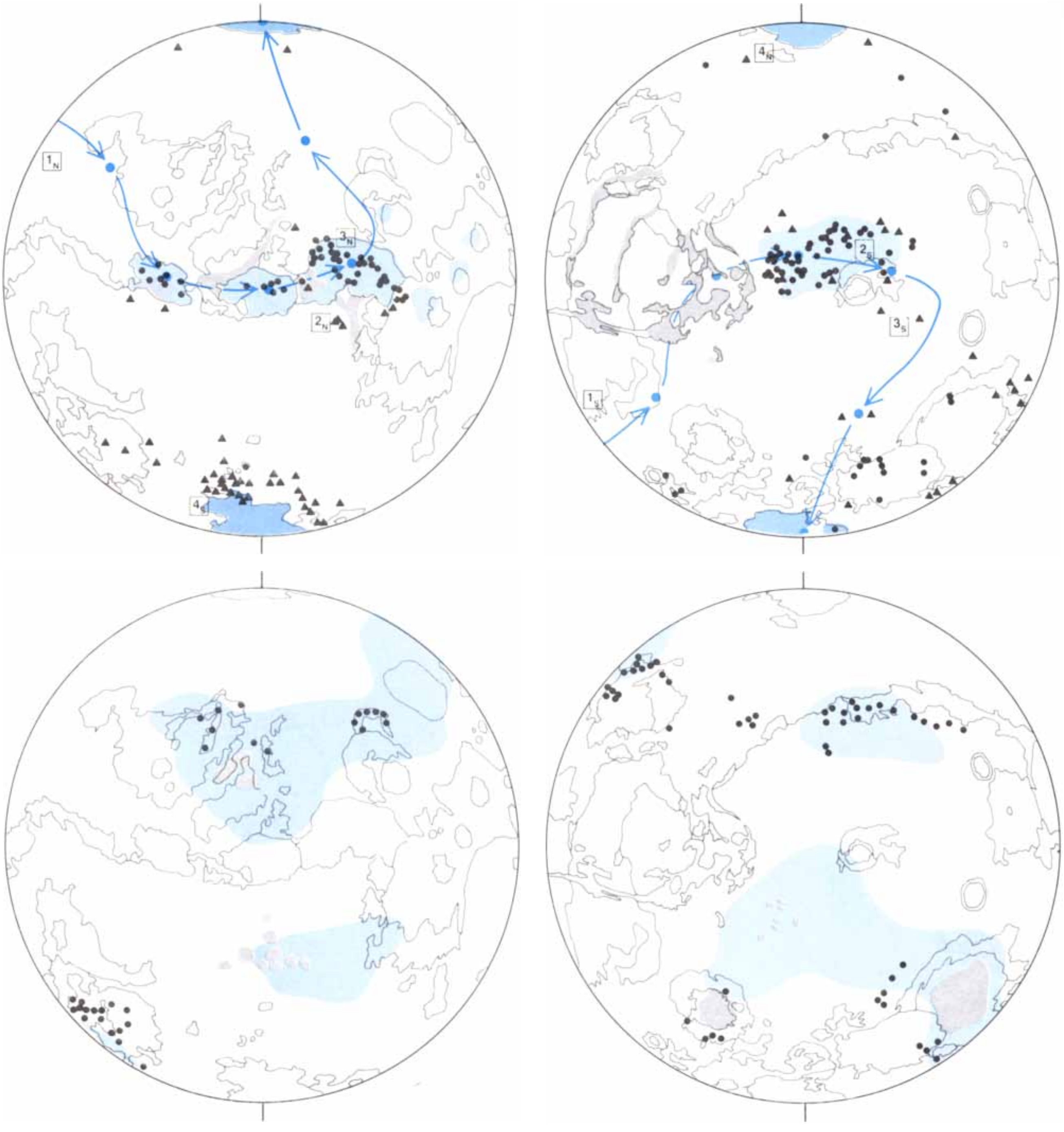
H. J. Melosh of the University of Arizona calculated the compressional and extensional stresses due to the reorientations of the poles. His calculations help to explain the existence, location and direction of large compressional ridges that extend for more than 4,000 kilometers across old cratered highlands. These ridges are much larger than the wrinklelike features that were produced in response to stresses caused when the lithosphere sagged under the weight of the Tharsis volcanoes. Similar correlations can be made for features that might have been caused by earlier polar shifts, but regional tectonism and volcanism, as well as the likelihood that Mars was expanding because of its internal heating, make such comparisons difficult.

The hypothesis that the Martian poles have wandered is supported by



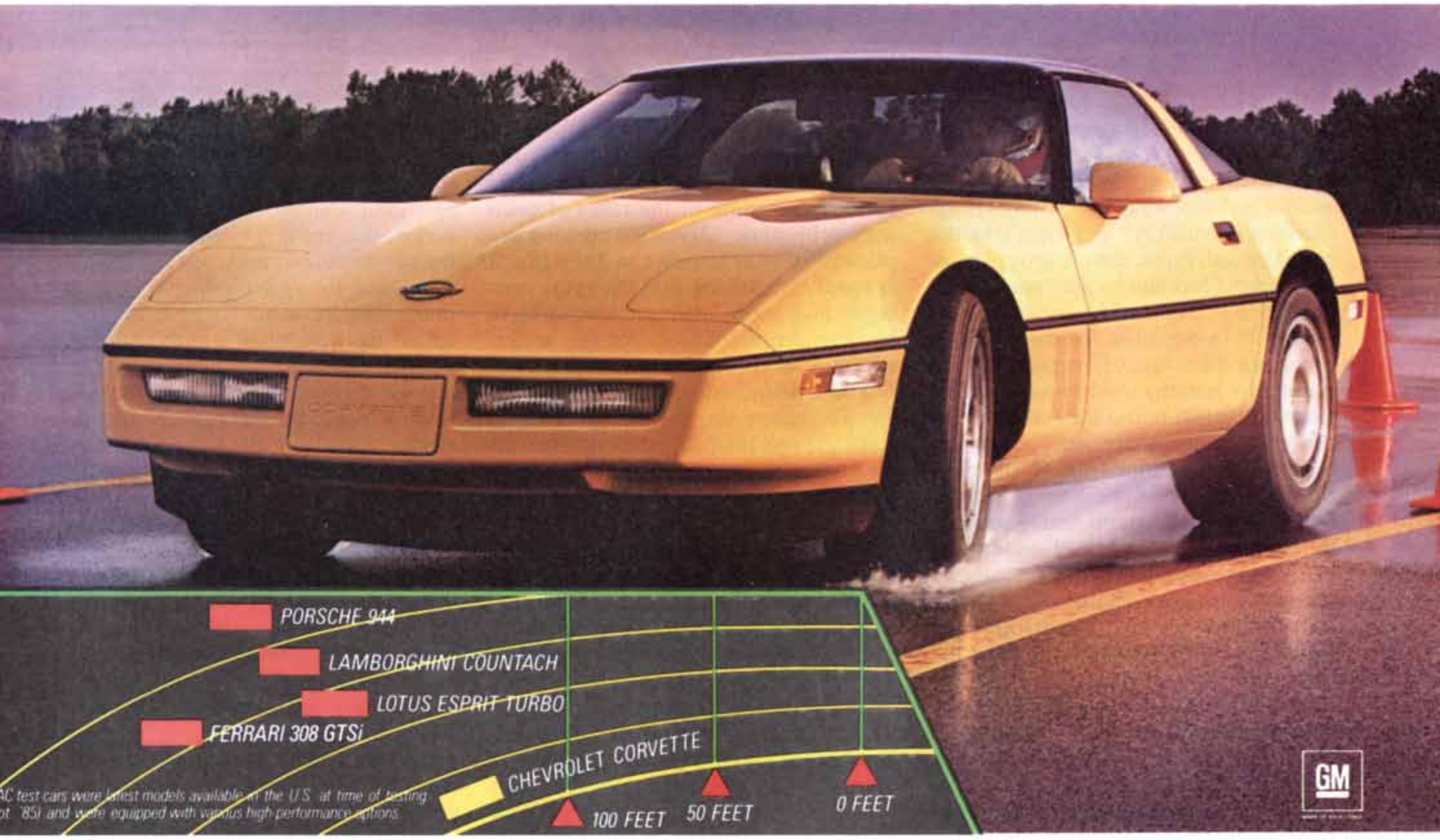
to escape through the surface, cutting deep outflow channels (*dark blue*). At about this time volcanic activity in the Tharsis region and in a region now known as the Elysium volcanic fields caused the crust to shift once again. By two billion years B.P. (*4*) the first of the major Tharsis volcanoes had begun to form; it is in the right hemisphere, just above the broken line marking the boundary of the northern periglacial region. Trapped deposits of ice and water continued to escape through the surface, cutting major systems of outflow channels in regions that had once been polar. About one billion years B.P. (*5*) the last stages of volcanic mountain building formed the three major Tharsis volcanoes (*near the northern periglacial region of the right hemisphere*) and Olympus Mons, Mars's highest mountain (*north of the Tharsis volcanoes*). Olympus Mons, which is now near the equator, was thus formed in a region where there were thick layers of ice, developing a foundation similar to that of volcanic mountains found on the earth in Iceland. During this period Olympus Mons and the Tharsis volcanoes accumulated large deposits of ice and frost on their northern slopes. Today (*6*) they have moved to the equator and changed orientation, and so the slopes that formerly faced north (which had accumulated extensive deposits) now face northwest. Vestiges of ancient polar deposits are visible near today's equator (*pale blue*). The erosion of these deposits yields the fine dust that is currently being gathered up by dust storms and deposited on the present polar regions (*medium blue*).

a good deal of evidence, and it offers new explanations for many of the planet's enigmatic features. For example, it explains how networks of channels could have formed in regions that are now polar and how large reservoirs of water and carbon dioxide may have been trapped in many regions of the planet. It also explains why ice-related features appear well outside the present polar zones and why today's polar deposits are so young. In order to prove or disprove the hypothesis, however, it will be necessary to have more data from planetary probes such as the planned Martian Observer. In particular it would be helpful to have samples returned from the Martian surface. Such samples would help to establish the conditions under which the equatorial deposits formed, and they would also provide a much more precise scale with which to calibrate the age of Martian features.



POLAR AND PERIGLACIAL FEATURES are found in high concentrations in areas that were within the Martian polar or periglacial zones for long periods. The top two images show the approximate path of the poles (*dark blue*). The path is marked by such features as pedestal craters (*dots*), thick surface deposits (*light blue patches*), outflow channels (*gray patches*) and thick deposits on crater floors (*triangles*). Also shown are the pole points that can be deduced from the location and orientation of grazing-impact craters (*numbered boxes*). The bottom images show periglacial features. Lobate flows, aprons of debris that have been interpreted as remnants of ice-rich surface deposits, are represented by dots. (The lobate flows were mapped by Steven W. Squyres of the National Aeronautics and Space Administration's Ames Research Center.) Thick wind-etched deposits that appear in isolated clusters and that may originally have been icy dust are represented by gray patches. Light blue patches represent areas that were within periglacial zones longest.

Anti-lock braking power you can grade on a curve.



You're driving 55 MPH on a rain-slick curve. Suddenly the unexpected: You stand on the brake pedal and steer to stay in your lane. You might expect Europe's most exotic cars to handle such a crisis effortlessly. Yet for all its awesome straight-line braking ability, Ferrari 308 GTSi failed to negotiate a 150-foot radius curve at maximum braking in USAC-certified testing. Lamborghini Countach failed. Lotus Esprit Turbo failed. Porsche 944 failed. Only the 1986 Corvette demonstrated the ability to steer and stop in these conditions at the same time. Only Corvette made the turn while coming to a controlled stop. When conditions turn foul, Corvette's new computerized Bosch ABS II anti-lock braking system is designed to help improve a driver's ability to simultaneously brake and steer out of trouble.

Why does the Corvette feature the world's most advanced braking technology? Because a world-class champion should give you the edge in an emergency.

Corvette. A world-class champion.

TODAY'S CHEVROLET Live it!

The Enormous Theorem

The classification of the finite, simple groups is unprecedented in the history of mathematics, for its proof is 15,000 pages long. The exotic solution has stimulated interest far beyond the field

by Daniel Gorenstein

How could a single mathematical theorem require 15,000 pages to prove? Who could read such a proof? Who could pass judgment on its validity? Yet there it is: the proof that all finite, simple groups have been found has run to between 10,000 and 15,000 pages. Of course, no one person is responsible for the achievement, nor is the size of the proof attributable to lengthy computer calculations (although computers are used at one place in the analysis). The work is instead the combined effort of more than 100 mathematicians, primarily from the U.S., England and Germany, but also from Australia, Canada and Japan. The complete proof is scattered across the pages of some 500 articles in technical journals, almost all published between the late 1940's and the early 1980's.

The idea of a mathematical group has been fundamental in mathematics since it was first introduced by the French mathematician Évariste Galois in the 1830's. Galois used properties of groups to settle in the negative a question that was then 200 years old: Can the solutions of polynomial equations of the fifth degree and higher be expressed by formulas similar to the familiar one for quadratic equations (and the less familiar ones for third-degree and fourth-degree equations)? Although the application Galois had in mind was limited to polynomial equations, the concept of a group has turned out to occur throughout both mathematics and nature.

The rotations of a sphere, the periodicities of a crystal and the symmetries of atoms are all examples of groups. Even the interactions of elementary particles and the "eightfold way," whereby certain particles are described as composite objects made up of quarks, need group-theoretical formulations in order to be understood.

The system of ordinary integers together with the operation of addition forms a group; indeed, the rules for combining the members of a group are simply borrowed, in more general form, from some of the rules of ordinary arithmetic. Given the striking applicability of arithmetic in daily life, it is hardly surprising that the same concepts in a more abstract setting have become powerful tools for the understanding of the universe.

The fundamental building blocks for all groups are the simple groups. Such groups bind together like the atoms in a molecule to generate ever more intricate patterns of groups. The term "simple" has no structural connotation: just as the atom has a highly complex internal structure, so does the simple group. A simple group is simple only in the sense that, like the atom, it cannot be decomposed into smaller entities of the same kind.

Many groups have an infinite number of elements, or members; again the number system is a good example, for there is no limit to the size of the numbers that can be added together. The rules for combining group elements are also satisfied, however, by many systems of finite size; such systems are called finite groups. For example, the hours registered by a clock can be added like numbers, but the result is expressed as an hour between one and 12: five hours later than 10 o'clock is three o'clock, not 15 o'clock. The 12 hours together with clock addition form the so-called clock group with 12 elements. Clock groups having "periods" other than 12 can also be readily constructed [*see illustration on page 106*].

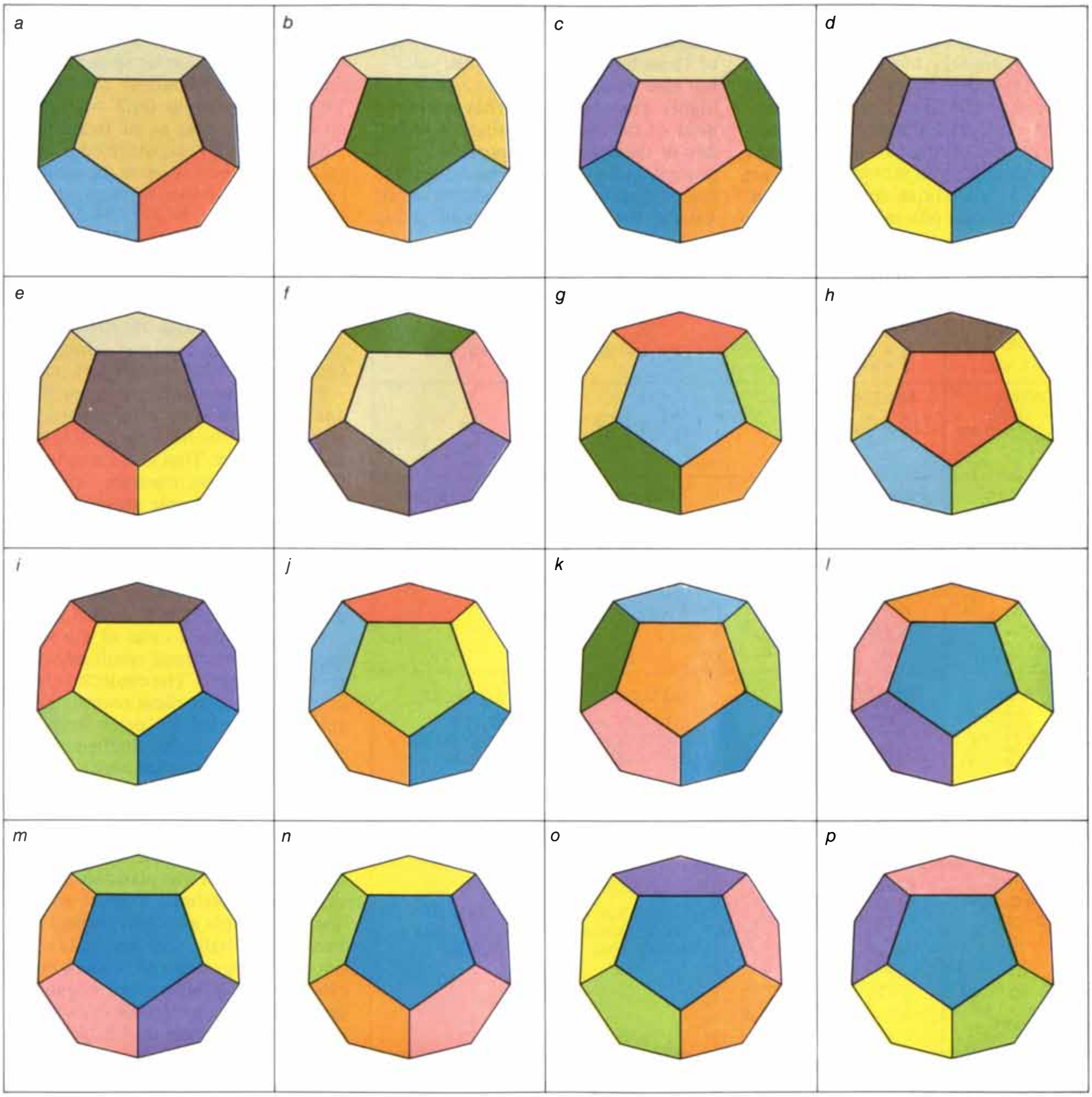
What then is a group? The formal definition is precise, but it is perhaps best understood through specific examples. A group is a set of elements together with an operation designated by an asterisk ($*$) for combining them. Thus for any two elements a and b in the set, the "product" $a * b$ must also be a member of the set. This property is called closure. Furthermore, to be a group the operation $*$ must satisfy three rules. First, the set must include a so-called identity element, designated e, such that for any element a in the set the products $a * e$ and $e * a$ are equal to a. Second, for each element a in the set there must be some element in the set called the inverse of a and designated a^{-1}, such that the products $a * a^{-1}$ and $a^{-1} * a$ are equal to e. Finally, the operation $*$ must be associative. In other words, for every three members a, b and c in the set the sequence in which the operation $*$ is carried out does not affect the result: $a * (b * c)$ is equal to $(a * b) * c$.

Note that the definition does not require the operation $*$ to be commutative: $a * b$ need not in general be equal to $b * a$. In fact, most of the interesting groups arising in mathematics as well as in nature are noncommutative [*see illustration on page 109*].

One can now appreciate how the rules for combining the elements in a group are the basic laws of arithmetic in more abstract form. Thus consider how the integers form a group when the combining operation $*$ is taken to be ordinary addition. The group product of two integers a and b is then their ordinary sum; for example, $5 * 7$ is $5 + 7$ or 12. Moreover, the identity element is the number 0, the inverse of the number a is the number $-a$ and for any numbers a, b and c the associative law holds; for example, $3 + (4 + 5)$ is equal to $(3 + 4) + 5$. Similarly, the nonzero rational numbers together with ordinary multiplication also form a group: the number 1 is the identity element and the number $1/a$ is the inverse of the number a. (The number 0 must be excluded because it has no inverse.) The 12 hours of a clock form a group in which zero o'clock, that is, 12 o'clock, is the identity element and the inverse of, say, five o'clock is seven o'clock. These examples of arithmetic groups are commutative because neither the addition nor the multiplication of numbers depends on the order in which the numbers are combined. On the other hand, groups whose elements are certain rotations or reflections of regular geometric figures such as spheres or equilateral triangles are almost always noncommutative.

There is an analogy between group theory and ordinary multiplication that can help in picturing how complex groups are built up from simple groups. Every whole number is either prime or composite. A prime number such as 2, 3, 5, 7 and 11 is evenly divisible only by itself and 1. If a whole number is not prime, it is composite, and the so-called fundamental theorem of arithmetic states that every composite number can be factored into a unique set of prime numbers. The composite number 12, for example, has the prime factors 2, 2 and 3.

In group theory there is a process



SYMMETRY OPERATIONS on the regular dodecahedron form a mathematical object called a group, which has been a major topic of interest to mathematicians for the past 150 years. The theory of groups has also found important applications outside mathematics, notably in crystallography and in the physics of elementary particles. All finite groups are built up from so-called simple groups, which play a role in group theory similar to the role of atoms in physics or prime numbers in arithmetic. The rotations of the dodecahedron that preserve its orientation in space form the smallest simple group whose elements do not commute with one another: in other words, the final position of the dodecahedron with respect to a fixed direction in space depends on the order in which the rotations are carried out. There are 60 elements in the group, which correspond to the five positions of each of the 12 faces of the dodecahedron; the illustration shows each face in the front position (*a–l*) and the five front positions of the dark blue face (*l–p*). The group has played a major role in the history of mathematics. Évariste Galois, the mathematician who invented group theory, showed that the simplicity of this group and the fact that it has a nonprime number of elements lead to the resolution of a classic mathematical problem: the general polynomial equation with rational coefficients whose highest-power term is x^5 cannot be solved with radicals.

called telescoping that splits an arbitrary finite group into a unique set of simple groups in much the same way as a composite number is decomposed into its prime factors. Moreover, the number of elements in each of these simple-group components is a factor of the number of elements in the parent group, and the product of all these numbers is equal to the number of elements in the parent group.

One must be careful, however, not to push the analogy too far. Indeed, the simple-group components of a parent group may contain a composite number of elements. Even more significant, whereas the product of all the prime numbers in a given set (such as the set 2, 2, 3, 5) is a unique composite number (in this case 60), in general a given set of simple groups can be combined to form many nonequivalent groups.

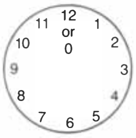
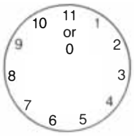
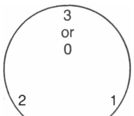
It is not only the extraordinary length of the proof that has made the classification of the finite, simple groups so unusual in the annals of mathematics but also the intriguing nature of the solution. As research on the problem progressed, group theorists not only were discovering infinite families of simple groups—and in the end a total of 18 such families were constructed—but also were discovering a number of highly irregular groups that fitted into none of the regular families. The first five of these puzzling sporadic simple groups, as they came to be called, had been found by Émile Mathieu in the 1860's; the smallest Mathieu group includes exactly $8 \times 9 \times 10 \times 11$, or 7,920, elements. The sixth sporadic group, which includes exactly 175,560 elements, was uncovered by Zvonimir Janko, then at Monash University in Australia, a full century after Mathieu's work.

Thereafter the strange new sporadic creatures began to pop up at the rate of about one per year, keeping pace with the intense theoretical developments of the 1960's and 1970's. The excitement of the discoveries spilled over to the larger mathematical community. The climax came in 1982, when Robert Griess, Jr., then at the Institute for Advanced Study, constructed a group that came to be known as the monster, because the number of elements in it is 808,017,424,794,512,875,886,459,904,961,710,757,005,754,368,000,000,000, or roughly 8×10^{53}. Ultimately 26 sporadic simple groups were discovered. The monster is the largest of them, but because its structure exhibits so many internal symmetries, Griess renamed it the friendly giant.

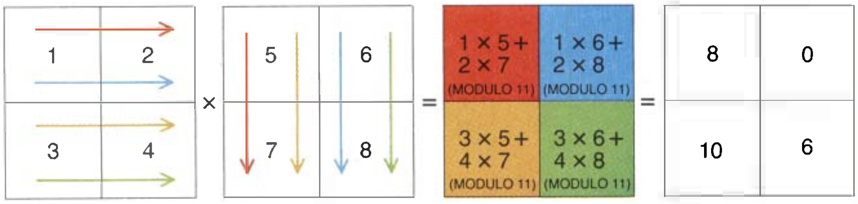
Finding a simple group is one thing, and discovery is its own reward, but proving that you have them all is quite another matter. That is the assertion of the classification theorem: The universe of finite, simple groups is made up of the 18 regular, infinite families of groups and the 26 sporadic groups, and no others! It is the proof of this statement that has taken the 500 articles and nearly 15,000 journal pages.

In the normal course of events an explicitly conjectured result may long precede its proof. The conjecture is not made without excellent reason: the result is suggested by prior investigations, or there exist mathematically nontrivial examples for which the conjecture is valid. Throughout most of the war game against the finite, simple groups, however, group theorists were unable even to estimate the size of the enemy troops. It was plausible to think that many—perhaps a great many—sporadic simple groups remained bunkered out of sight. Consequently it was not until many years after work on the classification problem had begun that group theorists were able even to formulate the theorem they hoped eventually to prove.

This reality forced a cautious strategy. Battles were fought over only limited types of simple groups. These hard-won victories yielded restricted classification theorems, which gradually began to root out the smaller of the remaining sporadic groups. The process is well illustrated by Janko's discovery of the sixth sporadic group, which is related to the 17th regular family of simple groups that had been

SET	GROUP OPERATION	IDENTITY ELEMENT	INVERSE ELEMENT
ORDINARY INTEGERS 0, ±1, ±2,...	ADDITION [2 + 3 = 5; 3 + (−7) = −4]	0 [3 + 0 = 0]	$-a$ [3 + (−3) = 0]
NONZERO RATIONAL NUMBERS p/q	MULTIPLICATION $\left[\frac{2}{3} \times \frac{4}{5} = \frac{8}{15}\right]$	1 $\left[\frac{2}{3} \times 1 = \frac{2}{3}\right]$	$\frac{1}{p/q}$ $\left[\frac{2}{3} \times \frac{1}{2/3} = 1\right]$
12 HOURS ON A CLOCK 	ADDITION MODULO 12 [3 + 4 = 7; 6 + 8 = 2]	0 O'CLOCK [3 O'CLOCK + 0 O'CLOCK = 3 O'CLOCK]	12 − a O'CLOCK [2 O'CLOCK + (12 − 2) O'CLOCK = 0 O'CLOCK]
11 HOURS ON A CLOCK 	ADDITION MODULO 11 [3 + 4 = 7; 6 + 8 = 3]	0 O'CLOCK [3 O'CLOCK + 0 O'CLOCK = 3 O'CLOCK]	11 − a O'CLOCK [2 O'CLOCK + (11 − 2) O'CLOCK = 0 O'CLOCK]
3 HOURS ON A CLOCK 	ADDITION MODULO 3 [2 + 1 = 0; 2 + 2 = 1]	0 O'CLOCK [2 O'CLOCK + 0 O'CLOCK = 2 O'CLOCK]	3 − a O'CLOCK [2 O'CLOCK + (3 − 2) O'CLOCK = 0 O'CLOCK]

EXAMPLES OF GROUPS are given in the table. Each group is made up of a set of elements and a group operation, customarily called group multiplication, whereby the elements of the set can be combined to yield another (usually a third) group element. The group operation, which can be designated by an asterisk (∗), acts much like ordinary multiplication or addition. It must obey the associative law: for any elements a, b and c in the group the triple group "product" $(a * b) * c$ must be equal to the triple product $a * (b * c)$. The set of elements must include an identity element, designated e, such that $g * e$ and $e * g$, the group products of any element g with the identity element, are equal to g. For each element g in the set there must also be an inverse element g^{-1} such that $g * g^{-1}$ and $g^{-1} * g$, the group products of g with its inverse, are equal to the identity element e. The set of elements in a group can be infinite, as it is in the first two examples, or finite, as it is in the last three. The fourth and fifth examples are the so-called clock groups of period 11 and period 3 respectively; their elements are added as if they were the hours on an 11-hour or a three-hour clock.



"MULTIPLICATION" of two two-by-two square matrices is carried out according to the procedure shown. Similar rules exist for multiplying two *n*-by-*n* square matrices for any number *n*. The matrix product is a square matrix that has the same number of entries as the two multiplying matrices. The entries of the matrices can be drawn from the elements of any given number system, which is a mathematical system whose elements can be added, subtracted, multiplied and divided in much the same way as ordinary numbers. The rational numbers and the complex numbers are familiar examples of infinite number systems, and the elements of clock groups of prime period generate examples of some of the possible finite number systems. In the matrices shown here the entries are drawn from the clock group of period 11. To calculate in this finite number system one must first compute the ordinary arithmetic result and then find the remainder after dividing that result by 11. For example, the element $1 \times 5 + 2 \times 7$ at the upper left in the product matrix (*third matrix from left*) is equal to 19, and its remainder after division by 11 is 8. Matrices are often considered as individual mathematical objects. All finite groups, and in particular all the finite, simple groups, can be represented as groups of matrices that are combined by matrix multiplication.

found by Rimhak Ree of the University of British Columbia in 1960.

