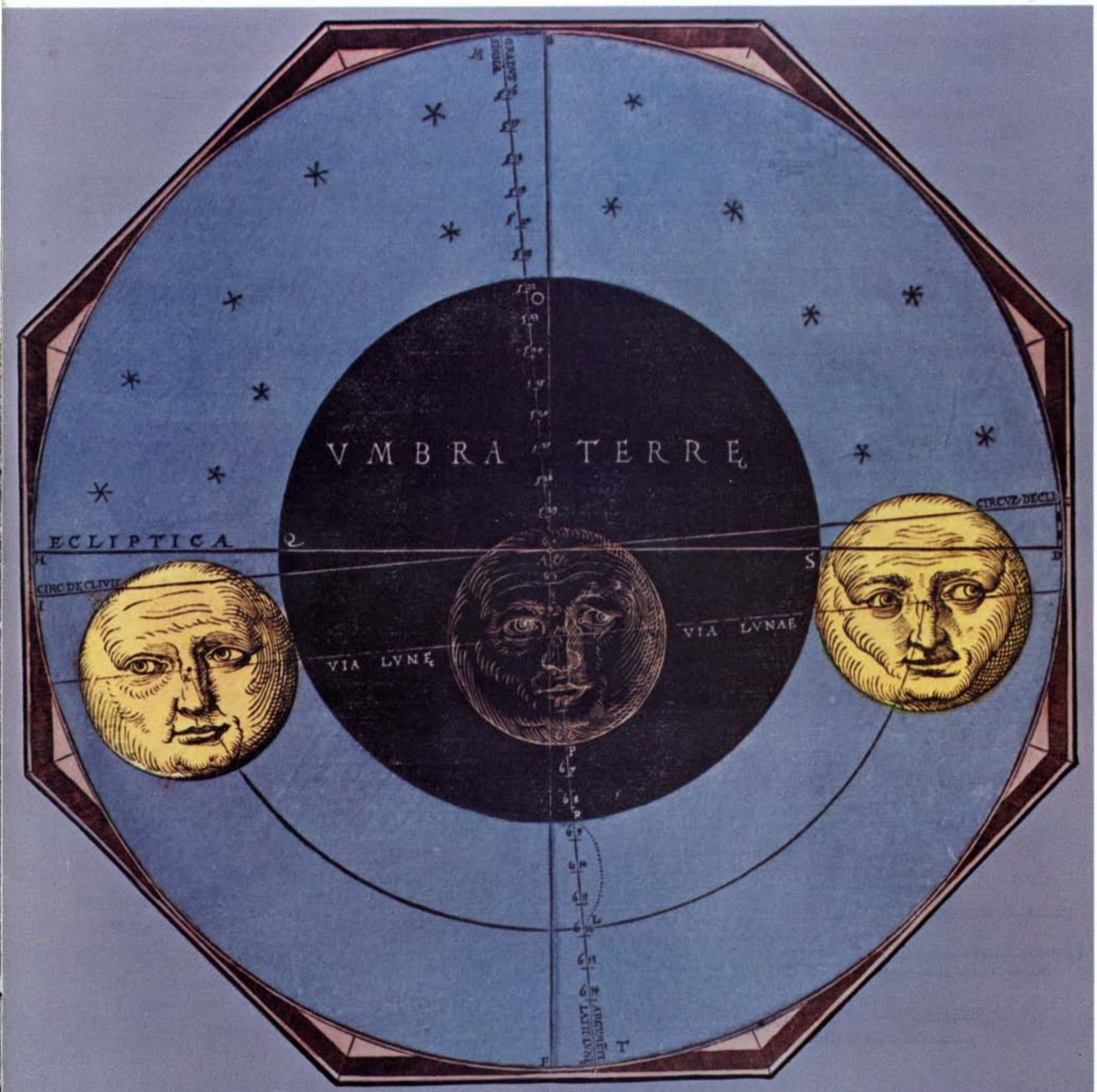
SCIENTIFIC AMERICAN



HISTORICAL ECLIPSES

$2.00

October 1982

The Mercedes-Benz 300 D Turbodiesel: America's best-performing diesel is also one of the world's most civilized sedans.