Within every simple group there are certain smaller groups, called centralizers of involutions, that are important for understanding the structure of the parent group. For the Ree groups the centralizers of involutions can be represented by two-by-two square matrices, whose four entries are drawn from a finite number system of size equal to some odd-numbered power of 3. For example, if the odd power of 3 is 1, the finite number system is made up of the three elements in the clock group of period 3. To prove one of the early, restricted classification theorems it was necessary to show that the Ree groups are the only simple groups that have the following property: their centralizers of involutions can be represented by two-by-two matrices whose entries are drawn from a finite number system of a size equal to an odd power of some prime number p. It was a natural first step toward that goal to try to prove the following conjecture: If an arbitrary simple group has the above-mentioned property, the size of the finite number system from which the entries of the two-by-two matrices are drawn is an odd power of the prime number 3. In time this conjecture was verified, except in the single case for which p is equal to 5 and the odd power is 1.

Janko set out to study the exceptional case of the Ree group problem with every expectation that there were no simple groups of the given type having a number system of size 5^1, or 5. In spite of all his efforts, however, he was not able to eliminate this possibility and thus complete the proof of the conjecture. On the contrary, with considerable effort he managed to demonstrate that if such a simple group existed, it would have to contain exactly $2^3 \times 3 \times 5 \times 7 \times 11 \times 19$, or 175,560, elements.

It was too much for Janko to believe he could establish such a sharp conclusion unless an actual group were hovering in the background. He pushed on with heightened expectation, arguing next that if such a group existed, it would have to be generated by two seven-by-seven matrices whose row and column entries are drawn from the clock group of period 11 [*see bottom illustration at right*]. In other words, if the two generating matrices are called A and B, the group is made up of all possible matrix products of the two matrices, such as AA, BB, ABA, $BBAABABBB$ and so forth.

The only question remaining was whether the group of such matrices has exactly 175,560 elements; if it did not, Janko's analysis would have yielded the contradiction he had initially sought. At first glance it might seem remarkable that all possible matrix products of A's and B's would be equivalent to one of only 175,560 matrices, for that is precisely what it means for a group of the required kind to exist. For example, matrix products of more than a million of the A's and the B's must belong to the set. Indeed, there are roughly 11^{49}, or 10^{51}, seven-by-seven matrices with entries from the clock group of period 11, and so the products formed by the two generating matrices yield only a small fraction of the total number of possible matrices. The calculations were nonetheless carried out entirely by hand, and they verified the existence of the sixth sporadic group, which is now called J_1 in honor of Janko.

Janko's construction of J_1 exhibits a certain parallel with the physics of elementary particles. Theoretical analysis provided evidence for a new simple group, and then explicit calculations, analogous to experimental verification, proved that such a group really exists.

This two-stage process is even more dramatically illustrated by Janko's second success. If J_1 was discovered by "fudging" the properties of the centralizers of involutions in the Ree groups, why not try a similar tack with some other known simple groups? Janko hit the jackpot almost immediately; in fact, he made a double hit. From a single centralizer of an involution he found evidence for two possible new simple groups, one with 604,800 elements and the other with 50,232,960. This time, however, Janko was not able to provide the experimental verification. Using the information derived by Janko, Marshall Hall, Jr., and David Wales of the California Institute of Technology constructed the smaller group, J_2, and Graham Higman of the University of Oxford and John McKay of Concordia University in

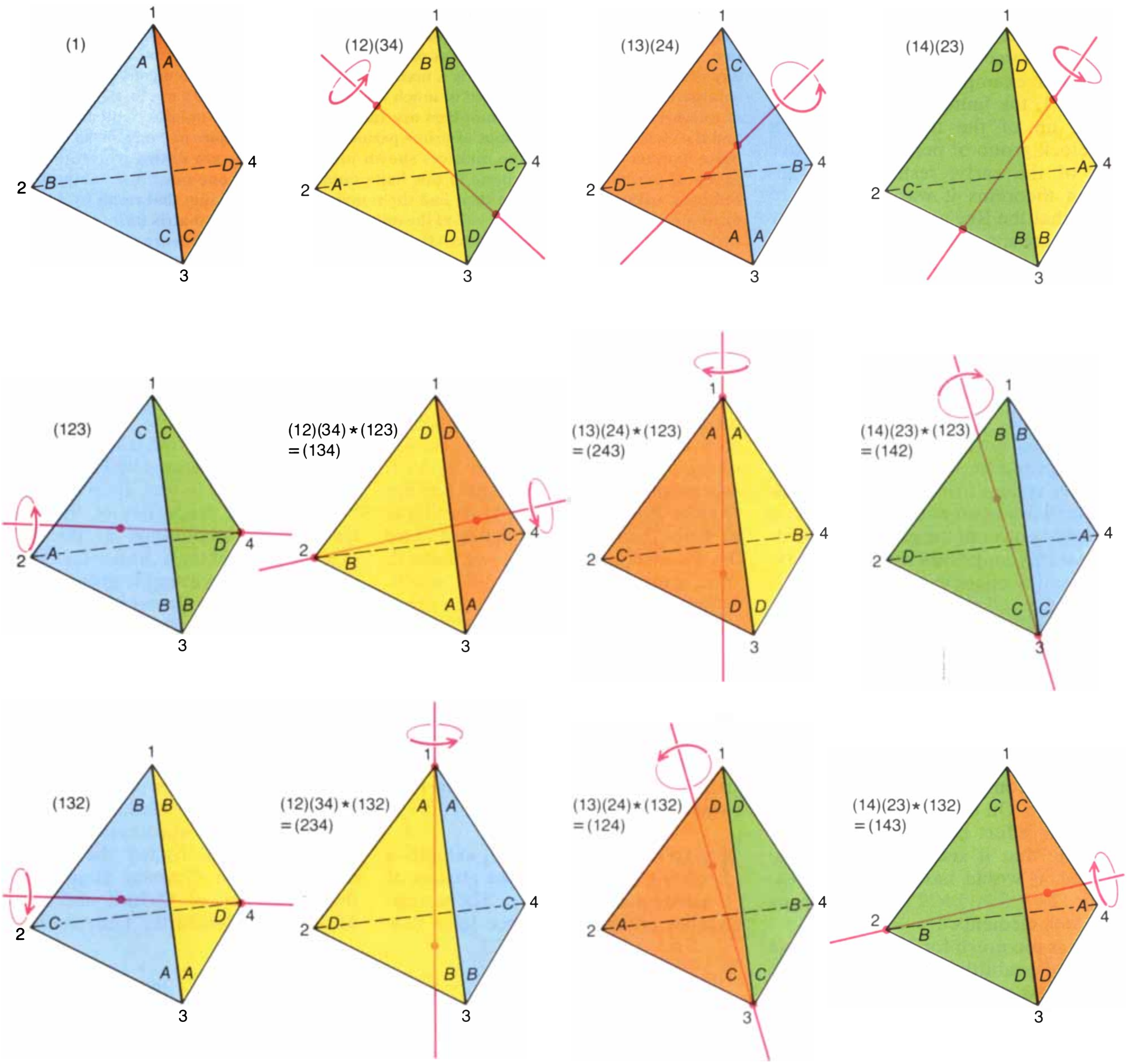
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 8 & 2 & 10 & 10 & 8 & 10 & 8 \\ 9 & 1 & 1 & 3 & 1 & 3 & 3 \\ 10 & 10 & 8 & 10 & 8 & 8 & 2 \\ 10 & 8 & 10 & 8 & 8 & 2 & 10 \\ 8 & 10 & 8 & 8 & 2 & 10 & 10 \\ 1 & 3 & 3 & 9 & 1 & 1 & 3 \\ 3 & 3 & 9 & 1 & 1 & 3 & 1 \end{pmatrix}$$

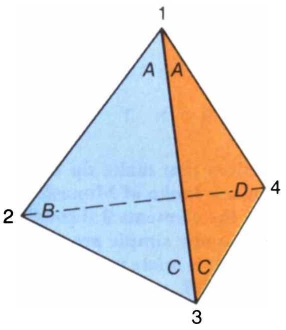
TWO SEVEN-BY-SEVEN MATRICES generate all the matrices that make up the elements of the sporadic simple group discovered in 1965 by Zvonimir Janko of Monash University in Australia. The entries of the matrices are drawn from the elements 0 through 10 in the clock group of period 11. Janko's group was the sixth sporadic simple group to be found, and it was the first to be discovered in more than 100 years. It is made up of all possible matrix products of the two generating matrices, such as *AA, BB, ABA, BBAABABBB* and so forth. There are exactly $2^3 \times 3 \times 5 \times 7 \times 11 \times 19$, or 175,560, matrices in the group.

Canada constructed the larger one, J_3. The calculations for J_3 were too complicated to be done by hand, and so they were carried out on a computer.

The scene now shifts, for Hall and Wales had been able to construct J_2 as a familiar kind of group of permutations. Group theorists flocked to the new site, and by the time the rush was over four more sporadic groups had been uncovered. The speed with which these early discoveries were made only reinforced the impression that the number of still undetected sporadic groups was large—possibly even infinite. With hindsight, however, mathematicians have come to realize that Janko's discovery of J_2 and J_3 was almost miraculous. Out of a myriad of potential variants from which to select, he had chosen one of the few possibilities capable of flowering.

I have stressed that all finite groups can be broken down into simple com-





ROTATIONS OF A REGULAR TETRAHEDRON that preserve its orientation in space form a group with 12 elements. The rotations are combined sequentially, that is, the group product of two rotations is the result of doing one rotation and then the second. The four rotations in the top row are the identity rotation, which does nothing, and the three "edge" rotations. Each of the edge rotations interchanges two pairs of vertexes of the tetrahedron; for example, the rotation (12)(34) interchanges the vertexes at positions 1 and 2 and the vertexes at positions 3 and 4. The middle row shows the group product of each edge rotation followed by the "face" rotation (123), which sends the vertex at position 1 to position 2, the vertex at position 2 to position 3 and the vertex at position 3 back to position 1. The bottom row shows the group product of each edge rotation followed by the face rotation (132). Each product is equivalent to a face rotation. The rotations are applied to the tetrahedron at left.

ponents. What I have not done so far, however, is to say exactly what a simple group is or how the decomposition is accomplished. Both points are related to the mathematical procedure called telescoping, which I mentioned briefly above. When a telescopic image of a group is made, multiplication in the parent group G is reflected in the image group G', although normally in a reduced form. The process can be compared to observing an object through the reverse end of a telescope: the general features of the object are preserved but its apparent size is diminished.

In the mathematical process of telescoping each element a of the group G must be associated with an element a' of the group G'; the element a' is called the image of a. Several elements of G may have the same image in G', and this accounts for the reduction in size. In addition every element of G' must be the image of at least one element of G. If such a way of associating the elements of a group G with the elements of a group G' is to be called telescoping, the group operations in G and in G' must be closely interrelated: if the elements a' and b' of the group G' are the images of the elements a and b of the group G, the product $a' * b'$ in G' must be the image of the product $a * b$ in G.

An arbitrary group G always has at least two telescopic images. One of them is its mirror image: G' is just the same group as G and each element in G is its own image. In effect, G is left untouched, and so the product condition for telescoping is satisfied. The other telescopic image of any group G is its point image: here G' is the group that has only one element, namely the identity e', and e' is the image of every element of G. Then no matter how the elements a and b are chosen in G, the three elements a, b and $a * b$ all have the same image e' in G'. Because $e' * e'$ is equal to e' in G', the condition for telescoping is again satisfied.

SECOND ELEMENT

FIRST ELEMENT

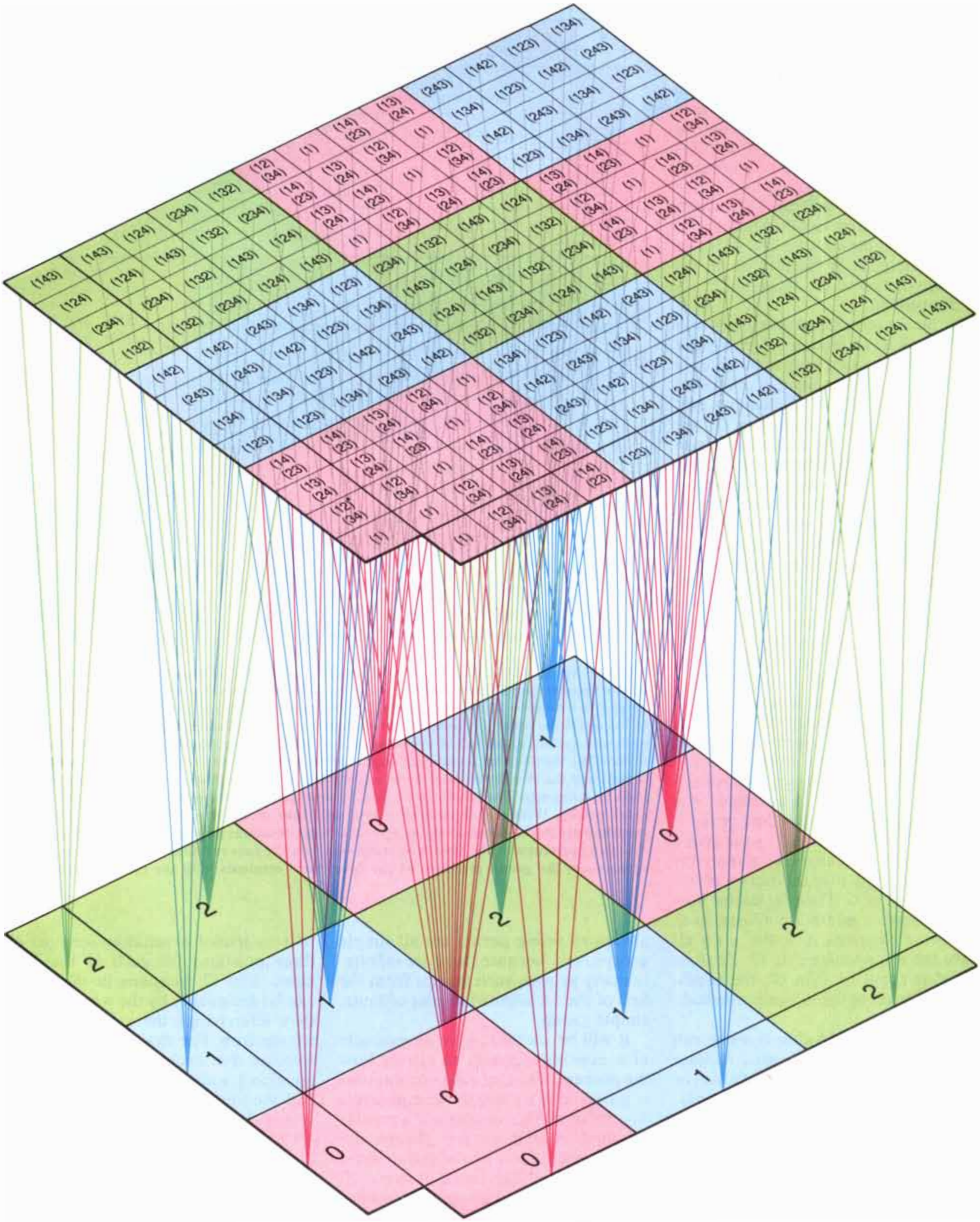
*	(1)	(12)(34)	(13)(24)	(14)(23)	(123)	(134)	(243)	(142)	(132)	(234)	(124)	(143)
(1)	(1)	(12)(34)	(13)(24)	(14)(24)	(123)	(134)	(243)	(142)	(132)	(234)	(124)	(143)
(12)(34)	(12)(34)	(1)	(14)(23)	(13)(24)	(134)	(123)	(142)	(243)	(234)	(132)	(143)	(124)
(13)(24)	(13)(24)	(14)(23)	(1)	(12)(34)	(243)	(142)	(123)	(134)	(124)	(143)	(132)	(234)
(14)(23)	(14)(23)	(13)(24)	(12)(34)	(1)	(142)	(243)	(134)	(123)	(143)	(124)	(234)	(132)
(123)	(123)	(243)	(142)	(134)	(132)	(124)	(143)	(234)	(1)	(13)(24)	(14)(23)	(12)(34)
(134)	(134)	(142)	(243)	(123)	(234)	(143)	(124)	(132)	(12)(34)	(14)(23)	(13)(24)	(1)
(243)	(243)	(123)	(134)	(142)	(124)	(132)	(234)	(143)	(13)(24)	(1)	(12)(34)	(14)(23)
(142)	(142)	(134)	(123)	(243)	(143)	(234)	(132)	(124)	(14)(23)	(12)(34)	(1)	(13)(24)
(132)	(132)	(143)	(234)	(124)	(1)	(14)(23)	(12)(34)	(13)(24)	(123)	(142)	(134)	(243)
(234)	(234)	(124)	(132)	(143)	(12)(34)	(13)(24)	(1)	(14)(23)	(134)	(243)	(123)	(142)
(124)	(124)	(234)	(143)	(132)	(13)(24)	(12)(34)	(14)(23)	(1)	(243)	(134)	(142)	(123)
(143)	(143)	(132)	(124)	(234)	(14)(23)	(1)	(13)(24)	(12)(34)	(142)	(123)	(243)	(134)

GROUP MULTIPLICATION TABLE is given for the tetrahedron group in the illustration on the opposite page. Note that the group contains elements that do not commute. For example, the rotation (12)(34) followed by the rotation (123) is equivalent to the rotation (134), but if the order of the first two rotations is reversed, their group product is (243). The red region at the upper left of the table shows that the group product of any two edge rotations of the tetrahedron is either an edge rotation or the identity rotation. The three edge rotations and the identity rotation form a group by themselves. Such a group within a group is called a subgroup. The table can be divided into nine square regions, as is indicated by the color coding. Within each colored region only four group elements are listed: in the red squares only the elements in the subgroup of edge rotations appear; in the blue squares are the group products of the four "red" elements with the face rotation (123), and in the green squares are the group products of the four "red" elements with the face rotation (132).

The definition of a simple group can now be given: it is a group whose only telescopic images are its mirror and point images. The most elementary examples of simple groups are the clock groups whose period is a prime number p. Indeed, an elementary theorem in group theory states that the number of elements in a telescopic image must evenly divide the number of elements in the parent group. Since a prime number has no divisors other than 1 and itself, any telescopic image of a clock group of prime period p can contain only one or p elements. This means that the only telescopic images of such clock groups are their mirror and point images. Hence the clock groups of prime period are all simple groups, and, because there are infinitely many primes, such groups form the first of the 18 infinite families of finite, simple groups.

It will be useful to give an example of a nonsimple group to clarify how the process of telescoping decomposes a group into its simple components. The set of all the rotations of a regular tetrahedron that do not change the spatial orientation of the tetrahedron forms a group. The result of two such rotations can be expressed as another (possibly a third) kind of rotation, and for every possible rotation there is an inverse rotation that brings the tetrahedron back to its starting position. The "rotation" that does nothing to the tetrahedron serves as the identity element. There are 12 elements in this group, which correspond to the 12 positions of the tetrahedron that maintain its spatial orientation: one set of three positions for each of the four faces. The 12 rotations in the group can be designated by the way in which they interchange the vertexes of the tetrahedron. For example, the "edge" rotation that exchanges the vertex at position 1 with the vertex at position 2 and the vertex at position 3 with the vertex at position 4 can be written (12)(34); the "face" rotation that sends the vertex at position 1 to position 2, the vertex at position 2 to position 3 and the vertex at position 3 back to position 1 can be written (123) [*see illustration on opposite page*].

One can now show that this group G of rotations of the regular tetrahedron has a telescopic image that is neither its mirror nor its point image. The existence of such a telescopic image will imply that the tetrahedron group is not simple. One takes the clock group of



SIMPLE-GROUP COMPONENT of the tetrahedron group is obtained by finding a so-called telescopic image of the tetrahedron group. Each element colored red in the multiplication table of the tetrahedron group is associated with the element 0 in the clock group of period 3. Each element colored blue in the table is associated with the element 1 in the clock group, and each element colored green in the table is associated with the element 2 in the clock group. The illustration shows how the table for addition modulo 3 in the clock group reflects the multiplication table for the tetrahedron group. For example, the group product of any two "blue" rotations is equivalent to a "green" rotation, just as the sum of 1 and 1, the two "blue" elements in the clock group, is 2, the "green" element in the clock group. Note that the telescopic image of every element in the subgroup of edge rotations is 0, the identity in the clock group.

period 3 as the image group G'. The tetrahedron group G can then be mapped onto G' as follows: the identity element 0 of G' is to be the image of the four elements (1), (12)(34), (13)(24) and (14)(23) of G; the element 1 of G' is to be the image of the four elements (123), (134), (243) and (142) of G, and the element 2 of G' is to be the image of the four elements (132), (234), (124) and (143) of G [*see illustration on opposite page*].

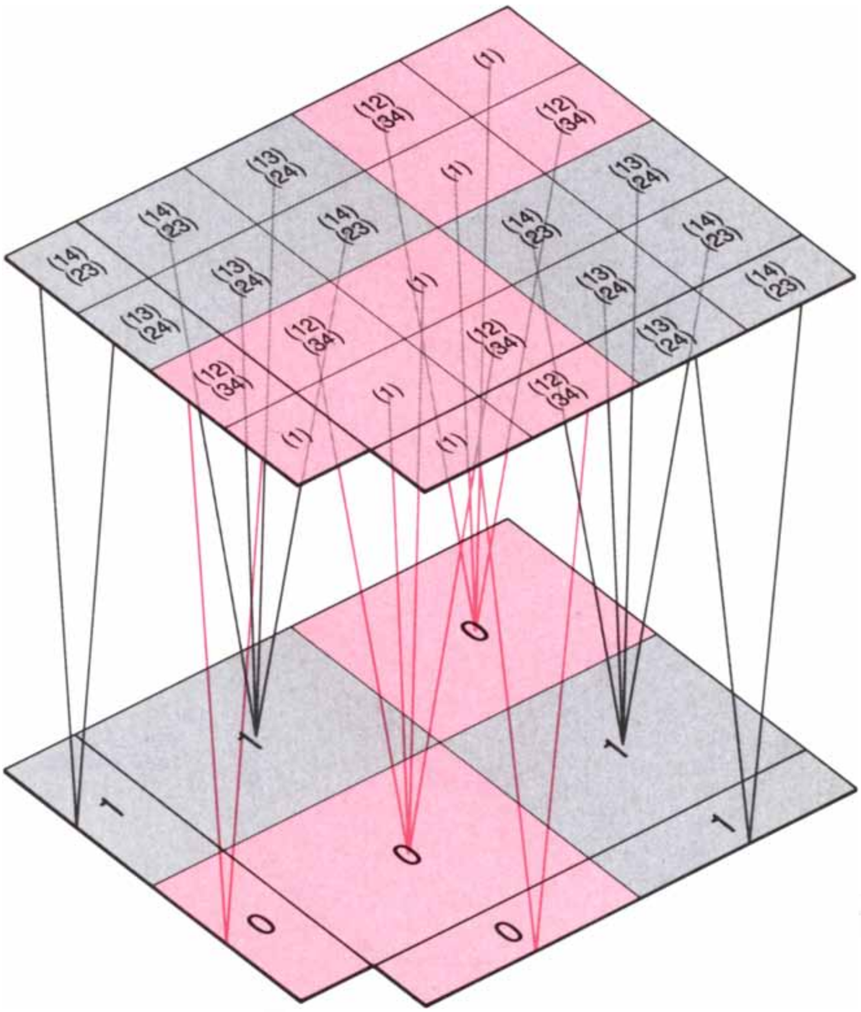
Considerable checking is needed to verify that this way of associating elements has the telescopic property. For example, consider the elements (123) and (134). Their group product $(123) * (134)$ in G, which is the result of applying first the rotation (123) and then the rotation (134) to the tetrahedron, is equal to the rotation (124) [*see illustration on page 109*]. These three elements have the images 1, 1 and 2 respectively in G'. Because the group product $1 * 1$ in G' is $1 + 1$ (modulo 3), which is equal to the element 2 in G', the telescopic condition holds for the pair of elements (123) and (134). Since G contains 12 elements whereas G' has only three, G' is indeed neither the mirror image nor the point image of G. Thus G is not simple.

How then does one find the simple components of the rotation group G? One of these components is the clock group G' of period 3. Indeed, G' is a telescopic image of G, and G' is simple since 3 is a prime number. The next step is to focus on all the elements of G whose image is the identity element e' of G': the four rotations (1), (12)(34), (13)(24) and (14)(23). These four elements form a group by themselves; such a group within a group is called a subgroup. The group consists of the four edge rotations that bring each of the four faces of the tetrahedron to its front but do not rotate a face once it has reached the front position [*see illustration on page 108*].

The subgroup of edge rotations is not simple either: it too has a telescopic image in addition to its mirror and point images, namely the clock group of period 2. Here the identity element of the clock group is the image of the edge rotation (12)(34) and the identity element (1), and the second element of the clock group is the image of the edge rotations (13)(24) and (14)(23). Since the clock group of period 2 is also simple, that group is a second simple component in the decomposition of the parent tetrahedron group G [*see illustration at right*].

The final step in the decomposition of G is to focus on the two elements (12)(34) and (1) in the group of edge rotations whose image is the identity element of the clock group of period 2. These two elements also form a group by themselves, and the only telescopic images of a group with two elements are its point image and its mirror image. Hence the subgroup made up of the edge rotation (12)(34) and the identity is itself a simple group, equivalent to the clock group of period 2. Since no further telescoping is possible, the tetrahedron group G must have three simple components, namely the clock groups of periods 3, 2 and 2. Since G contains 12 elements and 3 times 2 times 2 is 12, this result checks with my remark above that the product of the number of elements in the simple components must be equal to the number of elements in the parent group.

It is easy to show that the clock groups of prime period are the only commutative simple groups. In fact, if a commutative group contains a composite number n of elements, then for every number d that divides n evenly, the group has a telescopic image containing exactly d elements. Since the number d need not be equal to 1 or to n, such a commutative group is not simple. For example, the group of the four edge rotations of the tetrahedron is a commutative group for which n is equal to 4. Since 2 divides 4, the number d can be 2, and so the group has a telescopic image with two elements, as discussed above. Every simple group except the clock groups therefore contains at least one pair of elements a and



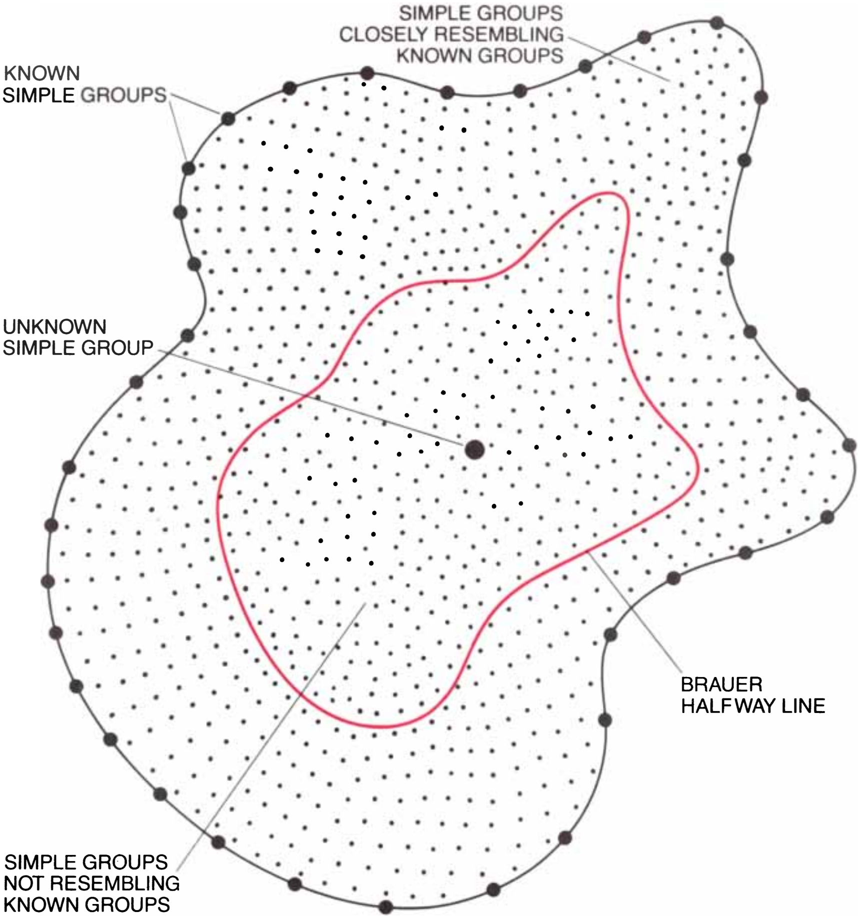
TELESCOPIC IMAGE of the subgroup of edge rotations gives the second simple-group component of the tetrahedron group. The edge rotation (12)(34) and the identity rotation (1) form a second, nested subgroup (*color*). They are associated with the identity element 0 in the clock group of period 2. The other two elements in the subgroup of edge rotations, namely the rotations (13)(24) and (14)(23), are associated with the element 1 in the clock group (*gray*). Addition modulo 2 in the clock group reflects group multiplication in the subgroup of edge rotations. For example, the group product of a "colored" rotation and a "gray" one is equivalent to a "gray" rotation, just as the sum modulo 2 of 0, the "colored" element in the clock group, and 1, the "gray" element in the clock group, is equal to 1, the "gray" element in the clock group. Thus the clock group of period 2 is also a simple-group component. A second clock group of period 2 is the third such component (not shown).

b for which $a * b$ is not equal to $b * a$.

The smallest example of a noncommutative simple group contains 60 elements. It can be described as the group of rotations of a regular dodecahedron that do not change the spatial orientation of the dodecahedron The 60 elements in the group correspond to the five possible positions for each of the 12 faces of the dodecahedron [*see illustration on page 105*].

The dodecahedron group is closely related to the rotation group of the tetrahedron. The tetrahedron group is structurally identical with the group of so-called even permutations of four letters. For example, if the letters *ABCD* are permuted to *CBAD*, the permutation is an odd one because it is out of order in three ways: *C* precedes *B*, *C* precedes *A* and *B* precedes *A*. On the other hand, *DCBA* is out of order in six ways, and so it is an even permutation of the letters. The dodecahedron group is structurally identical with the group of even permutations of five letters.

Although the tetrahedron group is not a simple group, all even permutations of five or more letters form simple, noncommutative groups. Such groups, which are called the alternating groups on *n* letters, make up the second infinite family of simple groups. Incidentally, it is this distinction between the alternating groups of degree less than or equal to 4 and the alternating groups of degree greater than or equal to 5 that underlies the work of Galois on the theory of equations. The distinction explains the sharp differences between the nature of the solutions of polynomial equations of degree less than or equal to 4 and the nature of the solutions of such equations of degree greater than or equal to 5.



PROOF OF CLASSIFICATION THEOREM for all finite, simple groups required showing that any simple group, with an unknown but perhaps arbitrarily complicated subgroup structure, has a subgroup structure equivalent to that of one of the known simple groups. The strategy of the proof is represented by the illustration, where the known simple groups are the dots at the boundary of the enclosed region. Toward the inside of the region the dots represent groups whose subgroup structures are increasingly far removed from the subgroup structures of the known simple groups. At the center of the region is the arbitrary simple group of unknown subgroup structure. Each stage in the proof of the theorem is represented by the movement of the unknown simple group toward the boundary, that is, at each stage the subgroup structure of the arbitrary group is shown to bear a closer resemblance to one of the simple groups on the boundary. The work of Richard Brauer led to a criterion that describes the stage about halfway to the boundary. The extreme length of the proof can be explained by the fact that some 100 paths to the boundary had to be explored.

The other families of simple groups are harder to describe. They all have representations as groups of square matrices of suitable size. In some cases the groups are initially defined by means of such matrices, whereas in others considerable work is required to obtain their matrix representations. The same statement holds for the sporadic simple groups, but there the situation is even more complicated: in some cases the matrix descriptions are based on computer calculations. Remarkably, even though the monster is the largest of the sporadic groups, Griess was able to find a representation for it entirely by hand. His construction is given in terms of square matrices of size 196,883 by 196,883 whose entries are taken from the complex numbers. The names and sizes of all the families of simple groups as well as the sporadic groups are listed in a table [*see illustration on pages 114 and 115*].

Many families of noncommutative simple groups had been discovered by the turn of the century, and each such group as well as Mathieu's five sporadic groups contained an even number of elements. This fact soon led to a natural conjecture: every noncommutative finite, simple group, known or unknown, must contain an even number of elements. It was not until 1962, however, that the by then celebrated conjecture was verified by Walter Feit and John Thompson, both then at the University of Chicago. The complexity of the proof of this easily understood statement foreshadowed the extreme length of the complete classification of the simple groups. The proof of the Feit-Thompson theorem filled one entire issue of the *Pacific Journal of Mathematics*, 255 pages in all. In 1965 Feit and Thompson were awarded the Cole Prize in Algebra for their work.

Their theorem can also be viewed as an example of a restricted classification theorem. In this form it states that the only simple groups containing an odd number of elements are the clock groups whose period is an odd prime number. Analogously the aim of the full classification theorem is to obtain a list of all the finite, simple groups without any restriction on the number of elements they may contain. As I noted above, the final theorem states that the complete list is made up of the 18 regular, infinite families of groups and the 26 sporadic groups.

Since more than 250 pages were required to prove the Feit-Thompson theorem, it is not surprising that a proof of the much more complicated conclusion of the full classification theorem has taken many thousands of pages. The length of the proof cannot be attributed to the difficulty of describing the known simple groups. The clock groups can be described in a few lines, yet it took Feit and Thompson several hundred pages to show that there were no other simple groups containing an odd number of elements. The length of the classification theorem must be traced to another source.

That source lies in the complexity of the subgroup structure of an arbitrary finite group compared with that of a simple group—just as a complex molecule has a far more tangled internal structure than a single atom does. To appreciate the nature of the classification problem, imagine that the finite, simple groups are spread throughout a region of the plane [*see illustration on opposite page*].

Group-theorists could not begin their proof with the assumption that any simple group has a subgroup structure approximately the same as that of a known simple group, for in essence that is what the theorem sets out to prove. Hence one obtains a more accurate picture of what had to be proved if one allows the subgroup structure of the simple groups in the planar region to vary over the same wide range of complexity as the structure of all finite groups. The dots on the boundary of the region represent the known simple groups, the dots near the boundary represent groups whose subgroup structures closely resemble the structures of the groups corresponding to the nearest dots on the boundary, and the dots deeper in the interior of the region represent groups whose structures bear progressively less resemblance to the structures of the known simple groups.

At the outset of the proof, one is given an entirely unknown simple group whose internal structure, as far as anyone can tell, is as intricate as that of any nonsimple group. Represent the given group with a dot in the middle of the planar region. To prove the theorem the given group must be forced, through a series of mathematical arguments, to coincide with one of the points on the boundary. Each little move of the dot toward the boundary is to be understood as the result of a refinement in the initially unknown subgroup structure of the given group. As the dot moves closer to the boundary, the internal structure of the given group begins to take on the shape of one of the known simple groups.

Since every boundary point is a possible destination, the analysis inevitably requires a large number of paths to be explored on the journey from the central point to the boundary. Distinct paths arise at forks in the journey; the forks correspond to stages at which the internal structure of the given group can have several alternative shapes. In the end roughly 100 different paths had to be followed before the proof was completed. Thus the classification theorem was actually the sum of about 100 separate theorems, which taken together yield the desired conclusion.

In the 1940's and 1950's the pioneering work of Richard Brauer on centralizers of involutions established one of the most important milestones along the journey to the boundary. Even though not all elements of a group necessarily commute with one another, it is useful to consider the set of all the elements in a group that commute with a given element *a*, that is, the set of all elements *g* in the group for which $g * a$ is equal to $a * g$. This set is called the centralizer of the element *a*. It is easy to check that such centralizers are themselves groups when their elements are combined according to the operation in the parent group, and so they are subgroups. A trivial example is the centralizer of the identity element *e*, which is always the entire parent group since one of the defining conditions of a group is that every element in the group must commute with the identity element. A more substantive example is the centralizer of the element (123) in the rotation group of the tetrahedron: the centralizer of (123) consists of the identity, the element (123) itself and the element (132).

Brauer's idea was to focus on those group elements *a* other than the identity element *e* for which the group product $a * a$ is equal to *e*. Group elements having this property are called involutions, and it is easy to show that any group with an even number of elements contains involutions. In view of the Feit-Thompson theorem, every noncommutative simple group therefore contains involutions. Brauer began by calculating centralizers of involutions in several of the 18 regular families and observed that they had the same general structure as the parent group, but in an embryonic form. He then wondered whether it might be possible to reconstruct the entire parent group solely from a knowledge of these centralizers, and in time he went on to resolve the question affirmatively in some important special cases.

Not only does the work of Brauer underlie the discovery of many of the sporadic groups but also it provided an operational procedure for subdividing the journey to the boundary of the planar region of simple groups into two parts. First, prove that the centralizer of an involution in an unknown simple group closely resembles the centralizer of an involution in one of the known simple groups. The analyses forming the basis for paths near the central dot are related to this aspect of the classification problem.

Knowing that the centralizer of an involution in the unknown group is the same as the centralizer of an involution in a known group is not at all equivalent to knowing that the unknown group is the same as the known group: the centralizer of an involution is just a small portion of the entire group. The aim of the second stage of the proof is to extend this local information about centralizers of involutions to a global equivalence, thereby showing that the unknown simple group is the same as one of the known simple groups. Paths near the boundary points are concerned with showing how the known simple groups are determined from information about the centralizers of their involutions. The separation of the proof of the classification theorem into these two phases can be pictured as a closed curve inside the planar region, about halfway from the central dot to the boundary.

Among the many mathematicians who took part in the final decade of work on the classification proof, I should mention the name of Michael Aschbacher of Caltech. In 1972, in a series of lectures at the University of Chicago, I presented a 16-step program for determining all the finite, simple groups. I suggested, in effect, the paths that would have to be followed on the journey to the boundary. I predicted a successful outcome by the end of the century, but my grand design and its accompanying optimism were met with considerable skepticism.

Neither the audience nor I had reckoned with Aschbacher. Fresh out of graduate school, he had just entered the field, and from that moment he be-

INFINITE FAMILIES OF FINITE, SIMPLE GROUPS

NAME OF FAMILY	NUMBER OF ELEMENTS
CLOCK GROUPS Z_p OF PRIME PERIOD p	p, WHERE p IS ANY PRIME NUMBER EXAMPLES: $\|Z_2\| = 2$; $\|Z_3\| = 3$; $\|Z_5\| = 5$; . . .
ALTERNATING GROUPS Alt_n, THE EVEN PERMUTATIONS OF n LETTERS, $n > 4$	$1/2 \times 1 \times 2 \times \ldots \times n$ EXAMPLES: $\|Alt_5\| = 1/2 \times 1 \times 2 \times 3 \times 4 \times 5 = 60$; $\|Alt_6\| = 360$; . . .
CHEVALLEY GROUPS LINEAR CHEVALLEY GROUPS $A_n(q)$, $n > 1$ OR $q > 3$	$q^{n(n+1)/2} \times (q^2-1) \times (q^3-1) \times \ldots \times (q^{n+1}-1)/d$, d = G.C.D.$(n+1,q-1)$ EXAMPLES: $\|A_1(4)\| = 4^{1\times(1+1)/2} \times (4^{1+1}-1)/[\text{G.C.D.}(1+1,4-1)] = 4 \times 15/1 = 60$; $\|A_1(5)\| = 5 \times (5^2-1)/[\text{G.C.D.}(1+1,5-1)] = 120/2 = 60$; $\|A_2(2)\| = 168$; . . .
SYMPLECTIC CHEVALLEY GROUPS $C_n(q)$, $n > 2$	$q^{n^2} \times (q^2-1) \times (q^4-1) \times \ldots \times (q^{2n}-1)/d$, d = G.C.D.$(2,q-1)$ EXAMPLE: $\|C_3(2)\| = 2^{3^2} \times (2^2-1) \times (2^4-1) \times (2^{2\times3}-1)/[\text{G.C.D.}(2,2-1)]$ $= 512 \times 3 \times 15 \times 63/1 = 1{,}451{,}520$
ORTHOGONAL CHEVALLEY GROUPS $B_n(q)$, $n > 2$ OR $q > 2$	$q^{n^2} \times (q^2-1) \times (q^4-1) \times \ldots \times (q^{2n}-1)/d$, d = G.C.D.$(2,q-1)$
$D_n(q)$, $n > 3$	$q^{n(n-1)} \times (q^n-1) \times (q^2-1) \times (q^4-1) \times \ldots \times (q^{2(n-1)}-1)/d$, d = G.C.D.$(4,q^n-1)$
EXCEPTIONAL CHEVALLEY GROUPS $G_2(q)$, $q > 1$	$q^6 \times (q^6-1) \times (q^2-1)$
$F_4(q)$	$q^{24} \times (q^{12}-1) \times (q^8-1) \times (q^6-1) \times (q^2-1)$
$E_6(q)$	$q^{36} \times (q^{12}-1) \times (q^9-1) \times (q^8-1) \times (q^6-1) \times (q^5-1) \times (q^2-1)/d$, d = G.C.D.$(3,q-1)$
$E_7(q)$	$q^{63} \times (q^{18}-1) \times (q^{14}-1) \times (q^{12}-1) \times (q^{10}-1) \times (q^8-1) \times (q^6-1) \times (q^2-1)/d$, d = G.C.D.$(2,q-1)$
$E_8(q)$	$q^{120} \times (q^{30}-1) \times (q^{24}-1) \times (q^{20}-1) \times (q^{18}-1) \times (q^{14}-1) \times (q^{12}-1) \times (q^8-1) \times (q^2-1)$
STEINBERG GROUPS $^2A_n(q)$, $n > 2$ OR $q > 2$	$q^{n(n+1)/2} \times (q^2-1) \times (q^3+1) \times (q^4-1) \times \ldots \times (q^{n+1}-(-1)^{n+1})/d$, d = G.C.D.$(n+1,q+1)$
$^2D_n(q)$, $n > 3$	$q^{n(n-1)} \times (q^n+1) \times (q^2-1) \times (q^4-1) \times (q^6-1) \times \ldots \times (q^{2(n-1)}-1)/d$, d = G.C.D.$(4,q^n+1)$
$^3D_4(q)$	$q^{12} \times (q^8+q^4+1) \times (q^6-1) \times (q^2-1)$
$^2E_6(q)$	$q^{36} \times (q^{12}-1) \times (q^9+1) \times (q^8-1) \times (q^6-1) \times (q^5+1) \times (q^2-1)/d$, d = G.C.D.$(3,q+1)$
SUZUKI GROUPS $^2B_2(q)$, $q = 2^m$, m ODD, $m > 1$	$q^2 \times (q^2+1) \times (q-1)$
REE GROUPS $^2G_2(q)$, $q = 3^m$, m ODD, $m > 1$	$q^3 \times (q^3+1) \times (q-1)$ EXAMPLE: $\|^2G_2(3^3)\| = 27^3 \times (27^3+1) \times (27-1) = 10{,}073{,}444{,}472$; . . .
$^2F_4(q)$, $q = 2^m$, m ODD [FOR $q = 2$, $^2F_4(2)'$]	$q^{12} \times (q^6+1) \times (q^4-1) \times (q^3+1) \times (q-1)$ $[\|^2F_4(2)'\| = 1/2\|^2F_4(2)\| = 1/2 \times 2^{12} \times (2^6+1) \times (2^4-1) \times (2^3+1) \times (2-1) = 17{,}971{,}200]$

COMPLETE LIST of finite, simple groups, together with the number of elements in each group, is given here. The table on this page lists the 18 infinite families of simple groups, and the table on the opposite page lists the 26 sporadic simple groups. The letters *A* through *G* designating the infinite families are derived from Lie theory, which is named after Sophus Lie. The families are named for Claude Chevalley, Robert Steinberg of the University of California at Los Angeles, Michio Suzuki of the University of Illinois and Rimhak Ree of the University of British Columbia. The sporadic simple groups are named for Émile Mathieu, Zvonimir Janko, now at the University of Heidelberg, Donald Higman of the University of Michigan, Charles Sims of Rutgers University, Jack McLaughlin of Michigan, Suzuki, Arunas Rudvalis of the University of Massachusetts at Amherst, Dieter Held of the University of Mainz, Richard Lyons and Michael O'Nan of Rutgers, John Conway of the University of Cambridge, Bernd Fischer of the University of Bielefeld, Koichiro Harada of Ohio State University, John Thompson, now at Cambridge, and Robert Griess, Jr., of Michigan. Individual simple groups that are members of a given infinite family are designated by numerical values of the variables *n*, *p* and *q*, where *n* can be any

came the driving force behind my program. In rapid succession he proved one astonishing theorem after another. Although there were many other major contributors to this final assault, Aschbacher alone was responsible for shrinking my projected 30-year timetable to a mere 10 years.

What does the future hold? My current work, along with that of several co-workers, is devoted to constructing a second-generation proof of the classification theorem. The first proof encompasses many early papers written long before there was any overall strategy for finishing the proof. Inevitably, therefore, the development of the first proof includes some false starts, inefficiencies and duplication. If we are successful in our current work, the second-generation proof will be only one-fifth as long as the first, with a commensurate improvement in conceptual clarity. By mathematical standards a 3,000-page proof will still be a proof of enormous length. Given the complexity of the problem, however, a substantially shorter proof will have to await the introduction of totally new methods.

The classification of the finite, simple groups is likely to have broad mathematical ramifications. Already the result has been applied to such diverse areas of mathematics as the theory of algorithms, mathematical logic, geometry and number theory. The monster group is now known to have deep connections, not yet fully unraveled, with the theory of elliptic functions.

The impact of the classification outside mathematics is less clear. There has, however, been some speculation that Griess's friendly giant, the monster group, may enter into the formulation of a possible unified field theory of elementary particles. But whatever the ultimate applications, finite-group theorists have succeeded in solving the most central problem of their subject, which existed implicitly from the moment Galois first added the concept of a group to mathematics.

SPORADIC SIMPLE GROUPS

NAME OF GROUP	NUMBER OF ELEMENTS
MATHIEU GROUPS	
M_{11}	$2^4 \times 3^2 \times 5 \times 11 = 7{,}920$
M_{12}	$2^6 \times 3^3 \times 5 \times 11 = 95{,}040$
M_{22}	$2^7 \times 3^2 \times 5 \times 7 \times 11 = 443{,}520$
M_{23}	$2^7 \times 3^2 \times 5 \times 7 \times 11 \times 23 = 10{,}200{,}960$
M_{24}	$2^{10} \times 3^3 \times 5 \times 7 \times 11 \times 23 = 244{,}823{,}040$
JANKO GROUPS	
J_1	$2^3 \times 3 \times 5 \times 7 \times 11 \times 19 = 175{,}560$
J_2	$2^7 \times 3^3 \times 5^2 \times 7 = 604{,}800$
J_3	$2^7 \times 3^5 \times 5 \times 17 \times 19 = 50{,}232{,}960$
J_4	$2^{21} \times 3^3 \times 5 \times 7 \times 11^3 \times 23 \times 29 \times 31 \times 37 \times 43 \doteq 8.68 \times 10^{19}$
HIGMAN-SIMS GROUP *HS*	$2^9 \times 3^2 \times 5^3 \times 7 \times 11 = 44{,}352{,}000$
McLAUGHLIN GROUP *Mc*	$2^7 \times 3^6 \times 5^3 \times 7 \times 11 = 898{,}128{,}000$
SUZUKI SPORADIC GROUP *Suz*	$2^{13} \times 3^7 \times 5^2 \times 7 \times 11 \times 13 \doteq 4.48 \times 10^{11}$
RUDVALIS GROUP *Ru*	$2^{14} \times 3^3 \times 5^3 \times 7 \times 13 \times 29 \doteq 1.46 \times 10^{11}$
HELD GROUP *He*	$2^{10} \times 3^3 \times 5^2 \times 7^3 \times 17 = 4{,}030{,}387{,}200$
LYONS GROUP *Ly*	$2^8 \times 3^7 \times 5^6 \times 7 \times 11 \times 31 \times 37 \times 67 \doteq 5.18 \times 10^{16}$
O'NAN GROUP *ON*	$2^9 \times 3^4 \times 5 \times 7^3 \times 11 \times 19 \times 31 \doteq 4.61 \times 10^{11}$
CONWAY GROUPS	
C_1	$2^{21} \times 3^9 \times 5^4 \times 7^2 \times 11 \times 13 \times 23 \doteq 4.16 \times 10^{18}$
C_2	$2^{18} \times 3^6 \times 5^3 \times 7 \times 11 \times 23 \doteq 4.23 \times 10^{13}$
C_3	$2^{10} \times 3^7 \times 5^3 \times 7 \times 11 \times 23 \doteq 4.96 \times 10^{11}$
FISCHER GROUPS	
F_{22}	$2^{17} \times 3^9 \times 5^2 \times 7 \times 11 \times 13 \doteq 6.46 \times 10^{13}$
F_{23}	$2^{18} \times 3^{13} \times 5^2 \times 7 \times 11 \times 13 \times 17 \times 23 \doteq 4.09 \times 10^{18}$
F_{24}'	$2^{21} \times 3^{16} \times 5^2 \times 7^3 \times 11 \times 13 \times 17 \times 23 \times 29 \doteq 1.26 \times 10^{24}$
HARADA GROUP F_5	$2^{14} \times 3^6 \times 5^6 \times 7 \times 11 \times 19 \doteq 2.73 \times 10^{14}$
THOMPSON GROUP F_3	$2^{15} \times 3^{10} \times 5^3 \times 7^2 \times 13 \times 19 \times 31 \doteq 9.07 \times 10^{16}$
FISCHER GROUP F_2 ("BABY MONSTER")	$2^{41} \times 3^{13} \times 5^6 \times 7^2 \times 11 \times 13 \times 17 \times 19 \times 23 \times 31 \times 47 \doteq 4.15 \times 10^{33}$
FISCHER-GRIESS GROUP F_1 ("MONSTER," "FRIENDLY GIANT")	$2^{46} \times 3^{20} \times 5^9 \times 7^6 \times 11^2 \times 13^3 \times 17 \times 19 \times 23 \times 29 \times 31 \times 41 \times 47 \times 59 \times 71 \doteq 8.08 \times 10^{53}$

positive integer and p can be any prime number. The variable q must be equal to the number of elements in a finite number system, and one can show that exactly one finite number system can be constructed for a set of q elements only if q is an integral power of a prime number. Hence the family names designate simple groups only for values of q equal to 2, 2^2, 2^3,..., 3, 3^2, 3^3,..., 5, 5^2, 5^3,... and so on, unless other exceptions are noted. For example, the linear Chevalley group $A_1(3)$ is excluded from the list by the conditions imposed on the values of n and q, because $A_1(3)$ is the nonsimple group of rotations of a regular tetrahedron. The integer n, which is called the rank of the family, is restricted to the values 2, 4, 6, 7 and 8 in the exceptional Chevalley groups. The Ree and Suzuki groups are defined only for values of q equal to an odd power m of 2 or 3. The number of elements in each group in a family is given by the algebraic expression or the number in the second column of each table. Here d is the greatest common divisor (G.C.D.) of the two numbers or algebraic expressions that appear in parentheses immediately following the abbreviation "G.C.D." In the examples that are evaluated in the first table the vertical bars enclosing the name of a group designate the number of elements in the group.

China's Food

After the many disruptions of recent decades the output of China's farms about equals the need for food. Whether the balance can be maintained depends on how the nation copes with some problems

by Vaclav Smil

The question is of global importance: How well can China, farming a mere fifteenth of the earth's arable land, feed more than a fifth of the world's population? Compounding the issue is the fact that much of the farmed land is exposed to frequent extremes of climate. My answer, based on years of close study of Chinese agriculture and its related activities, is that at the moment the nation is doing fairly well. Whether or not that achievement persists into the 21st century will depend on how the Chinese respond to several problems, including the shift toward a private-enterprise economy, the loss of farmland to erosion and the management of the negative effects of pollution.

During the centuries of dynastic rule China's traditional farming attained admirable levels of intensity and complexity, encompassing intricate irrigation projects and crop-rotation schemes and extensive organic recycling. The traditional system was nonetheless unable to ensure adequate nutrition during recurrent floods and droughts. Famines were frequent and devastating. The situation did not improve much during the 38 years of the Nationalist republic, most of which were also years of war.

When the People's Republic of China was established in 1949, the ruling party moved in Stalin's footsteps to collectivize agriculture and to exploit the rehabilitating but still very poor countryside as the source of wealth for extensive industrialization. Collectivization began in 1953, coinciding with the start of China's First Five-Year Plan. Although collectivization created social dislocations resembling the traumatic experience of the U.S.S.R. in the 1930's, and although the farm cooperatives received meager quantities of farming inputs such as fertilizer and modern equipment, the first period of internal peace in two generations combined with fairly good weather and much hard work to produce steadily increasing harvests.

By 1957 the amount of food available per person was about 15 percent above the 1950 level. The best pre-1949 harvests and yields, which had been achieved from 1931 to 1936, were surpassed for all major crops except soybeans. Further solid growth seemed most probable.

This expectation was shattered by Mao Zedong's failed experiment of the Great Leap Forward. Mao's objective was to achieve in a short time a level of development that would normally take decades. The pace was too swift and the pressures on the economy too severe. In agriculture the result, compounded by prolonged drought in some areas and extensive flooding in others, was an extremely costly setback to the nation's development. How costly it was the Chinese did not say. Outside observers could only guess, because by 1958 the usual economic and demographic statistics were nowhere available; indeed, they did not reappear until more than two decades later.

Only then could John S. Aird of the U.S. Department of Commerce and Ansley J. Coale of Princeton University independently reconstruct what had happened. They concluded that a famine lasting from 1959 to 1961 had caused at least 16.5 million excess deaths. More detailed demographic data released by the Chinese since 1981, together with the results of the 1982 census (which was by far the most reliable Chinese census), put the number of excess deaths in that period at 30 million and the number of lost or postponed births at about 33 million. No other known famine has been so devastating.

There was a fast recovery in the 1960's, but it was followed by the turmoil of the Cultural Revolution. During that decade-long period the official propaganda repeatedly claimed grand agricultural achievements. China was said to have solved its food problem. Such claims along with foreign impressions based on fleeting visits to a few showplace communes were soon deflated. The first domestic critical assessments of the record after 1957 appeared late in 1977; the first new set of official national statistics came out in 1979. By the end of 1979 the new spirit of "seeking truth from facts" opened the gates to a veritable flood of critical appraisals.

The first official statistics reported since 1958 showed that whereas industrial production had risen impressively by 1978, the availability of food per capita had not changed for a generation. Nicholas R. Lardy of the University of Washington took note of this remarkable incongruity, pointing out that China appeared to represent the only case of a modern nation where a doubling of real per capita national income (the equivalent of the gross national product) in 20 years was accompanied by a constant or even slightly declining average food consumption per capita.

Matters have improved considerably since 1978. The nation's agricultural output rose by an average of 8 percent per year between 1978 and 1984. The per capita availability of staple grain rose by 30 percent, that of vegetable oils nearly tripled and sugar refining and meat consumption almost doubled. Moreover, because the real per capita income of rural households increased 2.5 times in those years, peasants can now afford to eat more nonstaple foods and still reduce the portion of disposable income they spend on food by from 10 to 15 percent. Even a conservatively prepared balance sheet shows that the availability of energy and protein from food

ESSENTIAL FOODS in China are shown in approximate order of importance. They are (*a*) unmilled rice, (*b*) pickled cabbages, (*c*) lardy, nearly meatless pork, (*d*) northern green onions, (*e*) bean-curd cakes, (*f*) pears and persimmons, (*g*) Beijing duck, (*h*) fresh carp, (*i*) ginger root, (*j*) mushrooms, (*k*) rape greens and (*l*) fresh and preserved eggs. In the Chinese diet rice is by far the main cereal, cabbage is the main vegetable, pork constitutes 92 percent of the meat and carp is the main fish. Duck is the most favored poultry and green onions and ginger are the commonest flavorings in Chinese cooking; mushrooms and persimmons serve widely as delicacies.

(stated in terms of an annual mean per capita) is 25 percent above the level in 1978.

What has brought about this remarkable transformation, this rapid improvement that is unparalleled for its combination of growth rates and scope? Two factors in particular can be cited. One is a set of new reformist policies that brought farmers increased inputs, wider markets and higher prices. The policies also relaxed the emphasis on local self-sufficiency in food. That emphasis had militated against more efficient regional marketing and local specialization in crops that could be grown most advantageously.

Even more important is the policy summed up in the term *baogan daohu:* full responsibility to households. The roots of this system go back to the 1950's; its current vigorous revival started in 1979 when a few production teams (the basic units of the rural commune system set up in 1958) in Sichuan and Anhui provinces contracted with households to grow specified quantities of crops or to raise a certain number of animals. After fulfilling the quotas and selling the products at the state's fixed prices the households could keep the surpluses, sell them to the state at a higher price or put them on the free market.

By 1980 the practice had spread to several other provinces; in 1982 contracting households encompassed a majority of villages. By 1984 *baogan* was the norm for all but a small percentage of China's peasants. This de facto privatization of Chinese farming was the beginning of what may be the most far-reaching and orderly socioeconomic transformation of the 20th century. The prospects of 800 million peasants shifted quickly from subsist-

OTHER COMMON FOODSTUFFS in China are (*a*) noodles, here fresh ones being dried, (*b*) "mile long" beans, (*c*) sprouted beans, (*d*) mangoes, (*e*) mandarin oranges, (*f*) dried fish, (*g*) water chestnuts, (*h*) winter melon and (*i*) star anise, which is a popular flavoring.

ence stagnation to vigorous expansion and incipient prosperity.

The land is still collectively owned and the state continues to set output targets for basic commodities, but *baogan* links rewards with output and encourages initiative, innovation, investment, efficiency and risk taking by individuals. As Frederick Crook of the U.S. Department of Agriculture has pointed out, the practice gives families a degree of control over their well-being and hence a pride of stewardship: the critical ingredient of farming life that was denied to them for more than a quarter of a century.

As the *baogan* system spread out through the country it was extended from simple contracts for field crops and livestock to all conceivable rural activities, creating countless new job opportunities. Specialized households emerged in large numbers. This trend represents a highly significant move away from subsistence farming to concentration on local advantages and individual skills. Most of the households still grow grain, but since 1982 millions of families have left cereal farming for the production of such specialty foodstuffs as peanuts, mushrooms and oysters. There has also been a rapid expansion of specialized services. Among them are such previously unthinkable options as private trucking of food to cities and aerial spraying of crops. One also finds new providers of information: local and itinerant consultants who advise the new agricultural risk takers on how to fertilize, breed and manage in more rewarding ways.

Yet another effect is that for the first time since 1953 peasants can leave the land and are even encouraged to do so, although migration to large cities remains restricted. The rising productivity associated with *baogan* has been releasing millions of farm workers, who are being absorbed into expanding local service and industrial enterprises and are moving to nearby towns and smaller cities, where they work in the booming construction industry, offer an array of services (from peddling household articles to making bean curd and sesame paste) and open snack shops, teahouses and restaurants. China's rural-urban ratio, congealed for more than a generation, has begun to shift. In 1955 it was 87 : 13 and in 1975 it was still 87 : 13; now it is approximately 75 : 25.