A diesel that loves hills and sprints away from stoplights... a sedan that laps up sports car roads. The 300 D Turbodiesel is both. Meanwhile, it coddles you and your passengers in that secure comfort unique to the automobiles of Mercedes-Benz.

Mercedes-Benz engineers have extracted more sheer power from the 300 D Turbodiesel Sedan's five-cylinder engine than any other engineers have yet extracted from any other passenger car diesel, huge domestic V-8's included.

Their secret was to couple an advanced engine design with a power-boosting turbocharger. Even other turbocharged diesels are left flatfooted; this spacious, solid, 3585-lb. five-passenger Mercedes-Benz sedan ranks as the best performer of any diesel automobile sold in America.

The 300 D Turbodiesel is meanwhile perhaps the most *conscientious* performance automobile in America, because it is so inherently efficient that EPA figures show 27 mpg city estimated and 33 highway.* An almost uncanny *balance* of power and frugality is thus struck.

So exotic is engine technology that the interior of each moving piston is constantly cooled by precisely timed jets of oil. But deep down this is still a diesel: a rugged, reliability-minded diesel that will never require a conventional tune-up.

Do not disturb

Over-the-road performance is superb. This means not only that the 300 D can swallow up roller-coaster back-country roads, and nip rather than lurch around corners–but that its occupants usually feel so *undisturbed* in the process.

Generations of Mercedes-Benz engineers have toiled since 1931 to develop, refine, and refine again the principle of four-wheel independent suspension that is primarily responsible.

These engineers have endowed the car with an exquisitely responsive steering mechanism, guided by 24 recirculating ball bearings. It feels liquid-smooth and seemingly frictionless. The shift lever is mounted down on the transmission tunnel–but look again. The quadrant reads P-R-N-D-S-L. It is no mere automatic but a four-speed automatic, and is designed to serve equally well when shifted by hand.

America's most advanced automotive braking system is four-wheel *disc* braking. An 11-inch disc brake is fitted to every wheel of every 300 D Turbodiesel Sedan.

Mercedes-Benz engineers care little if the outside world knows that total swept brake area is a massive 456.5 square inches, or that the 300 D's front suspension geometry is devised to help minimize front-end "nose diving" in hard stops. Their main concern is that the driver is better served by these and myriad other technical flourishes.

Designed, not decorated

Even when fully padded and finished, the interior of the 300 D's welded steel body shell measures almost five feet in width–sufficient to easily accommodate three adults in the rear seating area, for example. Electronic cruise control, AM/FM stereo radio/cassette player, automatic climate control system and electric window lifts–whatever conventional luxury sedans furnish by way of real conveniences, so does this sedan.

Garishness, however, is absent. The instrument panel is no entertainment center twinkling with gadgetry but an exercise in ergonomics, meant to ease the driver's task by simplifying it.

The engineers continuously strive to make driver controls fewer and less awkward to use, and so reduce driving complexity. Item: the single steering-column lever that lets you perform nine different driving functions–without taking your hands off the wheel.

Seats are not designed to *look* luxurious but to so support your body that you *feel* relaxed, even after an all-day drive.

Interior wood trim is genuine wood from the Mercedes-Benz shops, handworked and fitted and finished. A small thing, but it epitomizes this car's unremitting honesty of character–as its 120 built-in safety features epitomize its seriousness of character.

The luxury of retained value

The 300 D Turbodiesel boasts a final and powerful distinction. As a Mercedes-Benz, it shares a name so coveted by American buyers today that after the first three years, the *entire Mercedes-Benz line*–not just a few isolated models–has been shown to retain an average of 84 percent of original value.