The reforms are still widening. As of this year peasants no longer have to sell part of their harvest to the state, which is gradually giving up its exclusive claim on basic foodstuffs and letting demand determine output. Another change is in contracts for cropland, which had usually been for short terms. Now many of them have been extended for up to 15 years in order to encourage soil-conservation measures and to ensure the proper maintenance of soil fertility. In the case of forestry the contract period can range from 30 to 50 years, and the right to inherit the trees is guaranteed. *Baogan,* as vice-premier Wan Li has said, is not merely an expedient policy to solve the problem of feeding the people but a fundamental reform of the economic system in the countryside.

The consequences of this reform for nutrition can now be evaluated with a fair degree of quantitative accuracy. Even though the recent flood of official statistics contains some discrepancies (and it would be naive to assume that inflated reporting, a ubiquitous practice in centrally run economies, has been eliminated), the new statistics appear to be surprisingly candid. They have good internal consistency, correspond in general to the observable realities and include plenty of unflattering details. Hence an economist can believe that for the first time in three decades he has an adequate basis for a realistic appraisal of agricultural performance.

The new data allow one to draw up fairly reliable food-balance sheets: comprehensive accounts of output, trade and changes in stock for all important plant and animal foodstuffs, taking into account nonfood uses and losses in storage and transportation. This exercise yields a picture of the average approximate availability of food per household per day.

I have prepared food-balance sheets for six key years of the People's Republic. They show a rise from 1,800 kilocalories per person per day in 1950 to 2,100 in 1957 (the end of the First Five-Year Plan), a plunge to 1,500 during the depth of the 1960 famine, recovery to 2,000 by 1965 (the last "normal" year before the decade of the Cultural Revolution), a small rise to 2,100 by 1977 (the last year before reform) and a big increase to 2,700 by 1983.

Is this much food finally enough? The calculation of nationwide food-energy requirements is prone to error; to be meaningful the result must be expressed as a range rather than a single value. To calculate energy requirements I used China's age-sex distribution and occupational data from the 1982 census, average adult weights of 60 kilograms for men and 50 for women, measurements of peasant energy expenditure made by the Chinese Academy of Medical Science and various assumptions about the prevalence of heavy physical exertion. The result is an average daily per capita requirement ranging from 2,200 to 2,400 kilocalories. (Calculations by the Food and Agriculture Organization put the average daily need in China at 2,360 kilocalories.)

Recent Chinese government surveys of food consumption (the one in 1983 included 9,060 urban and 30,427 rural households) make it possible to determine actual intakes with unprecedented reliability. The weighted nationwide average works out to a range of between 2,200 and 2,380 kilocalories per capita in 1983, or from about 19 to 12 percent below the mean availability of food.

The best conclusion I can offer—and the reader must bear in mind that this is a realm lacking quantitative certitude even in the Western nations—is that there is now a basic balance between need and average food consumption. The conclusion implies a significant achievement: since 1982 China's farming has provided on the average enough food to cover energy needs that are compatible with normal growth and with a healthy and vigorous life.

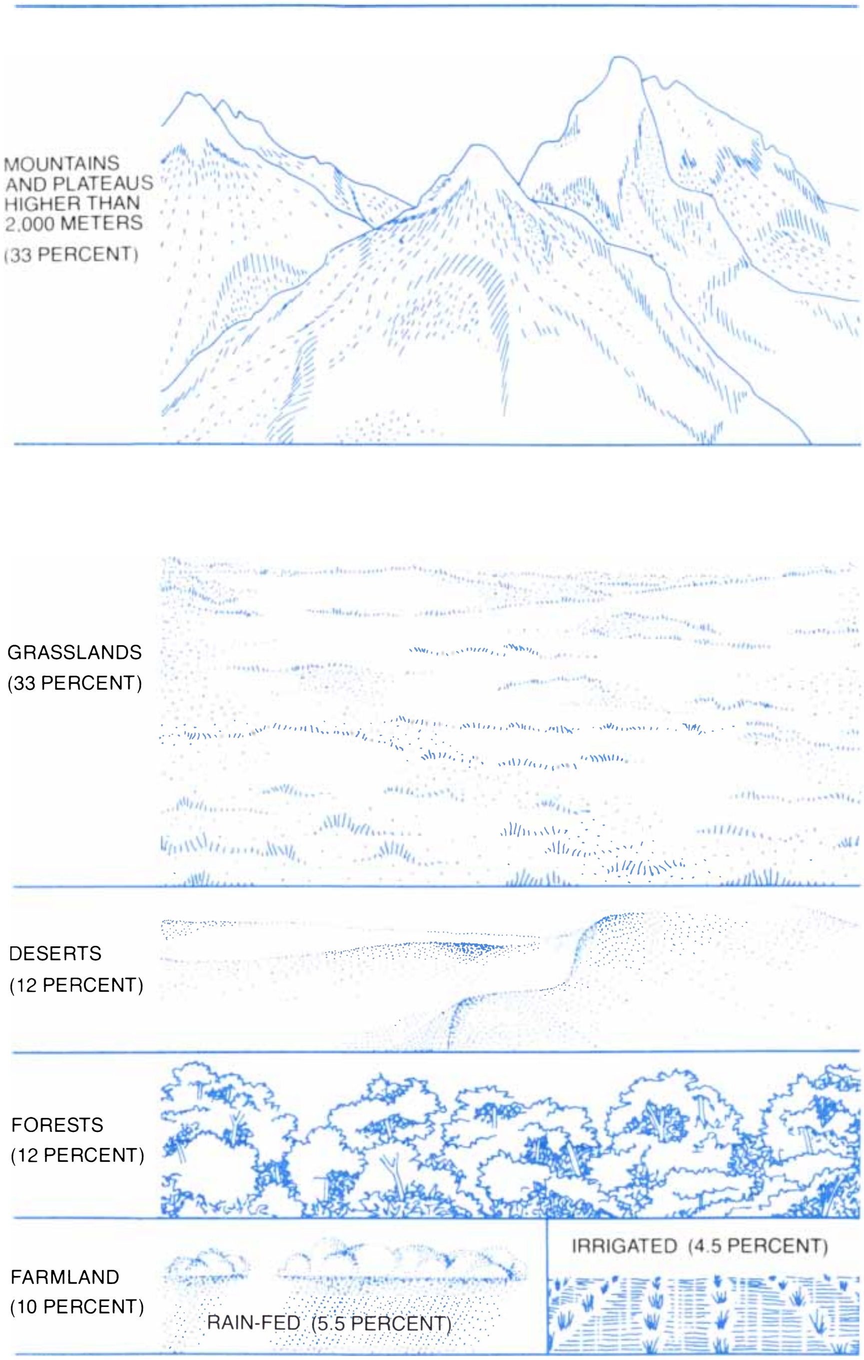
The situation with regard to protein is similar. My calculation—taking into account the fact that with virtually no milk, few eggs and a relatively small amount of mostly fatty meat the Chinese must get most of their protein from plants, mainly rice, wheat and coarse grains—is that a daily mean of at least 60 grams of protein as actually consumed in an average Chinese diet is needed to cover the growth and maintenance requirements. Recently published Chinese values stipulate 75 grams per day as adequate. In any case at least 75 grams have been available since 1982, making it possible to conclude that, on the average, protein needs are well covered.

"On the average" is, of course, the key phrase in both energy and protein assessments. Regional, local and individual disparities in the standard of life have undoubtedly increased since 1978, and nutritional inequalities are far from insignificant.

China's leaders are well aware of these disparities. In a speech published last January, Deng Xiaoping, the principal decision maker, acknowledged that "there are still some tens of millions of peasants in the countryside who do not have enough food." An article in *Liaowang* ("Outlook") in 1984 spoke of "11 percent of the rural population who still have not resolved the problem of dressing warmly and getting enough to eat." The figure implies that some 90 million people needed substantial nutritional improvement as of 1984.

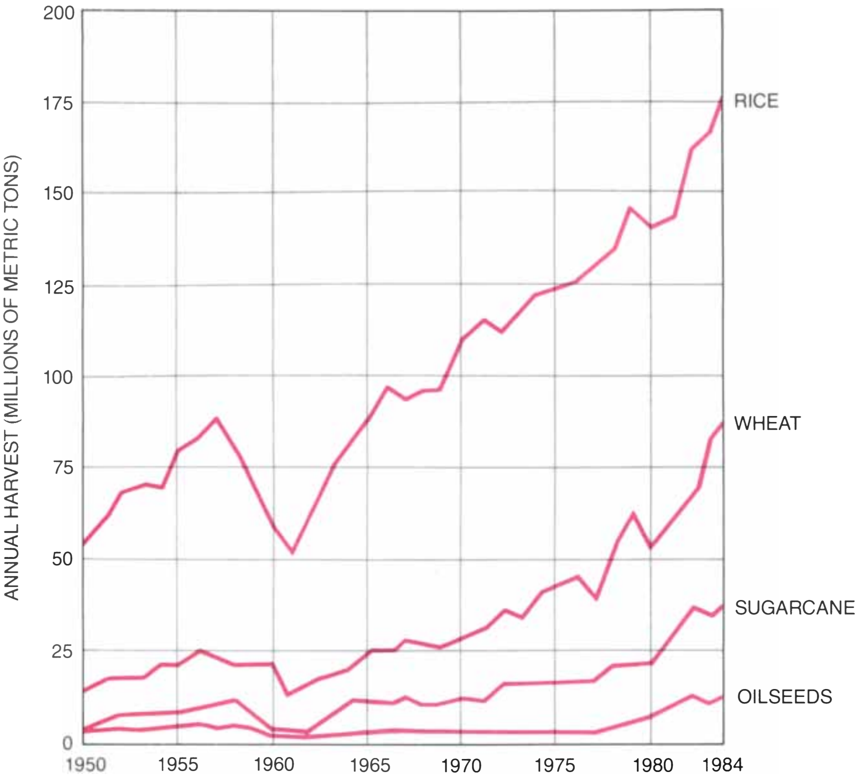
Moreover, differences between urban and rural diets are still considerable. Urban people may be envious of the newly rich farmers, particularly of their large new houses, but on the average the people in cities still eat much better. As in other places, better urban eating in China is marked by lower consumption of grains (in 1983 only about 75 percent of the rural mean) and higher intakes of fat (85 percent more plant oils than in the villages), meat (about twice as much), eggs (4.5 times as much) and sugar (more than twice as much). Even in the cities, however, the diets still provide too little high-quality protein and claim more than half of the disposable income. (The average in the cities is now about 60 percent.)

Another component of the food picture is a change in the nature of

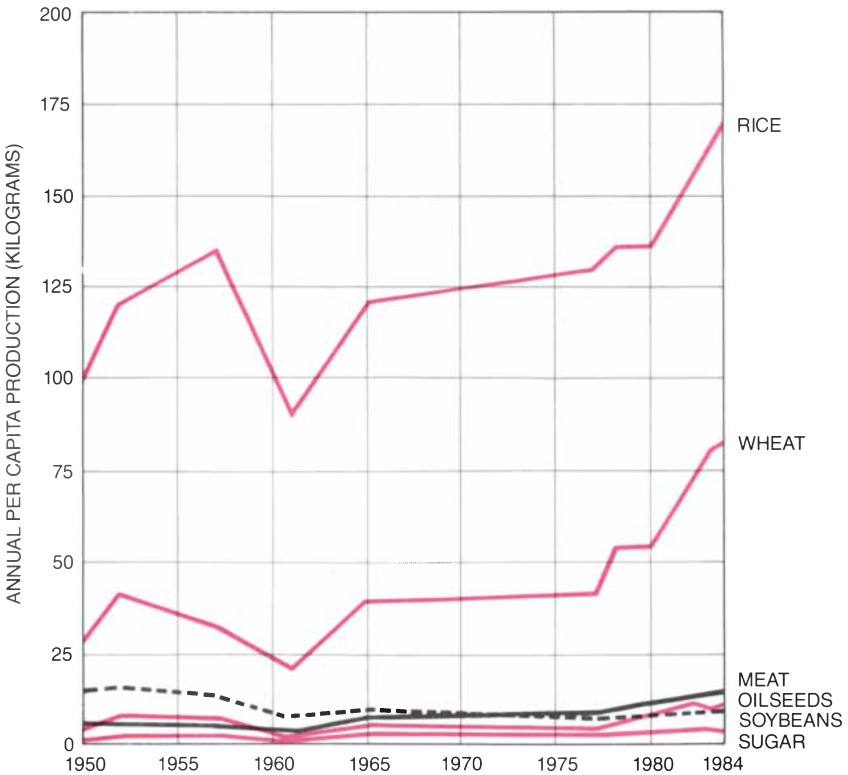


AGRICULTURE IN CHINA is limited to about a tenth of the land area. Environmental conditions are the circumscribing factors. For example, the land higher than 3,000 meters (and much of the land higher than 2,000 meters) is largely unproductive (*map at right*). A third of the land area is higher than 2,000 meters. Most of the grasslands are extremely dry. The map also identifies the Chinese provinces and the two major rivers: the Huang He (Yellow) and the Chang Jiang (Yangzi). The chart above shows how the land is apportioned.

the harvests. Fine grains, milled rice and wheat flour, which previously accounted for between 55 and 60 percent of the total cereal output, now represent about 65 percent of the grain harvest. Hence even rural households with average income can obtain staples in proportions available before 1978 only in privileged city rations. (Whole grains and high-fiber diets may be fashionable in the West, but the Chinese greatly prefer well-milled rice and white flour—particularly after eating the detested coarse grain and "black" rice for two decades.) With most meals vegetables remain the only accompaniment of staple grains. The seasonal variety of vegetables available from private sellers has increased considerably in recent years. Still, during the winter months the choice in the north comes down to not much more than cabbages, onions and turnips, and daily vegetable consumption is often



INCREASING HARVESTS reflect the recent gains (somewhat uneven) in Chinese agriculture. Severe declines in harvests brought about by Mao Zedong's Great Leap Forward caused a major famine in 1959–61. The harvest of grains recovered rapidly, but the harvests of nonstaple crops, notably oilseeds, started to rise only after agricultural reforms were instituted in the late 1970's. A key reform is the trend toward private-enterprise farming.



PER CAPITA HARVESTS of major crops in China have also improved. Contrasting, however, with the growth in total food production, the per capita harvests were below the best achievements of the 1950's until as recently as 1977. Since then most have improved.

only a third of the summer average.

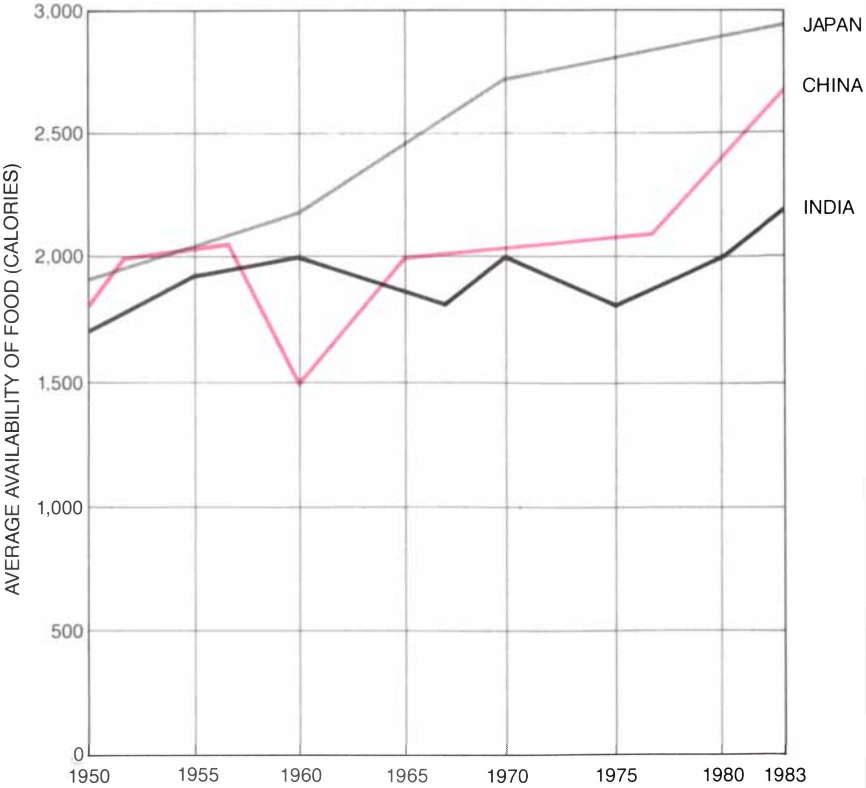
The most welcome gain is that about four kilograms of edible oils is now available per capita per year. Oils were in chronically short supply before 1978 as the cultivation of cereals displaced rapeseeds, peanuts, sunflower seeds and sesame seeds. Even with imports of soybean oil the availability of oil kept declining; in some cities the ration fell below 100 grams per month. Today's urban consumption of about 500 grams per month is still barely adequate. (It takes at least 30 grams of oil to stir-fry just one meat or vegetable dish for an average family meal.)

Animal foods provide only about 6 percent of food energy and 15 percent of protein. They amount to about 25 kilograms per year per person. Pork accounts for more than 90 percent of the red-meat supply. The pork comes overwhelmingly in fat chunks; the rising demand for lean cuts will not be satisfied for years to come. Specialized households have been increasing the output of freshwater fish by from 10 to 15 percent per year since 1980. Fish and shellfish, consumed at a rate of more than three kilograms per capita per year, constitute China's second most important source of animal protein. Beijing duck (or even a plain steamed chicken) is a treat: annual poultry consumption averages less than three kilograms even among the 75 million people in the largest cities. Even though the consumption of eggs has risen by 50 percent since 1978, it still works out to less than one large egg per week per person.

Impressive as all the gains in Chinese agriculture are, the real question is whether or not the upward trends maintained since 1978 can be sustained, not only in quantity but also in quality. The obstacles are enormous. A positive reply to the question is conditional on the moderation and substantial removal of those obstacles.

To a Western economist the irrational nature of China's price system must appear as the foremost hindrance. That view is now shared by China's top reformist leaders, who are focusing increasingly on the necessity of reforming the procurement system while lowering the prices farmers pay for key farming inputs and raising the urban retail prices of farm products. As Nicholas Lardy concluded in his analysis, in the long run neither the communes nor *baogan* farming will be capable of supporting a sustained growth in production if they remain confronted by distorted prices and restricted markets.

Perhaps the best illustration of the need for price reforms was the return



THREE ASIAN NATIONS show different patterns in the availability of food per capita, starting from similar positions in the 1950's. Units are in kilocalories per day per capita.

of pork rationing to China's large cities last February. Low purchase prices for pigs made pig farming a losing enterprise, and peasants turned to more rewarding angora rabbits or cereals. Another flaw became apparent when the country could not handle the record grain harvests of 1983 and 1984. Having run for years on bare local self-sufficiency, farmers were literally unable to get rid of their grain in the absence of adequate storage capacity and efficient interprovincial transport. Reports of large quantities of grain rotting in the open came in for months. Because the handling of grain harvests is inadequate and the country's feed industry is primitive, the system supplies less than five million tons per year of the quality mixed feeds needed for efficient production of lean pork and poultry. (The output of the U.S. feed industry is close to 200 million tons per year.)

The general index of state purchase prices (the payments to farmers) rose by 54 percent between 1977 and 1983 in order to encourage production, but the index of retail food prices went up only 15 percent. The gap is now so large that the state's subsidies to hold down retail food prices claim about 15 percent of total government expenditures. Large price increases last spring—70 percent for lean pork, 35 percent for ordinary pork, 100 to 300 percent for some kinds of fish—represent just a few of many needed price adjustments. Unusually low energy prices, which are a key cause of China's dismally high energy consumption and which distort the costs of essential farming inputs, will be harder to tackle.

Even given responsive pricing and flexible marketing, China's farming faces the enormous challenge of maintaining a sustainable agricultural ecosystem. The former one-sided emphasis on grain farming and local self-sufficiency in food can be seen to have been a reaction to the great famine of 1959–61, but the consequences of that policy have been severe. In a land where the forest cover was already low the "grain first" drive of the 1960's and 1970's led to indiscriminate deforestation. The provinces affected most—Sichuan and Yunnan—lost respectively 30 percent and 45 percent of their forest in a single generation. The reclamation of northern grasslands for grain led to advancing desertification. The filling in of lakes in alluvial lowlands for the same reason worsened seasonal flooding. After 1978 these ruinous steps received widespread condemnation from Chinese scientists pressing for sensible management of the agricultural ecosystem. The filling in of lakes and the conversion of grassland have for the most part stopped, but deforestation continues.

China's opportunities for obtaining more crops per year from the land are limited. The nationwide average of the multicropping index, in which 100 represents one crop per year per hectare, is already above 150. It ranges from about 100 in the dry and cold northwest to almost 220 along the lower course of the Chang Jiang (the Yangzi, formerly transliterated as Yangtze). The triple-cropping of grain, which was indiscriminately promoted in the east and south before 1978, can in fact be sustained in only a few parts of those regions.

For these reasons the conservation of farmland should be a priority. Yet China has lost about 11 percent of its arable land since 1957 through erosion, desertification and nonagricultural development. That is equivalent to all the farmland in Illinois. Erosion has been worsening even in the Yangzi basin, which was once relatively unaffected. The amount of farmland per capita is now .1 hectare, equal to that of Bangladesh and only two-fifths of India's average.

To sustain high yields and protect against erosion the Chinese must recycle organic matter vigorously. In traditional Chinese agriculture they did, but increasing amounts of wastes that are toxic or not biodegradable and greater reliance on synthetic fertilizers have undermined the ancient practice to the detriment of soil fertility and good tilth. The recycling of crop residues—essential for soil maintenance, retention of moisture and protection against erosion—is particularly inadequate. Whereas farmers in the U.S. recycle approximately 70 percent of their crop residues (mainly straw), China's acute shortages of energy in rural areas claim up to 75 percent of the straw for fuel.

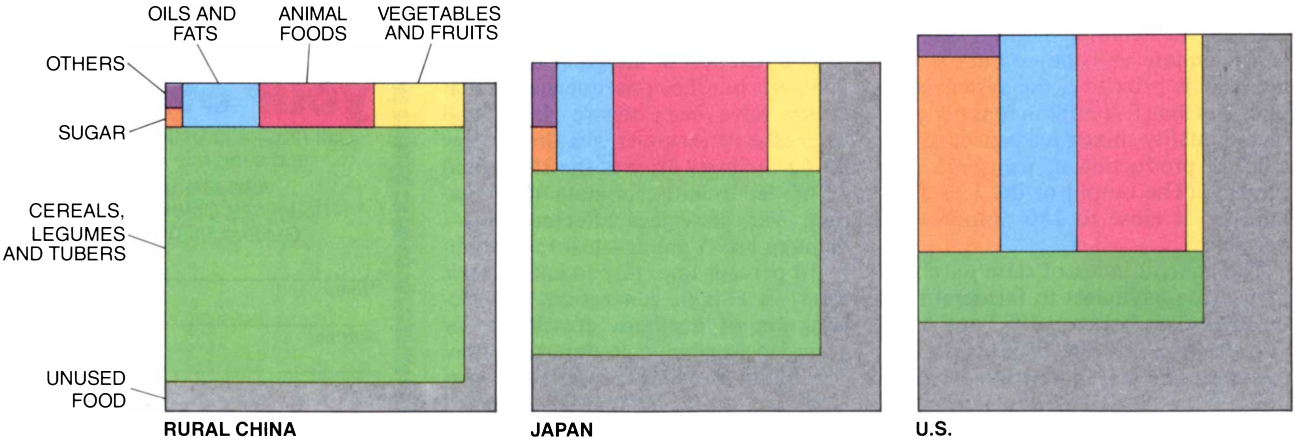
An additional worrisome degradative process imperiling the agricultural ecosystem is water pollution. It arises from mostly untreated urban wastes, from large industries, from the nearly five million small enterprises now located in villages and increasingly from nitrogenous fertilizers. Because nearly half of the farmland is irrigated, waters polluted by heavy metals, oil wastes, phenolic compounds, chlorinated hydrocarbons, arsenic and nitrates present a widespread danger of food contamination as well as a direct hazard in drinking. The large expansion of economic output planned for the next several decades can only heighten the pollution.

China also faces problems in satisfying the demand for irrigation water in the dry north. The nation now devotes about 400 billion cubic meters of water per year to irrigation (more than 80 percent of all water consumption). By the year 2000 this amount is expected to rise by one-third and the total use of water by one-half. That level of use would represent about a fourth of the nationwide runoff. In most of the northern watersheds, however, about two-thirds of the runoff is tapped already, and entire provinces have large water deficits.

It is inevitable that China will experience these natural constraints more acutely in the decades ahead. Although they can be significantly moderated by increasing the efficiency of water use, maximizing the practice of growing the crops that flourish best in each area, stressing organic recycling, instituting strict measures to protect farmland and increasing the quantity and quality of farm inputs, each of these steps carries its own considerable price. Increased efficiency of irrigation, for instance, calls for a massive investment in lined canals, sprinklers and center-pivot systems. Growing crops that take the most advantage of local conditions is impossible without flexible marketing and first-rate facilities for storage and long-distance transportation. Proper recycling of crop residues is unthinkable until the stunning rural energy crisis is eased, largely by the newly established private fuelwood groves.

Additional losses of prime farmland are inevitable, if only to accommodate the population increase. Moreover, increased farm inputs will not come cheap. I have calculated that the Chinese are already investing about 40 percent more energy per hectare than farmers in the U.S., in no small part owing to low efficiency of energy conversion.

Dwight H. Perkins of Harvard University and Shahid Yusuf of the World Bank ask the fundamental question at the end of a recent broad analysis of China's rural development: "How much control is the leadership willing to give up in exchange for how much of a rise in productivity?" The issue is even deeper: Will the thrust of the reformist policies be maintained? Deng Xiaoping himself acknowledges that convincing the people of the irreversibility of the new policies "is quite a problem." One must hope that the reforms will not only continue but also deepen further, both for the better life they would provide for the Chinese people and for the contribution to world stability that such an evolution would represent.



AVAILABILITY AND CONSUMPTION OF FOOD are compared for three populations in terms of kilocalories per day per capita. The difference between availability and consumption is smallest in China, where grains dominate the diet. The patterns are notably different in Japan and the U.S. "Unused food" includes food lost in transportion, storage and processing and through table waste.


GMC
The classic no enthusiast should be without.
A classic truck? Well, sure.
The pickup is a classic of sorts, with over five decades of service to mankind to its credit. But the GMC S-15 4X4 Pickup is a classic departure from trucks as they are commonly perceived.
Its rakish looks and sophisticated comforts allow this GMC truck to do virtually anything a car can do. And its Insta-Trac four-wheel drive, double-wall cargo box and optional V-6 engine allow our pickup to do a lot of things a car could never do.
The GMC S-15 4X4 Pickup. If you're into fine cars, it's time you discovered fine trucks. Buckle yourself into one soon at your nearby GMC Truck dealer. And please help preserve America's natural beauty whenever you're out four-wheeling.
GMC
TRUCK
A truck you can live with
GM

The Construction Plans for the Temple of Apollo at Didyma

The nature of the "blueprints" from which the Greeks built their temples has long eluded archaeologists. A recent discovery shows they were drawn on stone surfaces of the very temple they depict

by Lothar Haselberger

How did the Greeks design and build their classical monuments, which in the course of time have repeatedly been regarded as exemplary? What skills did the erection of these imposing marble edifices require? On what aesthetic theory were they based? The answers to these questions were sought by archaeologists for more than 200 years. Although it is known that the Greek architects themselves wrote detailed explanations of the conception and construction of their buildings, none of these written works has survived to the present day. No other sources reveal what kind of construction plans (if any) were followed by the ancient builders. In the absence of any descriptions of temple "blueprints" scholars had no idea where to look for them.

Actually the blueprints were under the archaeologist's nose all along. I recently discovered an entire archive of construction plans still in place at a site that has been under study since the turn of the century: the famous ruin of the Temple of Apollo at Didyma, south of the present-day Turkish town of Söke.

During a tour through the lands of classical antiquity that the German Archaeological Institute sponsors annually for young archaeologists, I took a side trip to the temple of the Didymaean Apollo. It was on that excursion in October, 1979, that I first came across many finely etched lines on some of the temple walls. What first sparked my curiosity during that casual visit to Didyma became the subject of a major archaeological project: it developed that the lines trace out the design of various temple structures. The incised blueprints cover an area of hundreds of square meters and constitute the most extensive and most complete set of plans that have come to light in all of ancient architecture.

The construction of the Temple of Apollo, the Greek god of light, art and prophecy, was begun under the patronage of the powerful trading city of Miletus, in the southernmost part of ancient Ionia. The temple, which was to house the most famous oracle in the eastern reaches of the Hellenic world, was meant to replace an earlier building (on the site of an even earlier shrine) that had been destroyed by the Persians.

Planning for the temple began soon after 334 B.C., when Alexander the Great arrived in Asia Minor. The Milesian architect Daphnis collaborated with one of the leading architects of Ionia, Paionios of Ephesus, to design the ambitious structure. Paionios at the time was completing his work on the renowned Temple of Artemis in his native city-state. This Ephesian temple, which was the largest Greek temple up till then, was numbered among the Seven Wonders of the ancient world. The new Temple of Apollo was foreseen as equaling the Temple of Artemis in size and splendor.

The realization of the plan, however, progressed very slowly, and the construction work was suspended after 600 years. The temple remained in this unfinished state until the Middle Ages, when an earthquake demolished most of what had been erected.

Even as a ruin the great marble edifice impressed me from the start. Three of the columns still stand at their original height of 19.7 meters atop the temple's stepped platform, which is 120 meters long and 60 meters wide. One of the columns is unfinished, in its mantle of rough-hewn stone; the other two were completed and still support part of the stone entablature, the beam on which the roof was meant to rest. A double ring of these tall columns (108 in all) was once intended to encircle the equally tall walls of the temple's inner sanctum. Twelve additional carved columns of similar size and form, arranged in three rows of four, stood in the portico, which faces roughly eastward. The 10 columns of the facade were even more richly carved than those of the portico. Most of the columns were in position at the time construction was halted.

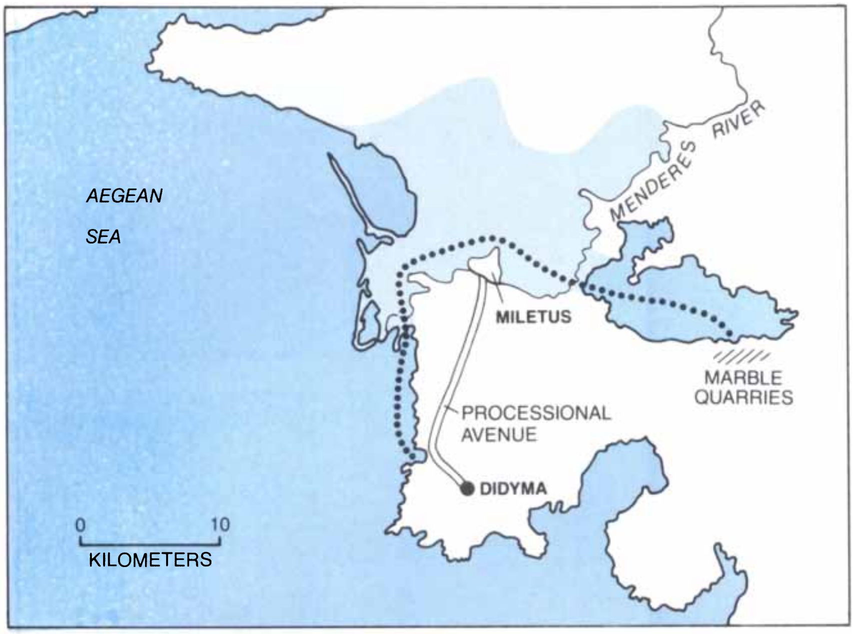
Behind the forest of columns at the front of the temple towered an immense doorway called a manifestation portal. This doorway, however, was obstructed: its threshold was raised so high above the floor that one could not step over it. On the other side of the portal, from an elevated, stagelike hall oracular pronouncements were made.

One might suppose the double row of marble columns along the temple's length was meant to support a majestic roof spanning the entire width of the temple. Actually there is a huge open-air courtyard with no flooring in the temple's interior. This was the adytum, or "forbidden" area, one of the most remarkable enclosures ever created by Greek architects.

Here, near a laurel tree under which Apollo had reputedly been conceived, a sacred oracular spring rose from the ground. Over the spring stood a small, richly worked inner temple, the *naiskos,* which contained a bronze statue of the god. Although the temple was razed in the early Christian Era, it was possible to reconstruct it on paper from the hundreds of fragments.

In order to enter the adytum (something only a select few were allowed to do in antiquity) it is necessary to go through one of two narrow, dark, steep tunnels that lie under each side of the manifestation portal. Each tunnel leads to a tomblike antechamber, and

RUINS OF THE TEMPLE OF APOLLO at Didyma are impressive in their own right. A frontal view (*top*) shows the three remaining columns of the more than 80 that were originally standing. The column on the left was not finished: the flutes, or vertical grooves, are missing. The other two columns still support a marble beam 5.3 meters long that was once part of the entablature. The width at the base of the temple is 60 meters. The adytum, or inner sanctum of the temple (*bottom*), was a large, unpaved courtyard sunk into the raised temple platform. It was accessible only through two narrow tunnels, the exits of which are on each side of the stairs. The stairs led to the stagelike "manifestation portal," from which oracular pronouncements were made. Many of the marble fragments that lie scattered are from the *naiskos*, the small inner temple. The rectangular foundation of the *naiskos* is clearly exposed (*left foreground*). The foundations of earlier structures have also been uncovered. Inscribed on the smooth lower walls of the adytum are full-scale plans (imperceptible here) for many temple structures. The upper walls have been partially rebuilt to about a third of their original height.



DIDYMA was situated in the southernmost part of ancient Ionia. The small map (*left*) locates the enlarged area (*above*) in relation to modern Greece and Turkey. Didyma's patron city, Miletus, was connected to it by a 20-kilometer processional avenue. The marble blocks for the temple were quarried in the mountains southeast of Miletus and transported over a considerable distance by sea (*dotted line*). Miletus is no longer a port: the region shaded in light blue is now land.

as one steps out of these chambers into the adytum one is momentarily blinded by the sudden light.

The podium walls along the base of the adytum alone are enough to fascinate the visitor. Indeed, it was the impeccable precision with which the walls' stone blocks had been fitted together that initially intrigued me. It was also here, on the wall only a few steps from the tunnel exit on the north side, that a large number of finely etched lines first caught my eye. I was particularly struck by the narrow, regular gaps of from 1.8 to 1.9 centimeters between them.

The glare of the sunlight reflected from the white marble wall dazzled me, and I could not closely examine the lines that first day. I returned the next day, when the light shone on the surface at a more favorable angle. This time more lines became visible on the marble surface, and half circles and quarter circles as well. The discovery forced me to spend more days scrutinizing the wall surfaces in the adytum and elsewhere among the ruins.

I found that a delicate web of overlapping lines and geometric constructions covers most of the podium walls encircling the adytum. Other smooth walls among the ruins that were originally within arm's reach also bear such lines. All told the drawings cover an area of approximately 200 square meters.

There are straight lines up to 20 meters long and circles whose radii extend as far as 4.5 meters. Parallel lines, polygons and subtending angles have been constructed, and distances have been accurately subdivided. The lines are as thin as a pencil mark and are incised a little more than half a millimeter into the marble surfaces. The lines and curves had been meticulously traced by means of a fine metal gouge guided by either a long straightedge or dividers. Imprecise lines can also be seen, but they were always corrected.

The lines, like the wall surfaces on which they were incised, have been affected surprisingly little by exposure to the elements. They have been eroded in places by rainwater and have also been covered by sinter, the thin mineral deposits left from the evaporated water. In some cases the stone surface has flaked and lines have been lost. Nevertheless, in many places the lines look as though they had just been drawn. Yet even where a number of legible lines exist the pattern they form as a coherent whole cannot be perceived: as one steps back for a larger view, the lines quickly become indistinguishable.

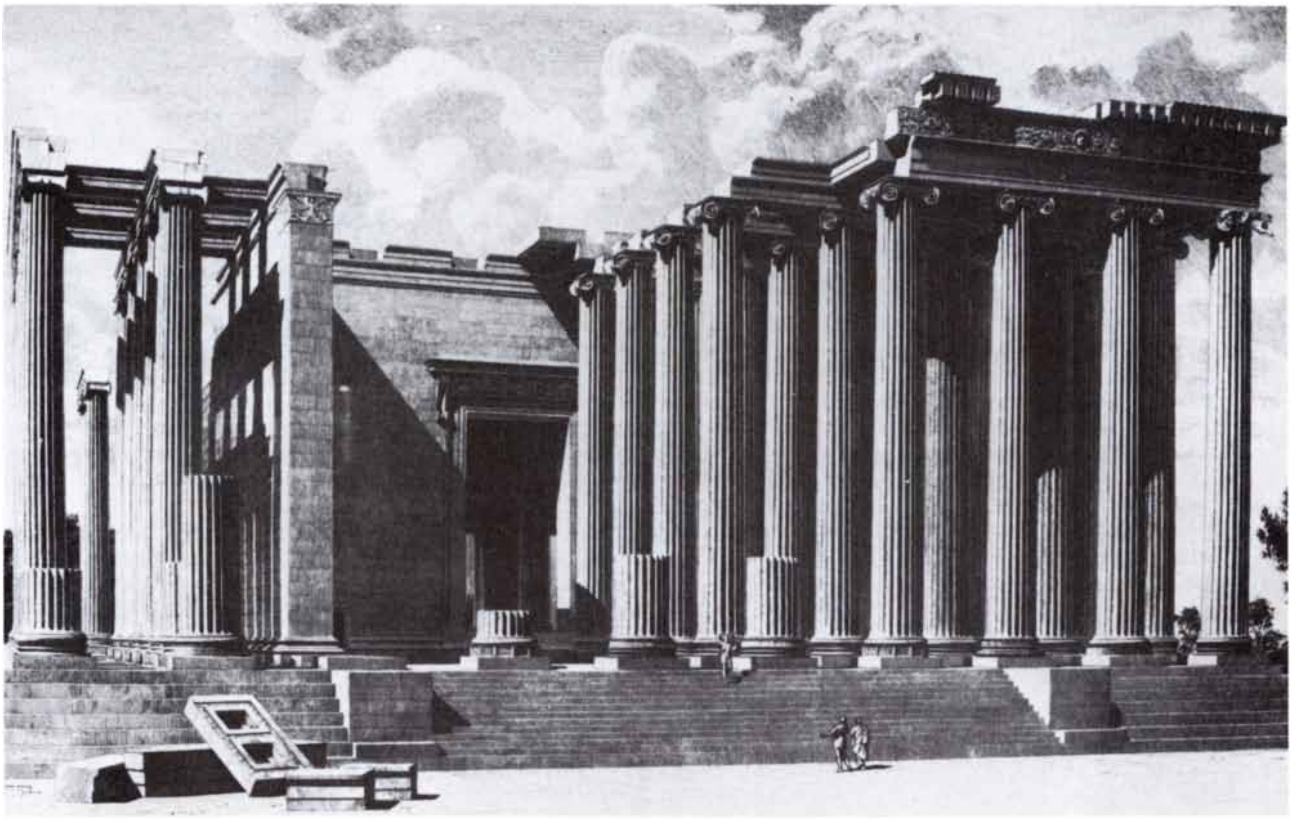
Numerous traces of reddish mineral pigment that have been preserved on these walls (mainly under sinter deposits) indicate that the drawing surface had originally been colored with red chalk to make the diagrams stand out. The chalk must have been applied before the drawings were made, so that the incised lines showed white against a dark red background and produced a needle-sharp outline. Moreover, the application of more chalk made it easy to "erase" and correct.

It was clear to me from the beginning that these sketches could not be dismissed as meaningless scrawls. On the basis of both their quality and their quantity they could only be regarded as the work of practiced draftsmen.

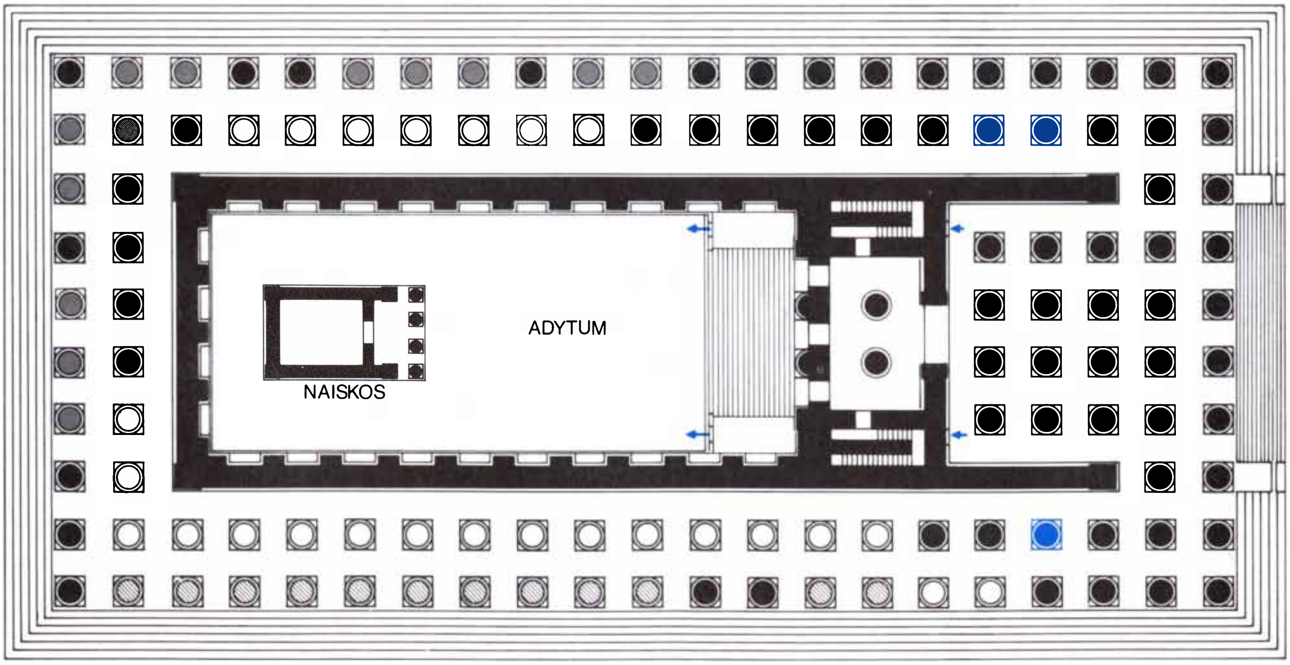
Today these finely carved lines cannot be seen with any clarity unless sunlight casts shadows into them; one can either wait for the right time of day or reflect the light with mirrors. Because the accuracy of detail cannot be attained by any photographic process, an overall image of these lines could only be formed by redrawing the lines to scale after making careful measurements.

The task of reading, measuring and recording the lines was begun in 1980 as part of a comprehensive study of Didyma directed by Klaus Tuchelt and sponsored by Edmund Buchner, president of the German Archaeological Institute. In the course of 39 weeks of meticulous and time-consuming work I have been able to get about half the drawings onto paper.

Once the first complete drawing had been transcribed the biggest question posed by the discovery was quickly answered: What do the drawings represent? If one compares the outline, or profile, of the base of a column with what was copied from one of the adytum walls, the similarity is immediately obvious [*see illustration on page 128B*]. This was, in fact, the Archimedean fulcrum that set us on the right track, showing us that the drawings are plans of the temple and its *naiskos*. The individual parts of the building were laid down on the wall surfaces in their full dimensions (on a scale of 1 : 1) with the utmost precision.



DRAWING OF TEMPLE as it may have looked before construction was discontinued was made by George Niemann in 1912 while the temple was still being excavated. The reconstruction includes some inaccuracies: the columns are drawn as slightly tapered but rectilinear cylinders, and the stepped platform as well as the beams are drawn as straight elements. Actually the entire temple, from foundation to roof, "hunched" slightly upward and the columns "swelled" slightly outward. In addition the columns (composed of stacked drum-shaped segments) were first erected in their entirety before the 24 flutes along their length were carved and polished.



TEMPLE FLOOR PLAN depicts the 122 columns (*circles*) envisioned for it. The columns shown in black were fully or partially erected; the same is probably true for the gray columns. The hatched columns were probably also being raised when building stopped, although this can be stated with less certainty. The columns left blank had not yet been erected. The three standing columns are colored. The *naiskos* stood in the roofless and floorless adytum. The two tunnels leading to the adytum are indicated by colored arrows.

Although the similarity between the sketch of the column base and its actual cross section is striking, there are differences too. The most conspicuous of these is at the same time the easiest to explain: the horizontal grooves in the bulbous, protruding part of the base, called the torus, are not shown in the wall drawings. The grooves were always carved after the contours of the column base had been shaped. Only in this way could the grooves be made straight and true.

Less prominent features are more interesting. The semicircular molding immediately above the torus was drawn twice, first in a forward-projected position and then set 2.4 centimeters back, in its actual position. In addition two curved lines shift the lower part of the torus inward by less than a centimeter to make the figure congruent to the actual torus cross section.

Such small changes may seem trivial, particularly in relation to the overall dimensions of the column (nearly 20 meters tall and two meters thick). Yet these adjustments are not mere bagatelles in the context of the rigid criteria of form and proportion to which the Greek architects were thought to have adhered. The framework for the design of any component of a temple was dictated by very strict and precise geometric relations.

The profile design of the column base, for example, was determined by rectangular coordinate axes [*see illustration on opposite page*]. The vertical axis was meant to just touch the surface of the column shaft, and the horizontal axis specified where along the vertical axis the torus would reach the point of farthest radial projection. Two other horizontal lines marked the top and bottom of the torus profile; the lower one was exactly twice as far from the horizontal axis as the upper one.

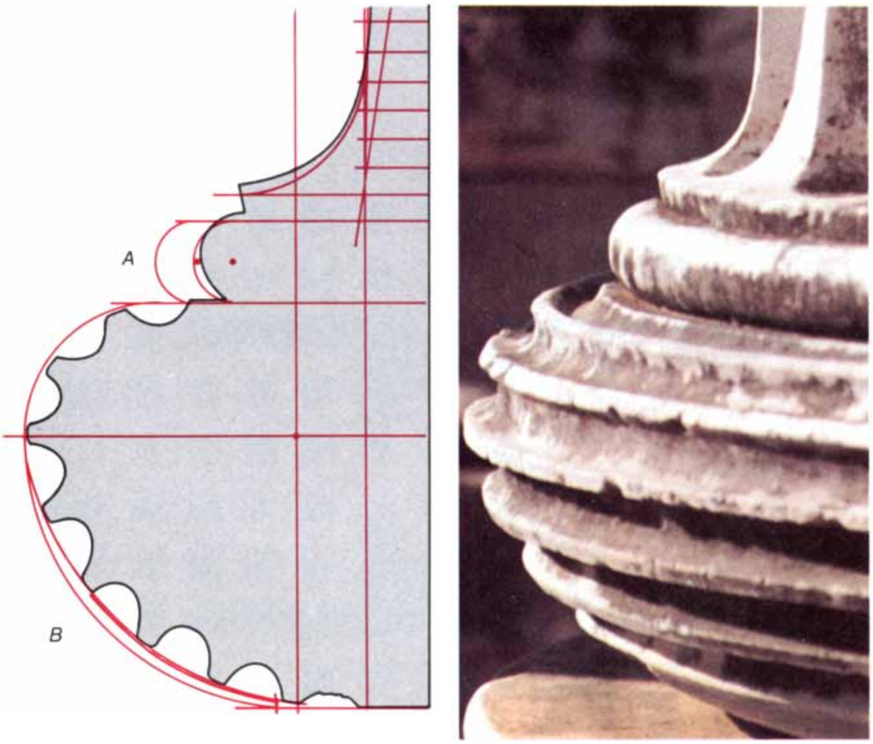
The subsequent construction of the "perfect" torus profile follows simple geometric rules and relations. A curve congruent to the rounded outer edge of the torus profile can easily be constructed by inscribing a small circle in a larger one so that the point at which they are tangent as well as their centers lies on the horizontal axis. If one traces a quarter of the smaller circle by drawing upward from the point of tangency and then traces a quarter of the larger circle by drawing downward from the point of tangency, the curve thus outlined is congruent to the torus's curve in profile. As drawn on the wall at Didyma, the larger circle is centered on the origin of the coordinate axes (the hole for the dividers is still visible) and has a radius equal to the distance between the horizontal axis and the horizontal line drawn below it (two-thirds the height of the torus). The smaller circle has a radius equal to the distance between the horizontal axis and the horizontal line drawn above it (one-third the height of the torus).

A vertical line drawn upward from the center of the smaller circle should have been tangent to the outer edge of the semicircular molding above the torus. A vertical line from the center of the semicircular molding in turn was meant to determine the curvature and point of tangency of the hollow groove that sweeps upward from it to the column shaft.

This geometrically perfect design was not the final one, however; it was further refined in a way that was inconsistent with the previous strict pattern. The semicircular molding was set back slightly from its geometrically specified position to achieve a less crowded, more distinct effect. In addition the lower part of the torus was pulled in with the help of freehand divider strokes of indeterminate radius and position, resulting in a tauter curve.

Apparently the master builder responsible for the design (there can be no question that he was a master of his art) was guided, but not bound, by the strict obligations imposed on him by the geometric design. He transcended these self-imposed rules whenever his aesthetics demanded it. On the other hand, he never fully rejected the underlying proportions of Greek design. Because the semicircular molding was set back, the vertical coordinate axis (which represented the column shaft) had to be redrawn. The new vertical axis was in fact redrawn a distance one-third the height of the torus profile (the same unit that figured so prominently in the original design process) from the center of the shifted semicircular molding.

The interacting considerations of stringent rules and dauntless whim forced the master builder to engage in an architectural balancing act between the two opposites. A golden mean between two extremes is precisely what Greek philosophers postulated as the fundamental source of beauty in the widest sense. Evidently architecture was permeated down to the smallest detail by a struggle to achieve beauty in this way too.



PLANNED AND ACTUAL COLUMN BASE can be compared in profile. The shaded outline (*left*) is a scaled cross-sectional diagram of the base of a temple column (*photograph at right*). The inscribed lines on the adytum wall are drawn here (*color*) to the same scale as the section insofar as they are visible or can be reliably reconstructed by extending visible lines. Disregarding what appear to be "first tries" at a base design, the match is close. The redrawing of the semicircular molding (*A*) and of the protruding curve below it (*B*) indicates that the sketches on the wall were drafts, the final versions of which were then executed.

It is clear that these working drawings were used to elaborate the component parts of the temple in an often highly involved process of design. Geometrically pure paradigms were drawn first, but the architects did not hesitate to alter them to suit their refined sensibilities. Construction work then proceeded according to the result. The visible sequence of design, revision and production indicates the drawings were definitely not incised into the walls to illustrate what had already been built. A more accurate dating of these drawings can thereby be determined. The construction plans must have been drawn before the individual components pictured were fashioned but after the stone "drawing boards" were smoothed.

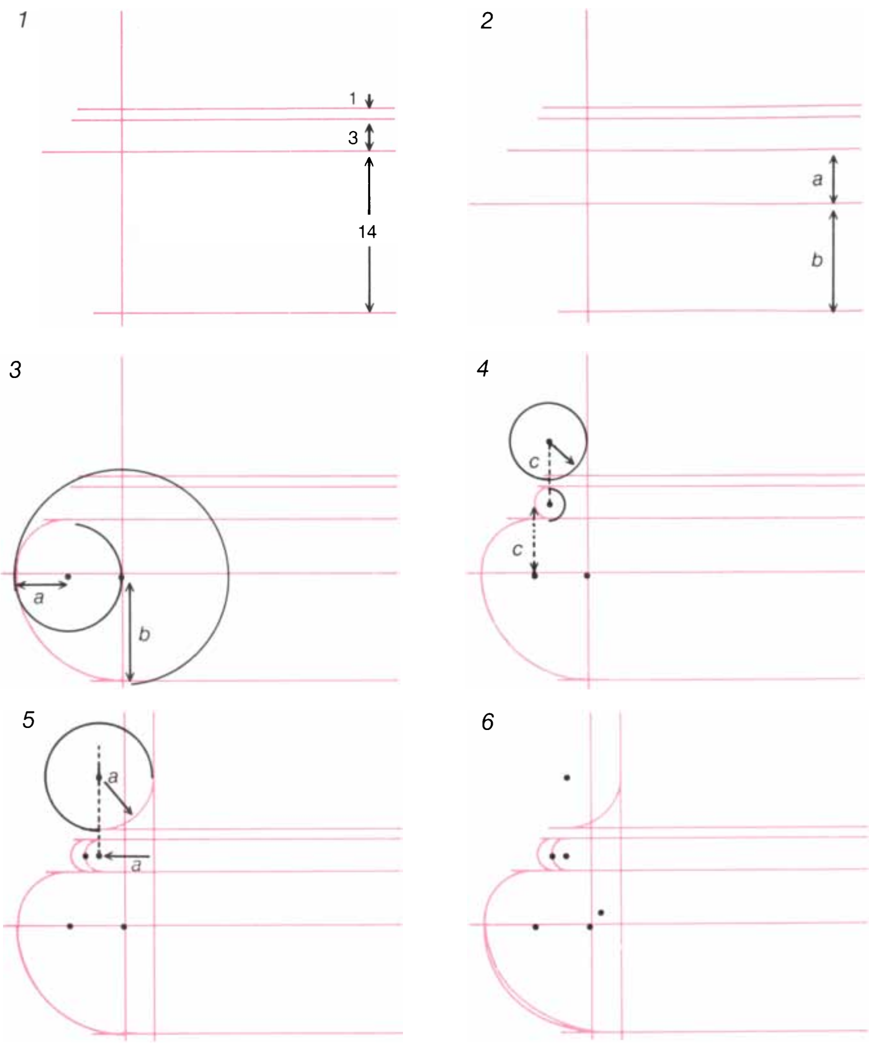
We know from one of the surviving "annual reports" of this great enterprise (the temple authorities publicly exhibited stone slabs on which the building's progress and cost to date were summarized) that its first columns were erected in about the middle of the third century B.C. The design of the columns and of their base profile must have preceded the event by at least a short time. According to other building accounts, the podium walls of the adytum, where the finished column and base designs were drawn, were built at more or less the same time, in about 250 B.C. Hence we can reliably date these and all other concomitant drawings to this period.

It was about this time, 80 years after ground for the temple was first broken and after construction work on the stepped platform was for the most part completed, that attention could be turned to the detailed design of the columns and the walls of the superstructure. Indeed, since the podium walls of the adytum were the earliest ones erected, most of the drawings are found on them. By this time the architects responsible for the design, Paionios and Daphnis, were long dead.

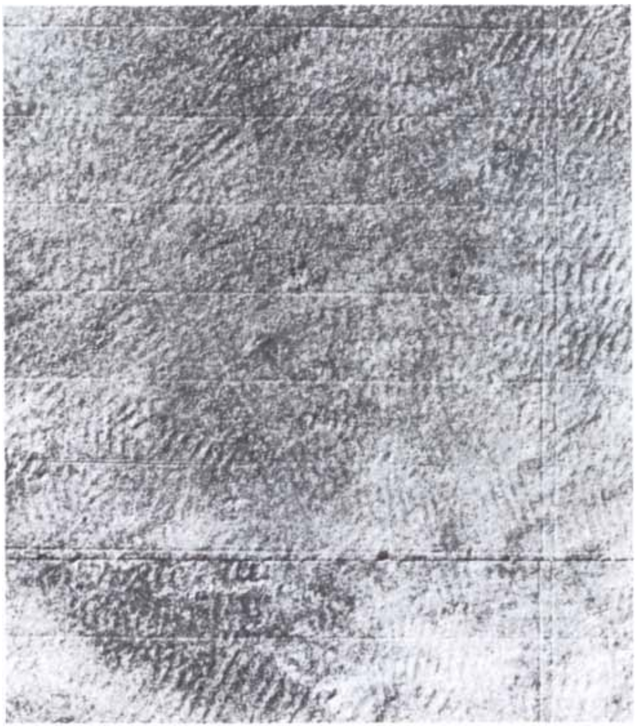
Because columns were the dominant feature of Greek temples, no expense was spared in their production. For the Temple of Apollo no fewer than 120 columns were required. From the bookkeeping records it is known that each column cost roughly 40,000 drachmas, or, at today's wages, nearly a million dollars. Much effort also went into column design.

To make a detailed plan possible, the shaft of these columns, almost 18 meters long, was drawn full scale on the podium walls of the adytum. Of course, because these walls are not 18 meters high, it was necessary to draw the column on its side.