If retained value is a form of luxury, then perhaps the 300 D Turbodiesel *is*, after all, a luxury car.

*EPA estimate for comparison purposes. The mileage you get may vary with trip length, speed and weather. Actual highway mileage will probably be less.

Engineered like no other car in the world


80 PROOF · CALVERT DISTILLERS CO. N.Y.C.
JOHN JAMESON & SON
ESTD 1780
SINE METU
IRISH WHISKEY
A BLEND
DISTILLED IN IRELAND BY
LIMITED
BOW STREET DISTILLERY
DUBLIN
IMPORTED
Now that you're ready for a change of pace
it's time to try John Jameson.
Take a sip of John Jameson. Note the light, delicate taste.
Luxurious and smooth as you would expect a premium whiskey to be.
But with a distinctive character all its own. Set a new pace
for yourself. Step ahead of the crowd with John Jameson, the
world's largest selling Irish Whiskey.

Established 1845 SCIENTIFIC AMERICAN October 1982 Volume 247 Number 4

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: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

executive health

the report that briefs you on what to watch

Volume XVIII, Number 11 • P.O. Box 589, Rancho Santa Fe, Calif. 92067 • Area 714:756-2600

Lot B. Page, M.D.:

ON MAKING SENSE OF SALT AND YOUR BLOOD PRESSURE

An authoritative report on how much salt intake may matter to you if excessive . . . and how much good moderation can do . . . by a preeminent expert who, literally, has gone almost to the ends of the earth to find out.

PUBLISHER'S NOTE: *Chief of Medicine at the Newton-Wellesley Hospital and Professor of Medicine at Tufts University School of Medicine, Boston, Dr. Lot B. Page has been called "the preeminent scholar in this field (hypertension) in the United States; in fact, in the world." His studies of high blood pressure and what influences it have taken him on medical expeditions to many places abroad, including the Solomon Islands, to investigate pre-industrial populations. He has also studied nomads in the deserts of Iran. We have asked him to assess hypertension and its importance, the factors that influence it, the role of salt in elevating blood pressure, and how much you may gain by examining the salty side of your life.*

Unless you've been off somewhere remote for much of the last decade or so you couldn't have escaped becoming aware of a whole series of facts about hypertension or high blood pressure—and about salt as well.

About hypertension, it comes to this: The disease, silent though it is, is a killer—an almost incredibly

THE COVER

The photograph on the cover shows one of the hand-tinted plates in a famous 16th-century work by the German cosmologist Petrus Apianus, published in Ingolstadt in 1540. Titled *Astronomicum Caesareum* ("Imperial Astronomy"), it was dedicated to the Holy Roman Emperor Charles V and his brother and chosen successor, Ferdinand I. The work was illustrated by the engraver Michael Ostendorfer. The plate shows three stages of the moon's progress during the course of a total lunar eclipse: as it approaches, passes into and leaves the shadow of the earth (see "Historical Eclipses," by F. Richard Stephenson, page 170). The engraver has exercised artistic license to suggest that the Man in the Moon has been somewhat distressed by the dark experience. The copy of *Astronomicum Caesareum* that furnished this illustration is in the Metropolitan Museum of Art. It was presented to the museum in 1925 by Herbert N. Straus.