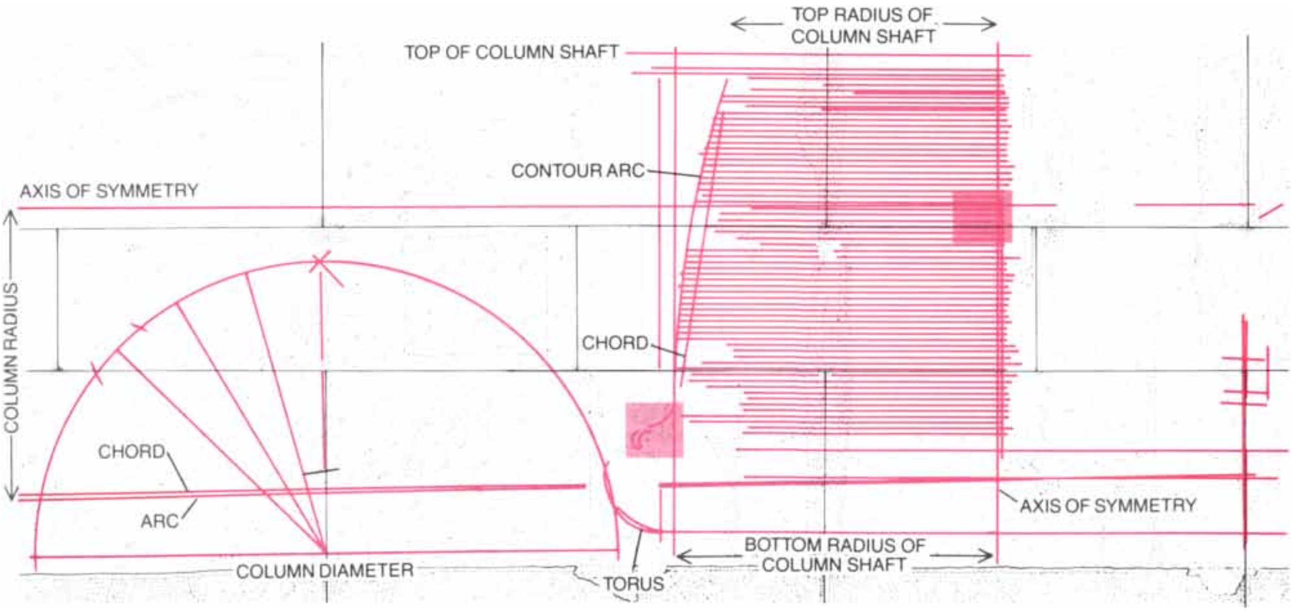
Three lines were enough for representing the bare shape of the column shaft. A pair of closely spaced lines, one straight and the other slightly curved, diverge from each other and, after reaching a maximum separation of 4.65 centimeters midway along their length, run together again. These represent the construction lines for the almost imperceptible curvature of the shaft. The third line lies some distance above the other two and runs almost parallel to them. (The distance between this line and the paired lines differs by one-sixth from one end of the column to the other.) The third line represents the column's central axis; it also is the column's axis of symmetry. Although these lines constitute only half of a vertical cross section, they are



DESIGN PROCESS for the base of the column was a combination of rigid geometric construction and subjective refinement. Four horizontal lines were drawn through one vertical line (*1*) so that the respective distances between each pair of lines were (*top to bottom*) one dactyl (1.85 centimeters), three dactyls and 14 dactyls. Another horizontal line was drawn (*2*) that split the distance between the bottom pair of lines in such a way that $a:b$ was equal to 1 : 2. Around the intersection point of this line and the vertical a circle of radius b was drawn (*3*). A smaller circle of radius a was then inscribed in the larger circle so that its center and the common point of tangency were also on the horizontal line. Starting from this point of tangency and drawing upward, a quarter of the smaller circle was outlined, and starting from the point of tangency and drawing downward, a quarter of the larger circle was outlined. The outlined portion was meant to depict the cross section of a geometrically "perfect" column base, called the torus. A vertical line drawn from the center of the smaller circle (*4*) was meant to fix the position of the semicircular molding drawn above the torus, between the middle pair of the original horizontal lines. The center of the molding, in turn, was used as an endpoint for another vertical line equal to c. A quarter of a circle centered on the other endpoint was meant to define the upward curve connecting the column base with the column shaft. Before this design was continued, however, it was arbitrarily changed (*5*). The semicircular molding was set back a short distance that is apparently unrelated to any of the geometric proportions used up to this point. Nevertheless, the designer did revert to the original proportions when he set the vertical line back a distance a from the molding's new center. In this way the new curve connecting the column base with the shaft could be defined by a circle of radius a. The final sketch, as it stands on the wall (*6*), also includes several freehand curves that tuck in the bottom two-thirds of the torus.



INSCRIBED LINES depicting temple structures can be seen in this pair of photographs. The curve (*left*) that sweeps downward to a pair of semicircles represents the foot of the column shaft as it joins the column base above a semicircular molding. The right-hand semicircle corresponds to the actual position of the molding on the base. Above and to the right of these incised curves is a series of parallel lines (*right*). They are measuring-unit divisions, each meant to represent a Greek foot in actual length along a scaled sketch of a column. Traces of the serrated chisel with which the marble blocks were smoothed are clearly visible. The joint between two blocks can be seen in the lower half of the photograph. A violent earthquake in the Middle Ages, which demolished most of what was standing of the temple, displaced the blocks somewhat, as can be seen from the slight misalignment of the inscribed vertical lines.



THREE OVERLAPPING DRAWINGS can be made out from the incised lines on a temple wall (*colored lines*). These drawings specify the shape of the column shaft and base: two vertical cross sections of the column (one of which is drawn on its side) and a horizontal section through the shaft. The upright vertical section is scaled in the vertical dimension to one-sixteenth the actual size; all the other drawings are full scale. The joints between the marble blocks of the wall (*black lines*) are also shown. Each drawing was meant to depict half of the structure it represents; the other half is simply its mirror image. Only the top fourth of the recumbent full-scale section of the column shaft is shown. The arc below the chord outlines the nearly imperceptible curvature of the shaft's contour. The diagrammatic horizontal cross section of the shaft includes three divisions that fix the spacing for the 24 flutes around the shaft's circumference. Each closely spaced parallel line in the third drawing represents one Greek foot in length along the column shaft. In this compressed version the convex contour of the column can be clearly seen. The varying radial dimensions of the shaft could be measured off this drawing directly. The colored patches indicate the area covered by the photographs at the top of the page.

sufficient because the other half is simply a symmetrical reflection. The draftsmen who carved these drawings rendered only one of two identical halves in almost all cases.

The diagrammatic vertical cross section of the column is complemented by a horizontal cross section that was also sketched in simplified form along its axes of symmetry. A large semicircle of the proper dimensions with three equally spaced radial divisions was enough to specify the distribution of the column's 24 flutes, or vertical grooves.

The most revealing of the column plans is the one we first considered: the profile of the column base. Above the profile of the base is the hatching that first caught my eye: the closely spaced parallel horizontal lines. These lines are drawn between two vertical lines that respectively represent the outer surface of the column shaft and its inner axis.

The curvature of the shaft in this drawing is depicted by a shallow convex arc, the endpoints of which are connected by a diagonal chord (partially effaced by erosion). The gap between the arc and the corresponding chord measures 4.65 centimeters at its widest point. This distance agrees exactly with the widest separation, halfway up the shaft, between the lower two lines in the big vertical section. The radii of the top and bottom of the shaft (standing in a relation of precisely 5:6 to each other) also agree in both drawings. Only the height of the shaft varies greatly between these two representations of the column: in the latter it has been shrunk from 18 meters to a little more than one meter, or one-sixteenth its actual size. This drawing, in contrast to all the others, was scaled down in size—but only in length, not in width. There was a reason for this.

One of the structural refinements in Greek temples is the gentle curvature that is apparent not only in the imperceptible curvature of the column shaft, as mentioned above, but also in the slight arch to which all the temple's horizontal layers conform. Over a distance of 90 meters the entire building—from foundation to roof—"bulges" upward. At the midpoint along this length the base of the temple is roughly 11 centimeters above the level of each end.

The Greeks called this subtle touch, appropriately enough, entasis, or tension. It constitutes the most vigorous expression of the "activity" produced by the individual parts of the structure, both the load-bearing parts and the sections that are themselves supported. It was because of this refinement that the parts could be forged into an organic whole instead of standing in isolation from one another. How the delicately curved contour of the column shafts was achieved had until now remained a mystery.

As seen in the drawings for the Temple of Apollo, the arc traced by the column's contour was intended to bulge outward a maximum distance of 4.65 centimeters from the straight chord connecting the top and bottom of the 18-meter shaft. Such a flat arc, in smaller dimensions, could be easily and precisely drawn by the swing of a pair of dividers. Working in the actual dimensions, however, the curvature of the column would have been impossible to describe: the dividers would have had to be set for a radius of nearly nine-tenths of a kilometer.

A geometric construction as simple as it was ingenious was therefore used: the height of the shaft was reduced. To do so a Greek foot in real measurement (29.6 centimeters) was equated with a dactyl, one-sixteenth of a foot (1.85 centimeters). The curve of the column's entasis could now be described with dividers that had a radius of approximately 3.2 meters. By leaving the width measurements unchanged it was possible to obtain full-scale measurements of the continuously varying cross-sectional radius at any point along the column. The corresponding height along the column for each radius measurement could be determined from the scale drawing by counting the hatched lines: the separation between each line is one dactyl and therefore each line represents a foot along the actual column's length.

These rough sketches would have been sufficient to guide the stonemasons as they began to carve the columns. Nevertheless, the plans are supplemented by a second set of more detailed plans across the adytum from the preliminary sketches: an 18-meter-long section through the length of the column shaft, and two horizontal cross sections—one through the top and one through the bottom of the shaft. In these drawings everything that had been sketched roughly in the previous plans is drawn in full detail. For example, the horizontal cross sections have been finished down to the details of all the flutes. Only then was the architect satisfied; the plans for the columns were as perfect as the columns were meant to be.

Construction plans for other parts of the temple have also been found. One extensive collection of drawings seems to have to do with the construction of the temple walls, whose exterior surfaces do not stand rigidly vertical but incline very slightly inward. Another drawing turned out to represent part of the entablature supported by the columns.

A 12-meter-wide drawing showing the entablature and pediment (the low frontal gable) of the *naiskos* was discovered on the transverse wall at the back of the adytum. It came as no surprise to find that the drawing of the entablature's construction corresponded exactly to what had actually been built. What was surprising, however, was the marked discrepancy between the dimensions of the pediment blueprint and its actual width, a difference that affected the whole structure. The *naiskos* was built approximately 2.8 meters narrower than the plans indicate, and its relative proportions were thereby drastically altered. A liberty of this magnitude is hard to explain, even in the light of all we now know about Greek architecture.

The 200 square meters of drawings discovered represent a large part, but by no means all, of the plans that must have been made. For example, no drawing has been found showing the triangular pediment of the larger temple. This is not entirely unexpected, because the construction of the main pediment was never started; quite likely it had not been planned and drawn in all its details either. On the other hand, the plans for the columns of the *naiskos,* which had been erected, are also missing—at least they are not to be found on the walls of the temple.

The most crucial set of plans not found anywhere on the walls were those for the foundation of the entire building. Clearly they must have preceded the plans for the temple's superstructure. This omission could not be dismissed easily. Where were the surfaces on which they would have been drawn?

If the Greeks drew plans of temple walls on the wall, could they have drawn floor plans on the floor? With this question in mind I examined closely the stone layers making up the foundation of the temple. The search revealed an entire array of thin lines engraved on the surface of the individual layers: alignments, axes and other construction lines. The details of the floor plan had been drafted on the very layers composing the stepped platform. Each layer was laid in accordance with the full-size plans inscribed on the preceding one. Furthermore, broader ideas for the building of the temple were tentatively traced out on a layer and, if they were accepted, the markings were copied from layer to layer. On reaching the topmost layer of the stepped base, the architect had come to the final ground plan on which the most important parts of the superstructure (the walls and columns) had to be accurately laid out in their final, binding form.

The placement of these parts was determined by a precise rectangular grid. The grid markers are found in the only suitable place: along the base of the wall surrounding the nucleus of the temple. The markers consist of a series of short vertical lines that, although they are inconspicuous, have been incised with a precision accurate to within a millimeter.

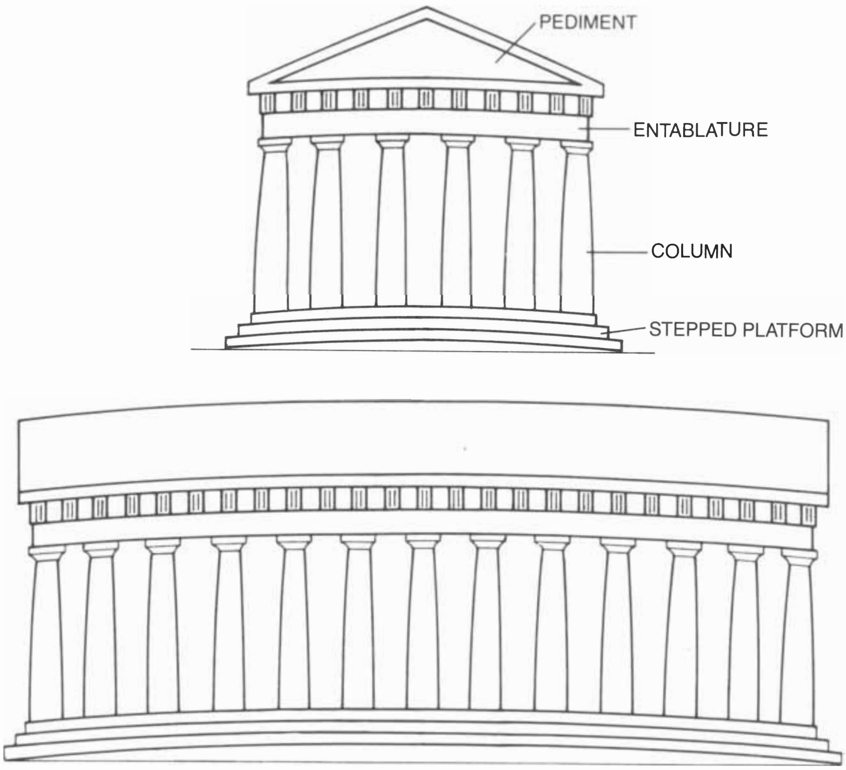
With the discovery of these plans the link between the ground plan and the vertical structures drawn on the temple walls had been established. We now have an unbroken sequence of plans before us.

It need hardly be said that the architects of the temple did not wait until the first stone layers of the foundation were laid before they started envisioning the final ground plan; nor did they delay setting the overall dimensions for the superstructure until after the walls had been erected. The drawings on the walls and other surfaces represent only the part of the planning process immediately preceding its execution: the site planning.

Before reaching this final stage, the plans for the structures must have gone through several drafts. It seems likely that papyrus, parchment, whitewashed wood tablets or even flat stones served as the material on which these preliminary designs were drawn. The last option was confirmed recently by my colleague Wolf Koenigs, who, while investigating the Temple of Athena in the town of Priene, discovered a scaled-down sketch of its pediment; it had been incised in the bottom of a block that was later fitted into the building itself.

Koenigs also found traces of large-scale drawings on one of the wall surfaces of the temple. I have found vestiges of similarly incised drawings on the Temple of Artemis at Sardis, and Wolfram Hoepfner has uncovered evidence of plans for a burial chamber that were drawn in red chalk.

The ancient blueprints at Didyma may no longer be the only ones known to exist in Greek architecture, but they certainly are the most complete. It is ironic that this architectural bonanza was preserved because the temple never came close to completion.



SLIGHT CURVATURES are found in all long vertical and horizontal members of Greek temples. Here the curvatures have been exaggerated in a front view (*top*) and side view (*bottom*) of such a temple. The tapered columns are somewhat "swollen," and the entire temple arches. The actual deviations from the straight vertical or horizontal are minuscule: the columns of the Temple of Apollo "bulge" outward a maximum distance of 4.65 centimeters over a length of nearly 20 meters; the foundation "hunches" a maximum distance of 11 centimeters over 90 meters. Yet these refinements clearly invigorate the temple's design.

Whereas the markings on the horizontal stone surfaces were for the most part built over and so disappeared automatically, the drawings on the walls remained exposed. Yet toward the end of the construction work they too were meant to be erased. The wall surfaces of the Temple of Apollo do not have the end finish intended for them; they still bear a thin protective layer of stone that was to have been ground off only in the very last stage of construction. The final polishing can be seen only on the inside corners of the wall (where it could not have been accomplished later with the desired perfection). The depth of the protective layer is approximately half a millimeter, about the average depth of the incised lines. If the walls had received their final polishing, the intricate building plans that had guided the construction of the temple would have been obliterated.

The reason for the apparently total lack of construction plans in Greek architecture lies above all in a characteristic penchant for uniqueness. At each stage of construction the plans for a building were constantly elaborated and refined. Once the building was completed, however, there was no longer any need for the plans. The refining process led to singular creations of timeless beauty, unsurpassable in their self-defining, Classical style.

The greatest irony of all is that in order to achieve such perfection the Greeks used techniques that were not significantly different from those employed earlier by Egyptian architects, and that were later used by the Romans, and later still by builders in the Middle Ages.

The technical and practical side of designing and drawing to a large scale was little different in Greek architecture from what was evidently common practice 2,000 years earlier in Egypt and 1,500 years later in central Europe. Construction plans have been found incised into the walls and floor surfaces of many Gothic churches (for example the cathedrals of Chartres, Reims, Bourges, York and Orvieto), and painted or incised "measuring and guiding lines" have been observed on early Egyptian monuments that, in light of the temple drawings at Didyma, can now be recognized as construction plans.

My discovery at the Temple of Apollo does not disclose any "lost" knowledge. It merely demonstrates that although the Greeks employed ordinary methods in translating their architectural ideas into tangible monuments, they achieved extraordinary and enduring results.

THE AMATEUR SCIENTIST

The kaleidoscope now comes equipped with flashing diodes and focusing lenses

by Jearl Walker

A kaleidoscope seems by magic to transform bits of ordinary colored glass and plastic into a geometrically precise array of colors and shapes. Until recently the kaleidoscope was mainly a toy, but new designs and novel applications of optical principles have brought the device into the world of art. Some modern kaleidoscopes are valued in the thousands of dollars.

The kaleidoscope was invented in 1816 by David Brewster, a British physicist who is remembered for his studies of the polarization of light. The traditional kaleidoscope consists of a tube containing two or three reflecting surfaces such as mirrors or shiny strips of metal that extend along the full length of the interior. The reflectors form either a V or a triangle. You look into one end of the tube, usually through a small aperture, so that your line of sight passes between the reflectors to the opposite end.

The materials at the far end of the tube come in a number of varieties. Bits of brightly colored plastic or glass are common, as are paper clips, pins and other pieces of metal. Part of what you see in the kaleidoscope is a direct view of the material at the far end. In a two-mirror kaleidoscope the direct view is shaped like a slice of pie. In a three-mirror version it is triangular. In both types you also see reflected images, usually of the direct view. I have long puzzled over these images. Why do some kaleidoscopes produce only a few of them whereas others give hundreds? I supposed the difference must be in the quality of the mirrors, but I was wrong.

Many present-day kaleidoscopes incorporate variations on the traditional designs. In some of them the objects at the far end of the tube are suspended in a viscous, clear oil so that the objects move sluggishly when the tube is rotated. Wheels at the far end are also popular. A wheel is mounted on an axis extending from the tube. Sometimes there are two wheels. Wheels can consist of clear plastic containing either butterfly wings or thinly sliced sections of colored minerals. Alternatively a wheel can be made of translucent colored glass or plastic, forming a composite that resembles a stained-glass window. The idea is to turn the wheels as you look into the kaleidoscope. Some kaleidoscopes contain music boxes. When you wind the instrument, the colored array at the far end rotates for several minutes to the accompaniment of a tune.

Another variation relies on the polarization of light. The far end is covered with a polarizing filter. Just inside the end are strips of clear, stretched plastic wrap. Another polarizing filter lies over the viewing aperture. The colorless plastic is made to appear brilliantly colored by the arrangement of filters. The colors depend on the thickness of the strips and the direction in which they are stretched.

In one version of this kaleidoscope the reflections are generated by a sheet of shiny metal that has been rolled into a loose cylinder, pulled into a spiral at one end and then inserted into the tube. The reflected images are not sharp reproductions of the direct view but smeared, extended stripes of color. I have a kaleidoscope in which the reflecting surface is the highly polished metal interior of the tube, which gives rise to concentric rings around the direct view.

The reflectors in one of my kaleidoscopes are made of light green plastic. They produce reflected images as mirrors do, but the images are swamped by the light leaking through the transparent sides of the kaleidoscope. I cover the instrument with a cloth to improve the visibility.

Many of today's kaleidoscopes are open at the far end, so that what you see is a variously transformed image of what you point the instrument at. Some of them have a converging lens glued to the far end, with the result that more of the scene is squeezed into the view. One of my kaleidoscopes has a hemispherical lens at the far end; its flat side faces the interior. Between the lens and the mirrors is a small spherical plastic ball. Most of the direct view consists of the scene in front of the kaleidoscope, condensed by the hemispherical lens, but part of it consists of the same scene condensed even more by the spherical ball.

I have saved my two favorite kaleidoscopes for last. One is from Kaleidoscopes by Peach of Austin, Tex. One of the partners there, Peach Reynolds, has been a pioneer in the development of modern kaleidoscopes. One version of his instrument is the first "high tech" kaleidoscope: the traditional bits of colored glass and plastic are replaced with an array of light-emitting diodes controlled by a sound-actuated transducer. When the transducer intercepts sound waves, it briefly lights up the colored diodes. To sample different parts of the array the viewer rotates the far end of the tube. I enjoy sitting in front of my stereo system while the diodes pulse to the beat of the music.

My other prized kaleidoscope is produced by Timothy Grannis and Jack Lazarowski of Prism Design in Essex Junction, Vt. A lens at the far end compresses the external scene. The mirrors form equilateral triangles at both ends. The novel feature is that the mirrors flare toward the viewer, so that the triangle at the far end is smaller than the one at the near end.

The resulting array of reflected images is stunning. They are not spread out on a plane passing through the direct view as in other kaleidoscopes. Instead they form a geodesic sphere composed of triangular sections. The illusion that the array is curved in three dimensions is startling. The illusion is enhanced by thin, bright threads that seem to project radially outward from the sphere into a bright, featureless background. When I point the kaleidoscope at a street scene or an active television screen, the sphere bursts into frenzied motion, each section on the sphere showing something different.

How are the beautiful displays created in traditional and modern kaleidoscopes? Although I knew reflections were involved, I did not understand how the images fitted so neatly into a geometric order. I was just as puzzled after examining several of my kaleidoscopes. They were too small to be enlightening. Determined to make a much larger kaleidoscope, I bought (in a department store) several flat mirrors 1.3 meters long and .3 meter wide.

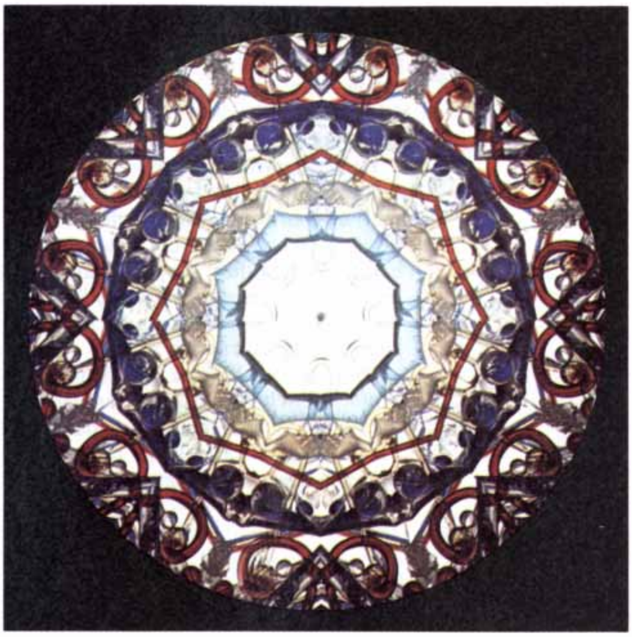
Ordinarily such inexpensive mirrors yield poor images. Most of the reflection is from a shiny metallic coating on the back of the glass. The front surface gives less reflection. The two reflections result in blurred edges in the image of an object. Nevertheless, my mirrors (made by Hamilton Glass Products, Inc., 200 Chestnut Street, Vincennes, Ind. 47591) gave remarkably clear images.

I stood two of the mirrors upright and taped them together so that they formed a V. The reflecting surfaces were on the inside. The tape hinge enabled me to vary easily the angle between the mirrors.

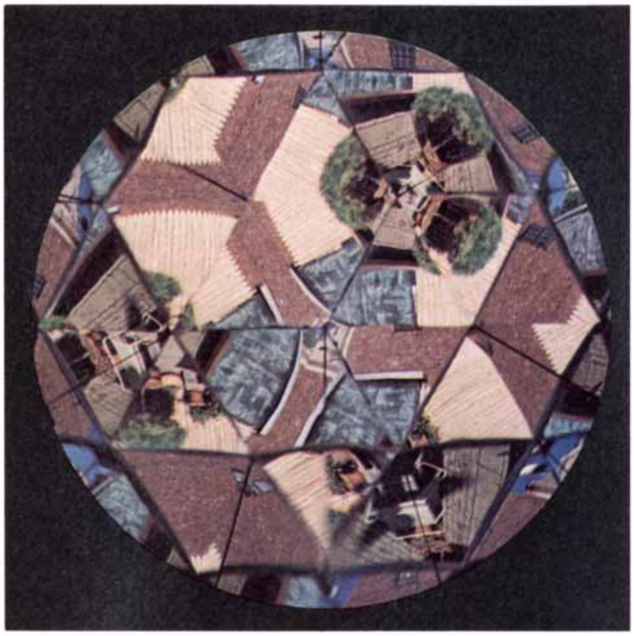
I put a small box on the floor between the mirrors and then counted the images of the box. The number rose as I decreased the angle. Five images were visible when the angle was 60 degrees. In addition I saw images of the mirror edges, two to each side.

Next I looked down at the box from above the mirrors. Again I saw five images of it as well as the direct view. Now I began to understand how a two-mirror kaleidoscope functions. When the angle between the reflecting surfaces is 60 degrees, the kaleidoscope yields a cluster of six views of the object at the far end. The views lie in sectors of pie-slice shape spread around a hub positioned at the intersection of the mirrors.

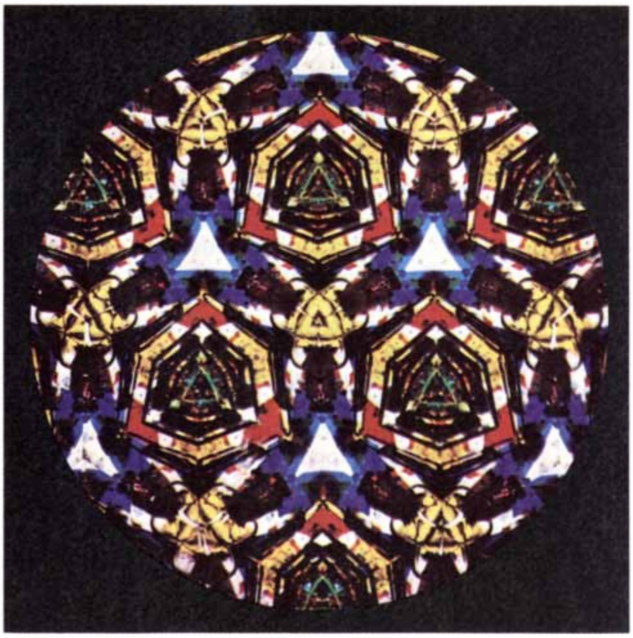
One of the views is direct and the



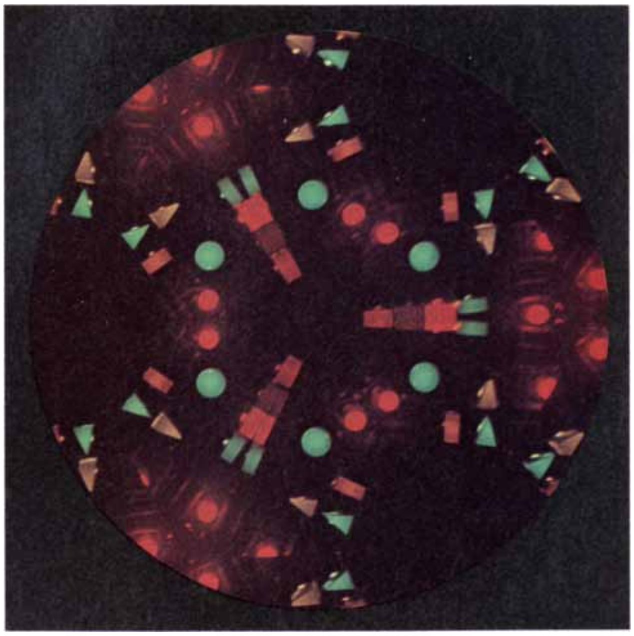
A 16-fold kaleidoscope image



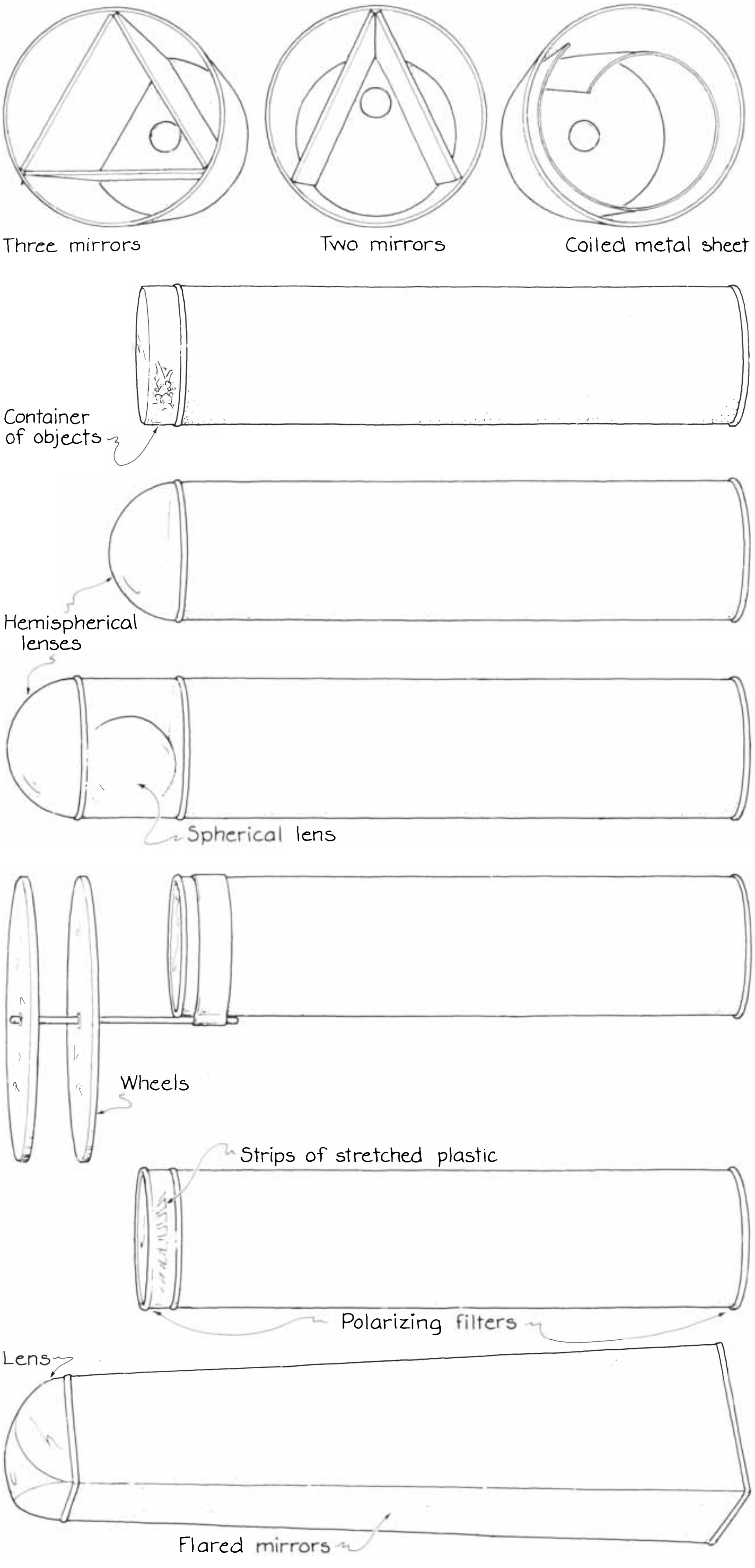
The effect of a lens



Kaleidoscope objects in clear oil



An image made by diodes responding to sound



Kaleidoscope designs

others are reflections. All seem to lie in the plane through the direct view, in this case the floor. (I call it the image plane.) In toy kaleidoscopes the reflecting surfaces are poor and the images are dull and blurry compared with the direct view. My mirrors yielded reflected images almost as bright and sharp as the direct view.