THE ILLUSTRATIONS

Cover photograph courtesy of the Metropolitan Museum of Art

Page	Source
21	Letraset USA Inc.
34	*The Watercolours and Drawings of Thomas Bewick and his Workshop Apprentices.* © 1981 The MIT Press
48–51	Alan D. Iselin
52	Lynn R. Sykes
53–55	Alan D. Iselin
56–58	Tom Prentiss
59	David W. Pumplin, University of Maryland at Baltimore School of Medicine, and Thomas S. Reese, National Institute of Neurological and Communicative Disorders and Stroke
60	Ilil Arbel (*top*), Tom Prentiss (*bottom*)
61–62	Ilil Arbel
63	Ilil Arbel (*top*); David W. Pumplin, University of Maryland at Baltimore School of Medicine, and Thomas S. Reese, National Institute of Neurological and Communicative Disorders and Stroke (*bottom*)
64–65	Ilil Arbel
66	Mount Wilson and Las Campanas Observatories (*top*), R. F. Bonifield (*bottom*)
68	Ilil Arbel
69	Texas Instruments
70	Texas Instruments (*top*), Ilil Arbel (*bottom*)
71–74	Mount Wilson and Las Campanas Observatories
119–131	Jerome Kuhl
137–139	Ronald E. Rosensweig
140–141	Gabor Kiss
142	Gabor Kiss (*top*), Ronald E. Rosensweig (*bottom*)
143–144	Gabor Kiss
147	Judith Tricker, courtesy of R. A. R. Tricker (*top*); David K. Lynch (*bottom*)
148	Susan Bartsch-Winkler (*top*); C. G. Kershaw, Severn-Trent Water Authority (*bottom*)
149–150	Andrew Tomko
151–154	Ian Worpole
159–160	Karl Hartig
161–162	Thomas D. Seeley
163	Patricia J. Wynne (*top*), Karl Hartig (*bottom*)
166–168	Karl Hartig
171	Owen Gingerich
172–173	Andrew Tomko
174	© British Museum
175	F. Richard Stephenson
176–180	Andrew Tomko
182	E. H. Strach
186	Lorne A. Whitehead
188–192	Michael Goodman

LETTERS

Sirs:

While reading Owen Gingerich's excellent article "The Galileo Affair" [SCIENTIFIC AMERICAN, August], I, as he, experienced a sense of déjà vu when I read Cardinal Caetani's instructions for the censorship of Copernicus' *De revolutionibus orbium coelestium:* "If certain of Copernicus' passages on the motion of the earth are not hypothetical, make them hypothetical; then they will not be against either the truth or the Holy Writ." Professor Gingerich correctly pointed out that the creationists are trying to have biology textbooks present evolution in the same way, since the occurrence of evolution would be acceptable if presented as only a hypothesis. Unfortunately Professor Gingerich is incorrect in thinking that the creationists would have the same lack of success as the Holy Congregation of the Index, since he is apparently aware of only the creationists' failure in California. The creationists have been successful elsewhere.

Texas has its own Holy Congregation of the Index, the State Board of Education. Since 1974 the State Board's textbook-policy proclamation has required that science textbooks seeking adoption by the state "that treat the theory of evolution should identify it as only one of several explanations of the origins of humankind," that textbooks "shall be edited, if necessary, to clarify that the treatment is theoretical rather than factually verifiable," that "each textbook must carry a statement on an introductory page that any material on evolution included in the book is clearly presented as theory rather than fact" (Texas' version of the Nihil Obstat) and that "the presentation of the theory of evolution should be done in a manner which is not detrimental to other theories of origin." Since the factual occurrence of evolution is not in question among scientists, the prevarications contained in this textbook-adoption proclamation are a misrepresentation of science. The word "theory" in the context of the proclamation is not the scientific definition of theory but the popular idea of theory as a speculation or guess....

Since Texas, as an adoption state, is the nation's largest single purchaser of textbooks, and since other states adopt the same textbooks that Texas chooses, publishers have written their precollege science textbooks to comply with the Texas requirements. Over the past eight years the result has been not only the inclusion of equivocations and misrepresentations about evolution but also the reduction of coverage of evolution to a couple of pages or nothing, the omission of any connection between evolution and other biological phenomena and even the inclusion of pro-creationist statements....

STEVEN SCHAFERSMAN

Rice University
Houston, Tex.