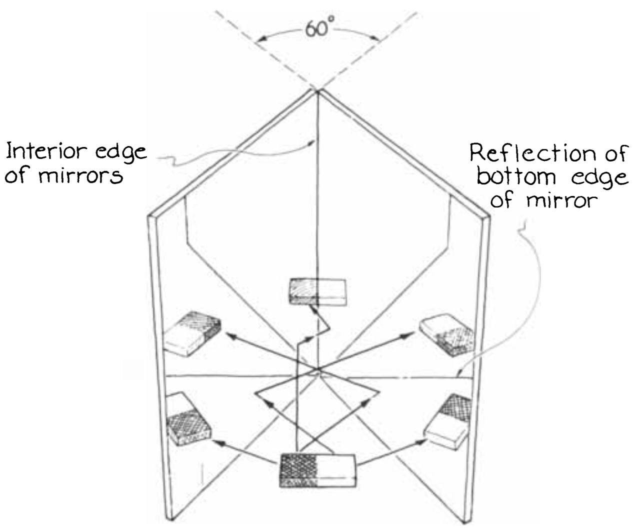
When I decreased the angle between the mirrors, the sectors shrank as more images appeared on the side of the hub opposite to the direct view. Usually toy kaleidoscopes are not constructed to yield more than five reflected images because the additional ones are too dull and blurry. With my mirrors I could see sharp images even when the angle was only 22.5 degrees, which produced a direct view and 15 reflected images.

How many images appear at a given angle? You might explore this question with two mirrors. (Some physics textbooks give an incorrect answer. Contrary to what they say, the number of images is not necessarily one less than the ratio of 360 degrees to the angle between the mirrors.)

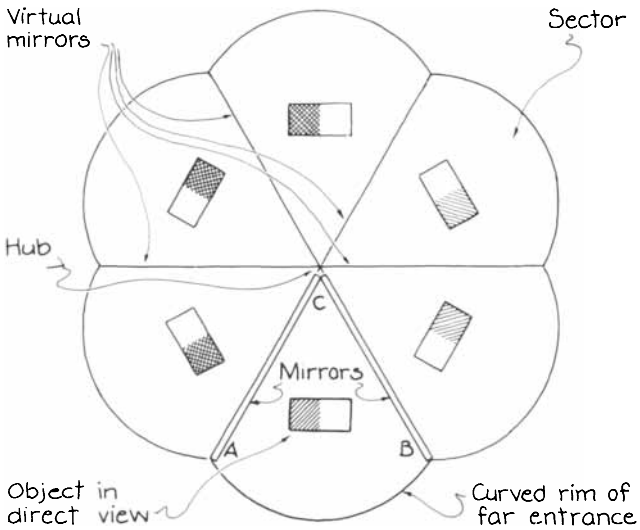
One easy technique is to build up a sketch of what you see, starting with the direct view. The apex is the hub of the full set of clusters of images. Reflect the direct view around one of its mirrors to form a new adjacent sector. This reflection amounts to rotating the direct view around the mirror until it lies again in the plane of the image field. Make another reflection on the other side of the direct view. Continue to reflect each new sector around its far side as you move both clockwise and counterclockwise around the hub until sectors meet or overlap. The radial boundaries of the sectors, which are truly the reflected images of the bottom edges of the mirrors, are called virtual mirrors because they effectively reflect one sector into the next.

If the sectors finally meet but do not overlap, the number of images is easy to calculate: one image per sector. If the sectors overlap, the calculation is more difficult. You should consult the study by An-Ti Chai of Michigan Technological University cited in "Bibliography" [*page 150*].

I continued my work by taping a third mirror to the first two to form an equilateral triangle. I thought the arrangement would yield two additional sixfold clusters because two more points (*A* and *B*) could then serve as hubs. When I peered down into the system, I found the entire image plane covered with sixfold clusters. I was so taken aback by the spectacle that I squeezed my head into the system as far as it would go. At least 20 distin-



A two-mirror array



The two-mirror cluster

guishable clusters stretched in every direction on the image plane, growing ever smaller with distance from the direct view [*see top illustration on next page*]. Farther out I saw dozens more that were too small to be clear.

The image field consisted of triangles outlined by reflections of the bottom edges of the mirrors. Within the triangles were images of the box. Superposed on the field were lines that were images of the vertical, interior intersections of the mirrors. Throughout the image field the clusters reproduced the clusters around points *A, B* or *C* in the direct view. The entire field could be made visible by beginning with the direct view and then repetitively reflecting it around each of its boundaries until the plane was filled.

As yet I did not understand how the light rays were reflected from the three mirrors to fill the image plane. I realized, however, that the optics must be generally related to the formation of images in a similar system of two parallel mirrors [*see bottom illustration on next page*].

Consider the rays leaving the small box at the far end of the system. The ones that go directly to the observer create the direct view. Others leave the box at various angles to the central axis of the system. Consider the ray that reflects at the midpoint on one of the mirrors. When it is intercepted, the observer perceives it to have originated from a point that is on a rearward extrapolation of the ray entering the eye. The point lies at the intersection of the extrapolation and the image plane. The image is perceived to be there.

Another ray leaves the box and is reflected first from the bottom mirror and then from the top mirror, each at a point a fourth of a mirror length from an end. Again the observer believes the ray originated on the image plane on a rearward extrapolation of the ray that enters the eye. This time the position is farther from the direct view than in the case of a ray reflected only once.

Another ray in the illustration is reflected three times from the mirrors, first at the farther one-sixth point on the top mirror, then at the midpoint on the bottom mirror and finally at the nearer one-sixth point on the top mirror. This ray creates an image that is even farther from the true position of the box. Thus images are created upward along the image plane. The greater the distance is between an image and the direct view, the more reflections are required. Ideally there is no limit to the number of images. Eventually, however, the many reflections make the images blurry. An identical set of images extends downward on the image plane. In kaleidoscopes that have two pairs of parallel mirrors (only a few are constructed this way) images also extend to the left and right along the image plane.

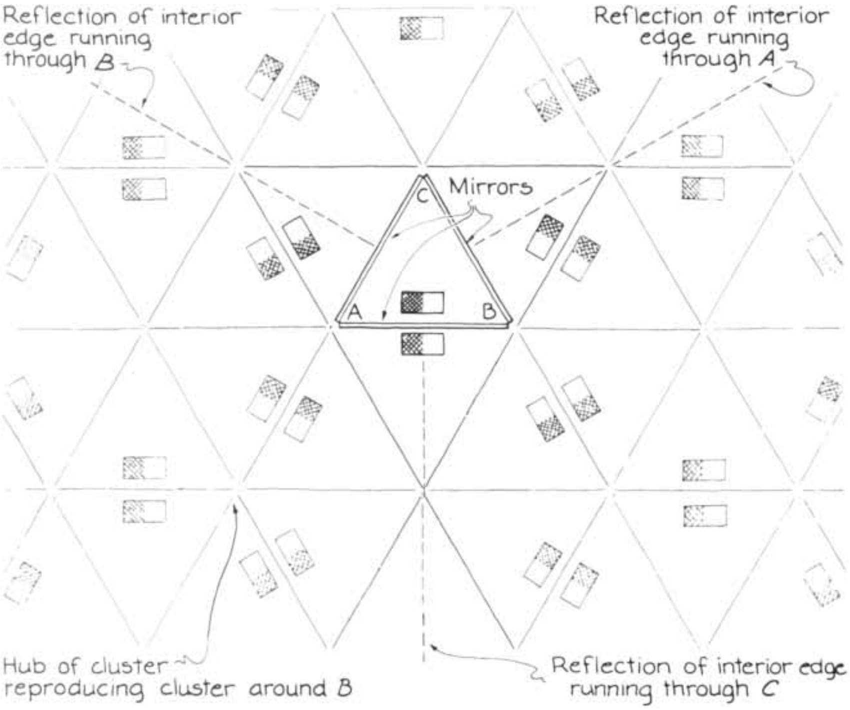
Surely a three-mirror system must function in a similar way, but exactly how? I was determined to follow the light rays that entered the system and were reflected from the mirrors to my eyes. I briefly considered placing a laser at the far end and studying how its beam was reflected as I changed its angle of entrance, but I abandoned the notion because I did not want the beam to end up in my eye. Instead, with the room lights off I began to probe the system with a small laser pointed downward.

I knew that the path of a light ray from the box to my eye (even with intermediate reflections) was the same as the path of a ray from my eye to the box. Therefore I pressed the side of the lightweight laser to one temple, closed the opposite eye and moved my head and the laser so that the beam was pointed to an image of the box. The beam continued downward through the mirror system, illuminating the box and thereby adding a red spot to all the images.

The arrangement was not perfect because the beam left the side of my face somewhat displaced from the path of the light rays traveling from a particular image to my eyes. Still, the small angle of error was usually unimportant except when the beam was reflected 10 or more times. I blew smoke into the system to make the path taken by the beam visible.

I first examined a cluster adjacent to the direct view, systematically pointing the laser at each image of the box in the cluster and then noting how the beam crisscrossed between the mirrors to reach the box. After sketching the paths I reversed the arrows that indicate the direction of travel of the light [*see upper illustration on page 144*]. I then had a sketch of how light rays normally leave the box and are reflected from the mirrors to create the sectors in the cluster. Sector *J,* which is adjacent to the direct view, is created by light rays that are reflected from the far end of the left mirror. Sector *K* is due to rays that are reflected from the far end of the bottom mirror and then from the left mirror. Sector *O* is similar except that the first reflection is from the far end of the right mirror.

Sectors *L* and *N* require three reflections. For sector *L* the rays are reflected from the right mirror, then from the



Images formed by mirrors arrayed as an equilateral triangle

bottom mirror and finally from the left mirror. For sector *N* they are reflected from the bottom mirror, the right mirror and the left mirror. Sector *M* is more complicated because the rays are reflected four times: bottom mirror, right mirror, bottom mirror again and left mirror.

Once I had made this illustration I began to appreciate why a three-mirror kaleidoscope produces a richer display than a two-mirror version. If the bottom mirror is removed, only sectors *J* and *O* remain; they are independent of reflections from that mirror.

I explored more of the image field with the laser beam, counting the number of reflections associated with each sector [*see lower illustration on page 144*]. As I had expected from my study of a parallel-mirror system, the sectors far from the direct view require more reflections. I also noticed a curious sequence in the reflections. Pick any hub and consider the reflections required by the sectors surrounding it. As you would expect, the sector closest to the direct view requires the lowest number of reflections (call it n) for that particular cluster.

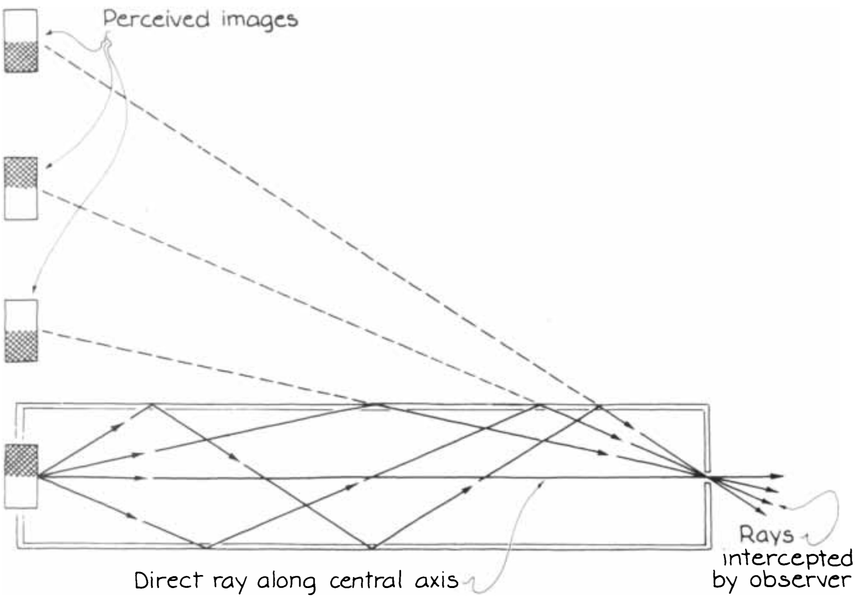
Now move around the hub by one sector both clockwise and counterclockwise. Because they are farther from the direct view, these sectors require one more reflection: $n + 1$. Move one more sector around the hub in opposite directions. These sectors require another reflection: $n + 2$. Finally move to the sector farthest from the direct view. It requires yet another reflection: $n + 3$. Every cluster turns out to have such a mathematical sequence around its hub.

When the direct view consists solely of objects lying in the plane across the far ends of the mirrors, every sector in the image field is identical. If the objects are more distant, the content of the sectors must differ. I had noticed this variation when I moved my mirror system outdoors, set it horizontally on a table and pointed it at a busy street. All the sectors differed in content.

This variation results from the perspective of the external scene available at the far end of the mirror system. A simple example is shown in the upper illustration at the right on page 145. Two objects, represented by a circle and a rectangle, are at the end of the system farthest from the observer. The circle cannot be seen in the direct view because it is too far off the central axis of the system. It can, however, send rays into the system at a sharp angle to the central axis. They are reflected three times before they reach the observer. He perceives the circle as lying in a sector of the image plane distant from the direct view. Similarly, other objects far from the central axis send their rays into the system at other angles and are perceived as being in other sectors.

The slanted mirrors in the kaleidoscope made by Grannis and Lazarowski distort the image field into what appears to be a three-dimensional geodesic sphere. To examine this distortion I stood in front of one of my mirrors while I tilted its top edge away from me. The image of the floor tilted away as if the bottom edge of the mirror were on the edge of a precipice.

Similarly, each sector in the kaleidoscope with flared mirrors seems to be tilted away from me because the shape of the sectors appears as an equilateral triangle viewed at a slant. As I move my view to one side of the direct view the sectors seem to slant more, creating the illusion that they are on the side of the geodesic sphere. The rays extending radially outward from the sphere are actually the images of the



Multiple reflections from parallel mirrors


Dance has a new partner.

A new look at

Summary:
GTE lighting research operates on many fronts: a space lighting lab to study the motion of gases in a gravity-free environment; the use of various isotopes to enhance the output of fluorescent lamps; the production of light directly from excited molecules.

The science of lights and lighting might seem to be rather mature. Indeed, the standard light bulb has changed very little in at least half a century.

But lighting science is on the brink of revolution. Recent work by GTE points the way to major improvements in every type of lighting.

Lighting research in space.

One of the most powerful and efficient light sources is the high-intensity-discharge (HID) lamp. Its light is derived from gases and ionized vapors which are excited in an electrical arc contained in a quartz arc tube.

The gases circulate by gravity-induced convection, which mixes the radiating species in the arc. This tends to obscure other vital processes such as diffusion, cataphoresis (motion of ions toward the negative electrode), magnetostriction and vapor condensation. Researchers have wanted to observe these processes at leisure, in the absence of convection, for many years.

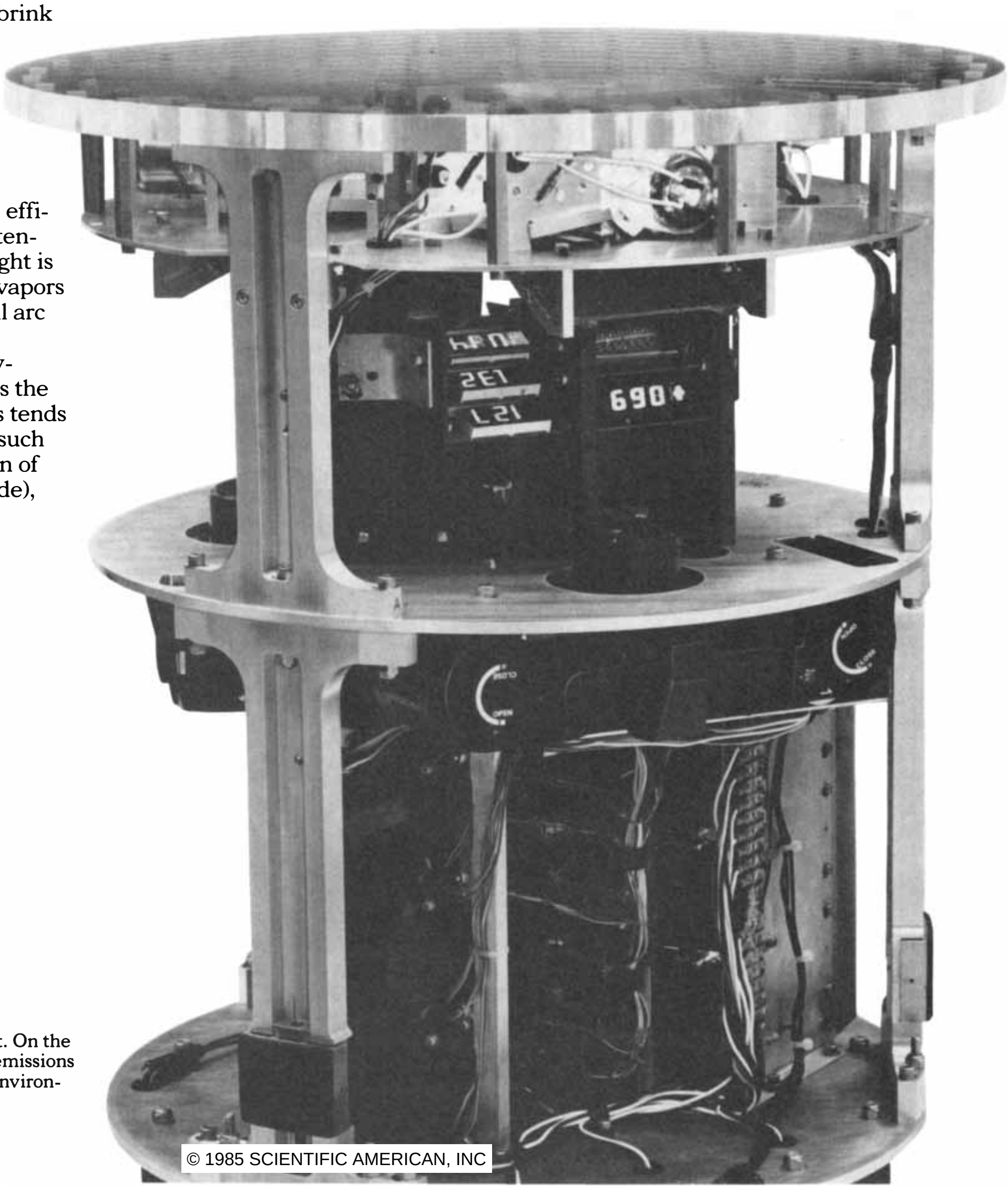
GTE has achieved this goal in a first-of-a-kind experiment aboard the space shuttle. A payload of three metal-halide HID lamps was operated in the microgravity environment of the orbiter. Each lamp was lit for half-hour periods while detailed spectroscopic, light output and electrical measurements were taken.

The results have substantially strengthened the technological underpinning drawn upon for lamp design. GTE scientists now have critical information and new insights that will produce lamps with brighter, whiter light.

Untrapping excited atoms.

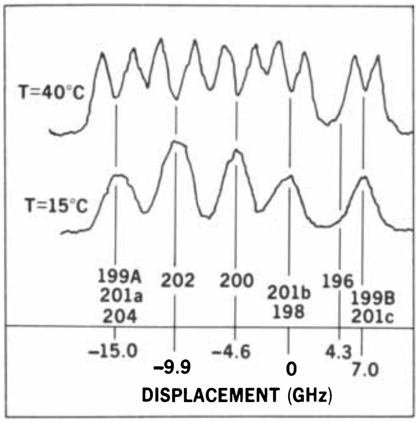
Improvements in fluorescent lamps are on the way, too. As just one example, GTE has discovered how to increase the efficiency of these lamps by about 5%.

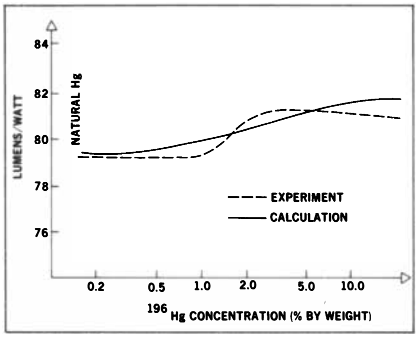
Mercury vapor in the lamp emits



Payload for the GTE space experiment. On the top layer are three HID lamps, whose emissions were investigated in the microgravity environment of the space shuttle.

an old science.





ultraviolet light when it is excited by the electric current. This light is transformed into white when it strikes the phosphors coating the glass tube but some ultraviolet is reabsorbed by the mercury vapor, limiting the lamp's efficiency. GTE researchers have found, however, that by increasing the level of 196 Hg isotope from its naturally-occurring 0.15% to 3.0%, more ultraviolet light escapes to the phosphor. Output improves about 5%.

Light from molecules?

In the future, light may be produced directly from excited molecules in low-pressure lamps. The light spectrum is in broad bands, rather than the narrow-line emission from mercury or sodium atoms.

GTE researchers are investigating ways to produce white light from molecules as the basis for a totally new lamp.

The chemical make-up of the molecules and their behavior in the excited state are undergoing critical studies. In many cases, GTE is applying electrodeless technology with RF power sources as exciters.

This new way of looking at light bulbs promises high-efficiency, long-lived, cool-running light sources with many industrial and residential applications.

The wonderful world of light.

At GTE, we are working on many projects aimed at bringing about the revolution in light. New electrode materials, improved sealants, excimers—these and more are on the GTE research agenda.

The box lists some current papers pertinent to GTE lighting research. For any or all of these, you are invited to write GTE Marketing Services Center, Department TP-L, 70 Empire Drive, West Seneca, NY 14224. Or call 1-800-828-7280 (in N.Y. State 1-800-462-1075).

Pertinent Papers

Convection and Additive Segregation in Metal-Halide Lamp Arcs: Results from a Space Shuttle Experiment
Symposium on Science and Technology of High Temperature Light Sources, Electrochemical Society Meeting, Toronto, 1985

Arc Discharge Convection Studies: A Space Shuttle Experiment
Proceedings of a symposium held at NASA Goddard Space Flight Center, Greenbelt, Maryland, August 1-2, 1984

Energy Conservation Through More Efficient Lighting
Science, Volume 226, pp. 435-436, October 26, 1984

Enhanced HgBr emission at low pressures
Applied Physics Letter 42, May 1, 1983

Bound-free emission in HgBr
Applied Physics Letter 41, November 1, 1982

GTE

No other system of keeping up can compare with SCIENTIFIC AMERICAN *Medicine.*

YOUR SYSTEM: a time-consuming, futile struggle to keep up with the information explosion.

The classic texts are convenient references—but the information they contain is obsolete before publication.

Like many physicians, you probably rely on the texts you first used in medical school. But even using the most recent editions, you find material that no longer reflects current clinical thinking—and they lack the latest information on such topics as herpes, oncogenes, AIDS, and photon imaging.

Reading stacks of journals alerts you to recent developments—but can't give you quick answers on patient management.

Struggling through the hundreds of journal pages published each month—even on only the really significant advances in the field—is arduous and memory-taxing. And it's a task that costs physicians valuable time—their most precious resource.

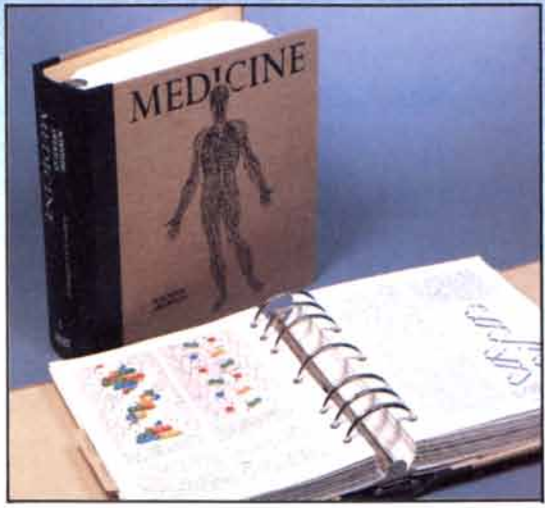
Review courses cover clinical advances—but, months later, do you recall the details of a new procedure or unfamiliar drug?

Seminars can also be costly and make you lose valuable time away from your practice—expenses that may amount to several thousand dollars. And, the speaker's skill often determines how much you learn.

SCIENTIFIC AMERICAN

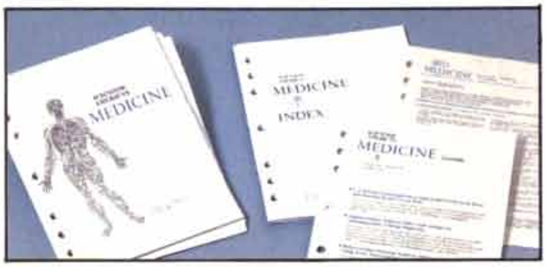
MEDI

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.



A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.



Each month, six to nine replacement chapters to update your text *plus* new references, a four- to five-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the floppy-disk version for your IBM® PC/PC Jr or Apple® IIe or II+ (or compatible models).

Your subscription includes:
the two-volume, 2,300-page loose-leaf text
and, each month, six to nine replacement chapters,
a newsletter, new references,
and a completely revised index.

call toll-free: 1-800-345-8112
(in Penn. 1-800-662-2444)

CINE

SCIENTIFIC AMERICAN MEDICINE
415 MADISON AVENUE, NEW YORK, N.Y. 10017

☐ **Yes,** I'd like to try the SCIENTIFIC AMERICAN *Medicine* system. Please enter my subscription at a first-year price of US$245*.

☐ Also enroll me in the CME program. I'd prefer:
- ☐ the printed version at **no additional charge.**
- ☐ the floppy-disk version at $US87* additional.

Computer type or compatible:
- ☐ IBM® PC/PC Jr. (128K RAM, DOS 2.0/2.1)
- ☐ Apple® IIe (64K RAM)
- Apple II+ (64K RAM) 80-col. card by: ☐ Apple ☐ Videx™

☐ Enroll me in the floppy-disk CME **only** at US$148.* (Note model above.)

☐ Check enclosed* ☐ Bill me

Signature ____________________

☐ VISA ☐ MasterCard Exp. Date ____________________

Account No. ____________________

Name ____________________

Specialty ____________________

Address ____________________

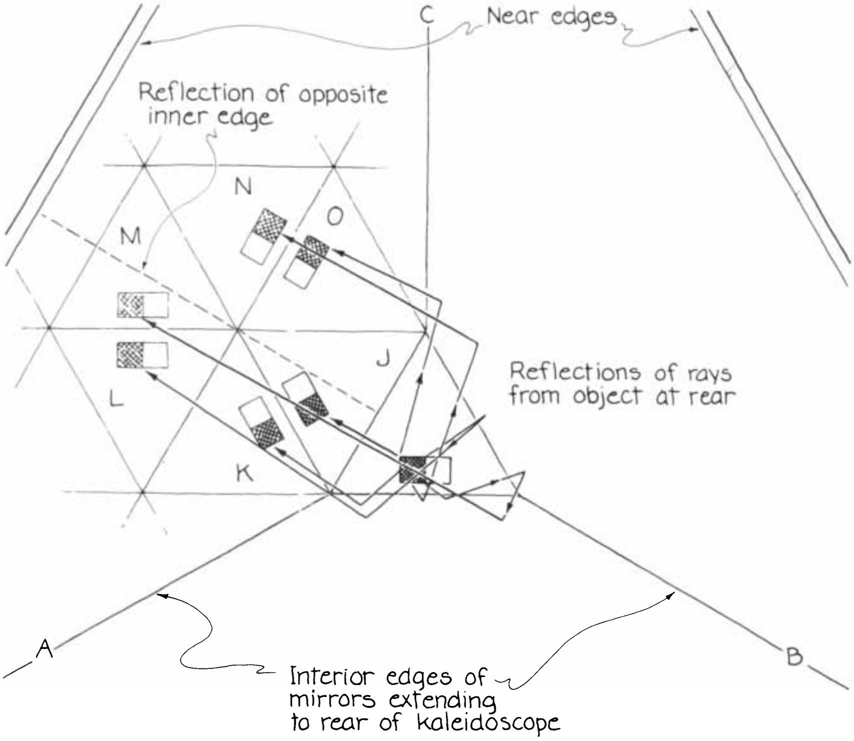
City ____________ State ____________ Zip Code ____________

*Add sales tax for Calif., Ill., Mich., or N.Y. Allow 8 weeks for delivery. Add US$10 for shipping to Canada. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computers, Inc. Videx is a trademark of Videx, Inc.

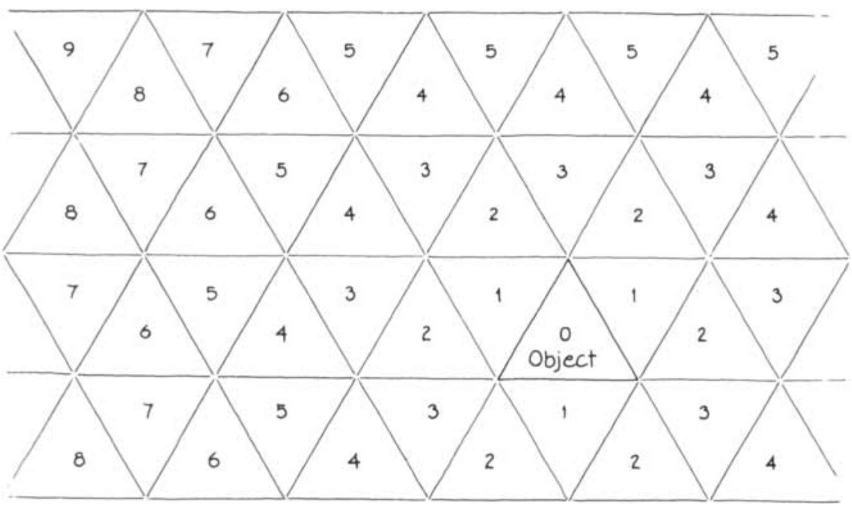
internal edges of the mirrors where they intersect one another. The bright, featureless background surrounding the sphere is the reflection by the mirrors of one another.

I also investigated image fields produced by mirrors in an isosceles triangle. (The mirrors were not flared.) Two mirrors formed the legs of the triangle and a third mirror served as the base. As I looked into the system I decreased the angle between the legs. More images appeared in the primary cluster around the intersection of the legs; the rest of the image field became cluttered with incomplete reflections of the scene.

When the angle was 22.5 degrees, the primary cluster consisted of the direct view and 15 copies of it (either exact or reflected copies). The rest of the image field consisted of only parts of that cluster [*see bottom illustration on opposite page*]. When I looked at the base mirror, I saw the center of the primary cluster. As I changed my angle of view I changed the part of the primary cluster I could see.



How light rays are reflected in an equilateral-triangle array



The number of reflections that create different sectors

The angle of view also determined what I saw in other parts of the field. One way to explain the rest of the field is to employ the concept of virtual mirrors. Note that the base mirror is reproduced in every reflection sector in the primary cluster. Each reproduction of the base mirror serves as a virtual mirror, capable of reflecting the primary cluster. (This reflection amounts to rotating the primary cluster about the virtual mirror until it is again in the image plane.)

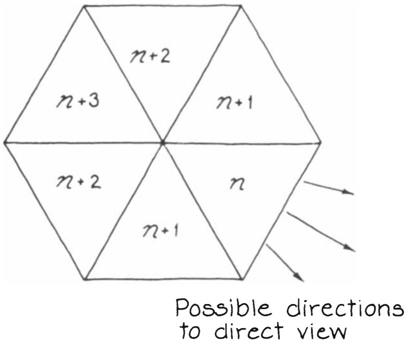
When I looked directly into a virtual mirror (as though it were real), I could see the center of the primary cluster, but adjacent virtual mirrors eliminated the rest of the reflection from that mirror. Where the rest of the cluster should have been I saw other perspectives of the cluster reflected by the adjacent virtual mirrors. (This dependence on perspective is not indicated in the illustration.) My most expensive kaleidoscope produces such a display. Initially I believed it must have 16 mirrors. (I dared not open such an expensive instrument to check.) Now I realize it has only three highly reflecting mirrors.

This type of image field is fundamentally different from the field formed by an equilateral triangle. The field is ambiguous, in the sense that the content at a particular place in the field depends on the angle of view into the system. A kaleidoscope with an equilateral triangle produces an unambiguous image field because the content at a particular place does not change when the angle of view is varied.

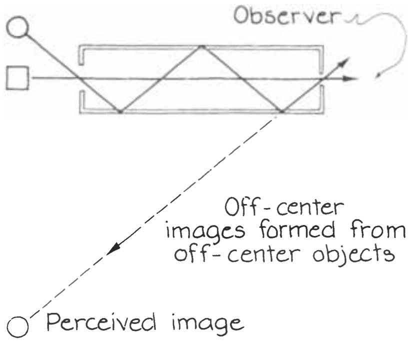
Ambiguous image fields are not as pleasing to me in spite of the beautiful symmetry in the primary cluster. Are there more configurations of three mirrors that give unambiguous fields? Can they be produced with systems containing more than three mirrors? I shall leave these questions for you with a simplifying question: Is there a geometric figure that can be reflected repetitively about each of its sides until a plane is filled with the figure without overlap? Next month I shall reveal the general solution I have discovered.

You might also enjoy studying systems of mirrors formed as a regular polygon with an even number of sides. Parts of the image fields from them should be similar to the scene generated by two parallel mirrors. Are there any differences? What happens when the polygon has an odd number of sides? What do you see when you look into a kaleidoscope in which the interior is a reflecting cylinder and thus is in effect a polygon with an infinite number of parallel sides? Would you expect a kaleidoscope with curved mirrors to yield unambiguous fields?

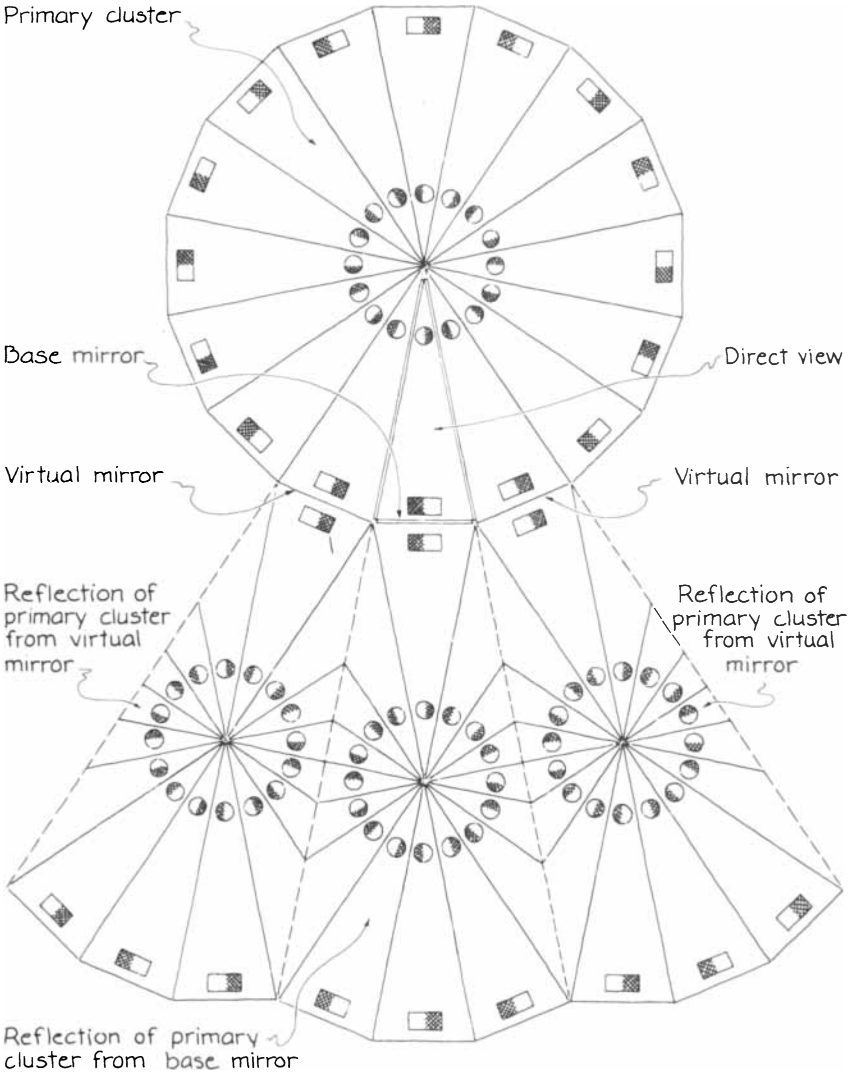
I bought my kaleidoscopes at High Tide / Rock Bottom, 1814 Coventry Road, Cleveland Heights, Ohio 44118. Many types of kaleidoscopes are also available from the Light Opera Gallery, Ghirardelli Square, No. 102, San Francisco, Calif. 94109. Kaleidoscopes by Peach is at 10731 Manchaca Road, Austin, Tex. 78748. Prism Design is at 87 Upper Main Street, Essex Junction, Vt. 05452. If you discover any unusual designs or invent some of your own, please let me know.



The sequence of reflections in a cluster



Imaging an object in front of a kaleidoscope



Kaleidoscope images from mirrors forming an isosceles triangle

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION (required by 39 U.S.C. 3685). *1A.* Title of publication: Scientific American. *1B.* Publication number: 00368733. *2.* Date of filing: September 27, 1985. *3.* Frequency of issue: monthly. *3A.* Number of issues published annually: 12. *3B.* Annual subscription price: U.S. and its possessions, 1 year $24; all other countries, 1 year $33. *4.* Complete mailing address of known office of publication: 415 Madison Avenue, New York, N.Y. 10017–1111. *5.* Complete mailing address of the headquarters or general business offices of the publisher: 415 Madison Avenue, New York, N.Y. 10017–1111. *6.* Full names and complete mailing address of publisher, editor, and managing editor: President, Jonathan Piel, 415 Madison Avenue, New York, N.Y. 10017–1111. Editor, Jonathan Piel, 415 Madison Avenue, New York, N.Y. 10017–1111. Managing Editor, none. *7.* Owner: Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017–1111; Armisac Profit Sharing Plan, c/o Michael R. Sonnenreich, Esq., Sonnenreich & Roccograndi, 600 New Hampshire Avenue, N.W., Washington, D.C. 20037; Atwell & Co. (for Estate of John Hay Whitney), P.O. Box 456, Wall Street Station, New York, N.Y. 10005–0456; Bache Halsey Stuart Shields Inc. for (a) H. R. Snellenburg Trust, (b) R. L. Rosenwald, Sr., Trust, 100 Gold Street, New York, N.Y. 10038–1605; George S. Conn, 20 Bittersweet Lane, Mount Kisco, N.Y. 10549–9666; Martin M. Davidson, 11 River Lane, Westport, Conn. 06880–1926; Dean Witter Reynolds Inc. for (a) L. J. Rosenwald Trust, (b) J. Rosenwald II, 130 Liberty Place, New York, N.Y. 10048; Dengel & Co. for E. G. Mannucci, c/o Fiduciary Trust Co. of New York, P.O. Box 3199, Church Street Station, New York, N.Y. 10008–3199; Richard E. Deutsch, c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008–1508; Egger & Co. (for B. Ewing), c/o The Chase Manhattan Bank, P.O. Box 1508, Church Street Station, New York, N.Y. 10008–1508; Evista B.V., c/o Michael R. Sonnenreich, Esq., Sonnenreich & Roccograndi, 600 New Hampshire Avenue, N.W., Washington, D.C. 20037; Dorothy D. Eweson, c/o Brunt & Co., P.O. Box 6626, Journal Square, Jersey City, N.J. 07306–0626; Excelsior Fund, 545 Madison Avenue, New York, N.Y. 10022–4212; William T. Golden, Room 4201, 40 Wall Street, New York, N.Y. 10005–2331; Jenk & Co. (for I. M. Scott and Joan R. Scott), c/o Industrial Valley Bank and Trust Co., York Road & W. Avenue, Jenkintown, Pa. 19046–2712; C. John Kirby, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017–1111; Frasier W. McCann, Suite 1706, 680 Fifth Avenue, New York, N.Y. 10019–5429; Eleanor Jackson Piel, Jr., 320 Central Park West, New York, N.Y. 10025–7659; Gerard Piel, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017–1111; Jonathan Piel, c/o Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017–1111; The President & Fellows of Harvard College Trust, 70 Federal Street, Boston, Mass. 02110–1906; Barbara K. Rosenwald, P.O. Box 3625, Norfolk, Va. 23514 3625; Stanley Schaefer, 418 Yerba Santa Avenue, Los Altos, Calif. 94022–2154. *8.* Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages or other securities: none. *9.* For completion by nonprofit organizations authorized to mail at special rates (Section 423.12, DMM only). The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes: Not applicable to this organization. *10.* Extent and nature of circulation: *A.* Total number of copies (net press run): average number of copies each issue during preceding 12 months, 787,404; actual number of copies of single issue published nearest to filing date, 761,558. *B.* Paid and/or requested circulation: *1.* Sales through dealers and carriers, street vendors and counter sales: average number of copies each issue during preceding 12 months, 137,865; actual number of copies of single issue published nearest to filing date, 141,047. *2.* Mail subscription (paid and/or requested): average number of copies each issue during preceding 12 months, 512,270; actual number of copies of single issue published nearest to filing date, 501,001. *C.* Total paid and/or requested circulation (sum of *10B1* and *10B2*): average number of copies each issue during preceding 12 months, 650,135; actual number of copies of single issue published nearest to filing date, 642,048. *D.* Free distribution by mail, carrier or other means, samples, complimentary, and other free copies: average number of copies each issue during preceding 12 months, 2,842; actual number of copies of single issue published nearest to filing date, 2,750. *E.* Total distribution (sum of *C* and *D*): average number of copies each issue during preceding 12 months, 652,977; actual number of copies of single issue published nearest to filing date, 644,798. *F.* Copies not distributed: *1.* Office use, leftover, unaccounted, spoiled after printing: average number of copies each issue during preceding 12 months, 12,283; actual number of copies of single issue published nearest to filing date, 3,124. *2.* Return from news agents: average number of copies each issue during preceding 12 months, 122,144; actual number of copies of single issue published nearest to filing date, 113,636. *G.* Total (sum of *E, F1* and *2*—should equal net press run shown in *A*): average number of copies each issue during preceding 12 months, 787,404; actual number of copies of single issue published nearest to filing date, 761,558. *11.* I certify that the statements made by me above are correct and complete. (Signed) Jonathan Piel, President and Editor.

ANNUAL INDEX

The following index lists all the authors and articles that appeared in SCIENTIFIC AMERICAN *during 1985. Also indexed are "Computer Recreations" and "The Amateur Scientist."*

AUTHORS

ARTICLES

COMPUTER RECREATIONS

THE AMATEUR SCIENTIST

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

MINIMUM-REDUNDANCY LINEAR ARRAYS. Alan T. Moffet in *IEEE Transactions on Antennas and Propagation*, Vol. AP-16, No. 2, pages 172–175; February, 1963.

APPLICATIONS OF NUMBERED UNDIRECTED GRAPHS. Gary S. Bloom and Solomon W. Golomb in *Proceedings of the IEEE*, Vol. 65, No. 4, pages 562–570; April, 1977.

GOLOMB'S GRACEFUL CURVE. Martin Gardner in *Wheels, Life and Other Mathematical Amusements*. W. H. Freeman and Company, 1983.

VARIATIONS IN THE ROTATION OF THE EARTH. W. E. Carter, D. S. Robertson, J. E. Pettey, B. D. Tapley, B. E. Schutz, R. J. Eanes and Miao Lufeng in *Science*, Vol. 224, No. 4652, pages 957–961; June 1, 1984.

THE DEVELOPMENT OF SOFTWARE FOR BALLISTIC-MISSILE DEFENSE

SAFEGUARD DATA-PROCESSING SYSTEM. *The Bell System Technical Journal*, Vol. 54, Special Supplement; 1975.

SOFTWARE ENGINEERING ECONOMICS. Barry W. Boehm. Prentice-Hall, Inc., 1981.

OPTICAL PHASE CONJUGATION

OPTICAL PHASE CONJUGATION. Edited by R. A. Fisher. Academic Press, 1983.

OPTICAL PHASE CONJUGATION. Special issue of *Journal of the Optical Society of America*, Vol. 73, No. 5; May, 1983.

PRINCIPLES OF PHASE CONJUGATION. B. Ya. Zel'dovich, N. F. Pilipetsky and V. V. Shkunov. Springer-Verlag, 1985.

CRICKET AUDITORY COMMUNICATION

TYMPANAL MEMBRANE MOTION IS NECESSARY FOR HEARING IN CRICKETS. Hans-Ulrich Kleindienst, David W. Wohlers and Ole Naesbye Larsen in *Journal of Comparative Physiology*, Vol. 151, pages 397–400; 1983.

AUDITORY BEHAVIOR OF THE CRICKET, 3: TRACKING OF MALE CALLING SONG BY SURGICALLY AND DEVELOPMENTALLY ONE-EARED FEMALES, AND THE CURIOUS ROLE OF THE ANTERIOR TYMPANUM. Franz Huber, H.-U. Kleindienst, Theo Weber and John Thorson in *Journal of Comparative Physiology*, Vol. 155, pages 725–738; 1984.

TEMPORAL SELECTIVITY OF IDENTIFIED AUDITORY NEURONS IN THE CRICKET BRAIN. Klaus Schildberger in *Journal of Comparative Physiology*, Vol. 155, pages 171–185; 1984.

THE IMMUNE SYSTEM IN AIDS

FREQUENT DETECTION AND ISOLATION OF CYTOPATHIC RETROVIRUSES (HLTV-III) FROM PATIENTS WITH AIDS AND AT RISK FOR AIDS. Robert C. Gallo, Syed Z. Salahuddin, Mikulas Popovic, Gene M. Shearer, Mark Kaplan, Barton F. Haynes, Thomas J. Palker, Robert Redfield, James Oleske, Bijan Safai, Gilbert White, Paul Foster and Phillip D. Markham in *Science*, Vol. 224, No. 4648, pages 500–502; May 4, 1984.

IMMUNOREGULATORY LYMPHOKINES OF T HYBRIDOMAS FROM AIDS PATIENTS: CONSTITUTIVE AND INDUCIBLE SUPPRESSOR FACTORS. Jeffrey Laurence and Lloyd Mayer in *Science*, Vol. 225, No. 4657, pages 66–69; July 6, 1984.

LYMPHADENOPATHY-ASSOCIATED VIRAL ANTIBODY IN AIDS: IMMUNE CORRELATIONS AND DEFINITION OF A CARRIER STATE. Jeffrey Laurence, Françoise Brun-Vezinet, Steven E. Schutzer, Christine Rouzioux, David Klatzmann, Françoise Barré-Sinoussi, Jean-Claude Chermann and Luc Montagnier in *The New England Journal of Medicine*, Vol. 311, No. 20, pages 1269–1273; November 15, 1984.

IMMUNOLOGIC ABNORMALITIES IN THE ACQUIRED IMMUNODEFICIENCY SYNDROME (AIDS). A. S. Fauci in *Clinical Research*, Vol. 32, No. 5, pages 491–499; December, 1984.

POLAR WANDERING ON MARS

SOME REMARKS ON POLAR WANDERING. Peter Goldreich and Alar Toomre in *Journal of Geophysical Research*, Vol. 74, No. 10, pages 2555–2567; May 15, 1969.

THE SURFACE OF MARS. Michael H. Carr. Yale University Press, 1981.

GRAZING IMPACTS ON MARS: A RECORD OF LOST SATELLITES. Peter H. Schultz and Anne B. Lutz-Garihan in *Journal of Geophysical Research*, Vol. B87, Supplement 1, pages A84–A96; November 15, 1982.

THE ENORMOUS THEOREM

ON THE STRUCTURE OF GROUPS OF FINITE ORDER. Richard Brauer in *Proceedings of the International Congress of Mathematicians*, Vol. 1, pages 209–217; 1954.

FINITE GROUPS. Daniel Gorenstein. Chelsea Publishing Co., 1980.

FINITE SIMPLE GROUPS: AN INTRODUCTION TO THEIR CLASSIFICATION. Daniel Gorenstein. Plenum Press, 1982.

THE FRIENDLY GIANT. Robert L. Griess, Jr., in *Inventiones Mathematicae*, Vol. 69, No. 1, pages 1–102; 1982.

CHINA'S FOOD

AGRICULTURE IN CHINA'S MODERN ECONOMIC DEVELOPMENT. Nicholas R. Lardy. Cambridge University Press, 1983.

THE BAD EARTH: ENVIRONMENTAL DEGRADATION IN CHINA. Vaclav Smil. M. E. Sharpe, Inc., 1984.

RURAL DEVELOPMENT IN CHINA. Dwight Perkins and Shahid Yusuf. A World Bank Publication, The Johns Hopkins University Press, 1984.

THE CONSTRUCTION PLANS FOR THE TEMPLE OF APOLLO AT DIDYMA

DIDYMA. Theodor Wiegand. Verlag Gebr. Mann, Berlin, 1941–1958.

VORARBEITEN ZU EINER TOPOGRAPHIE VON DIDYMA. Klaus Tuchelt. Deutsches Archäologisches Institut, Istanbuler Mitteilungen, Beiheft 9, Verlag Ernst Wasmuth, Tübingen, 1973.

GREEK ARCHITECTS AT WORK: PROBLEMS OF STRUCTURE AND DESIGN. J. J. Coulton. Cornell University Press, 1977.

THE AMATEUR SCIENTIST

THE NUMBER OF IMAGES OF AN OBJECT BETWEEN TWO PLANE MIRRORS. An-Ti Chai in *American Journal of Physics*, Vol. 39, No. 11, pages 1390–1391; November, 1971.

MULTIPLE IMAGES IN PLANE MIRRORS. Thomas B. Greenslade, Jr., in *The Physics Teacher*, Vol. 20, pages 29–33; January, 1982.

REFLECTIONS IN A POLISHED TUBE. Laurence A. Marschall and Emma Beth Marschall in *The Physics Teacher*, Vol. 21, page 105; February, 1983.

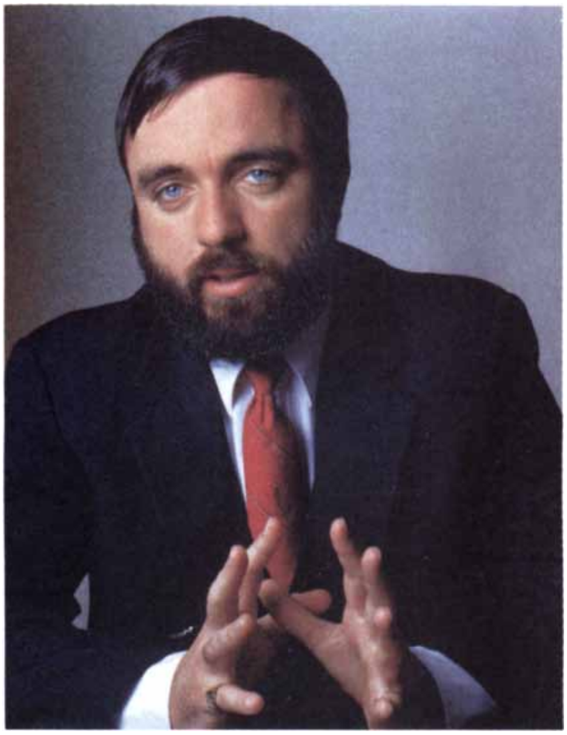
THROUGH THE KALEIDOSCOPE. Cozy Baker. Beechcliff Books, 100 Severn Avenue, Suite 605, Annapolis, Md. 21403; 1985.


DODGE LANCER
5/50 PROTECTION, STANDARD

A "business partnership" with Hewlett-Packard

► helps ETC Corporation reduce chemical analysis costs by one-third and double productivity.

► helps Container Corporation of America improve response time to customers--and save $600,000 a year.



John Fitzgerald, Vice President
Environmental Testing and Certification Corporation

"We use an HP 3357 lab automation system for chemical analyses and an HP 3000 business computer to generate reports and operate our database management services. The accuracy and ease of use of these HP systems have enabled us to make impressive productivity gains while reducing test costs.

"HP was the only vendor who could automate both the analytical and business sides of our operation and combine the systems into one network.

"Within a week of delivery, all HP systems were up and running: 24 hours a day, 7 days a week. Downtime costs us $30-$40 thousand an hour, so HP's near 100% reliability is essential to our success."

Jeff Norkin, Vice President
Container Corporation of America

"We have a network of 52 HP 3000 computers operating in 77 manufacturing sites. It enables us to develop estimates for customers in minutes—a process that previously took two days. Moreover, the network gives us vital information for enhancing the quality of our packaging products. All this with an annual savings of $600,000.

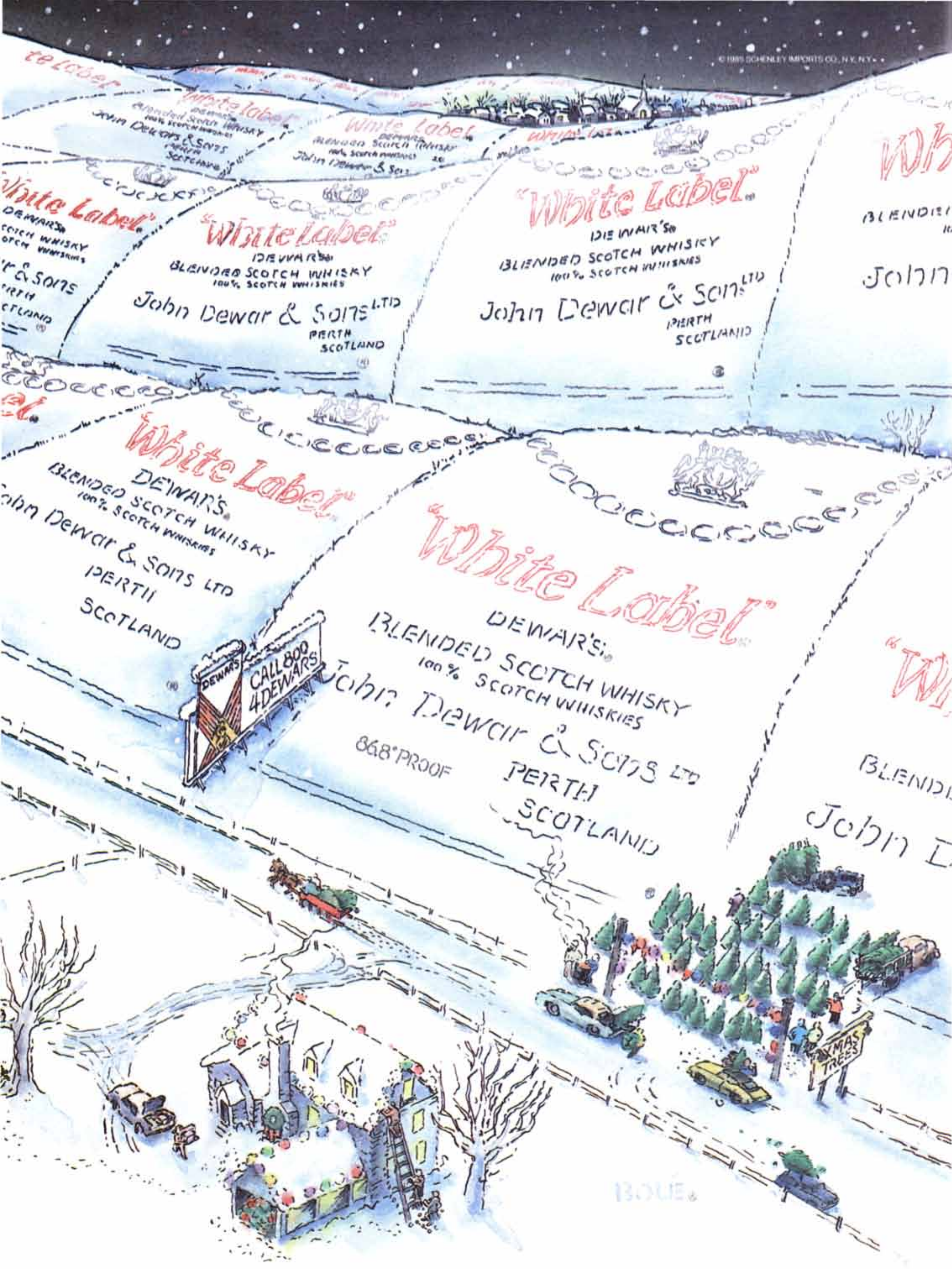
"With HP's support and the flexibility of our HP 3000 network, we can design and implement additional systems, as needed, without costly conversions.

"We can count on HP's technical expertise to provide solutions to new challenges, and we plan to expand our network with additional HP 3000s."

Isn't your business worth HP?

hp HEWLETT PACKARD

0002512A


© 1985 SCHENLEY IMPORTS CO., N.Y., N.Y.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
John Dewar & Sons LTD
86.8° PROOF
PERTH
SCOTLAND
DEWAR'S
CALL 800 4DEWARS
XMAS TREES
BOUE