Sirs:

In "Metamagical Themas" for August, Douglas R. Hofstadter (whose work is a delight) apparently confers on the gene theory of Richard Dawkins some form of acceptance, viz., "Dawkins uses the concept to show how *group* selection can seem to be taking place in a population when in fact mere *gene* selection can account for what is observed."

I would guess that Darwinians will forgive Hofstadter for utilizing Dawkins' insights into game theory but will frown on his acceptance, if it is that, that *gene* selection, not *individual* (*body*) selection is involved. Stephen Jay Gould, writing in his book *The Panda's Thumb,* notes: "No matter how much power Dawkins wishes to assign to genes, there is one thing that he cannot give them—direct visibility to natural selection. Selection simply cannot see genes and pick among them directly.... Bodies cannot be atomized into parts, each constructed by an individual gene.... Selection doesn't work directly on parts. It accepts or rejects entire organisms because suites of parts interacting in complex ways confer advantages."

ED BURNETT

New York, N.Y.

Sirs:

Arlette Leroi-Gourhan's article "The Archaeology of Lascaux Cave" [SCIENTIFIC AMERICAN, June] provides a fascinating overview of the range and finesse of the inferences to which the use of modern analytical methods in archaeological investigation can give rise. One tentative inference she might possibly have drawn is that the Paleolithic inhabitants who decorated the cave were predominantly right-handed. Not only do the majority of the animals depicted face right (an orientation generally preferred by right-handed draftsmen) but also the "fossilized" rope depicted in the article was obviously the work of a right-handed individual. Since the natural twisting motion involves the overhand rotation of the thumb away from the body, the twist in this case must have been imparted by the right hand.

THOMAS A. REISNER

Université Laval
Quebec

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 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Student 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 $2 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, Mass. 01970. Specify CCC Reference Number ISSN 0036–8733/82. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 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, N.Y. 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 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 ______

This Peugeot 505S is travelling at 80 miles an hour in the silence of the Sahara Desert. A computer is "listening" to noises that you should not be permitted to hear.

From Europe's second-largest carmaker comes the Peugeot 505 S, seen here undergoing silence tests in the soundproof chamber.

A silver car sits in a strange golden room. The engine of the car is humming, the rear wheels are spinning on rollers at 80 miles an hour, and something feels very, very wrong.

The room should be awash with sound. Instead, it is uncannily quiet. The walls and ceiling are receiving the sound waves... and swallowing them.

In pursuit of a quiet car

This is the anechoic chamber at Peugeot, a sonic-test room that replicates the almost deathly hush of the empty desert.

The room is tremor proof. Its door is five and a half feet thick. Golden dihedrons of glass-fiber panels smother up to 99 percent of the sound waves so that they cannot ricochet back and disturb the probing microphones at work.

Connected to these microphones is a computer. It is "listening" for noises that should not exist. If there is a panel vibration, or a buzzing from the dashboard or transmission, the computer will pinpoint the noise.

Peugeot has little tolerance for cars with mysterious rattles and squeaks.

The Peugeot Lion

The lion has represented Peugeot since 1858. It was Peugeot that built the first car ever sold commercially (1891)... the world's first diesel-powered car (1922)... and the first high-rpm diesel engine (1967).

Listen with your eyes

A word of caution: Do not be fooled by the quietness of a car. Quietness, by itself, is no guarantee of first-rate construction.

Trust your eyes—and an alert brain—before you trust your ears. Open the door of a Peugeot 505S and peer at the hinges. They are machined from hot-rolled steel. Peugeot believes that its owners will not abide doors that sag later in life.

Notice how smoothly the seats move back and forth. The reason: they glide on tiny rollers instead of simple metal slides.

The body is welded in 4,032 places, then treated with 21 anti-corrosion measures. One of these measures involves dipping the body in an electrically charged bath that attracts the primer to the metal like a magnet. Hydrogen is used so there are no tiny bubbles of oxygen that can breed rust.

The meaning of "S"

Peugeot does not believe in nibbling you to death with extra-cost options. When you spend $13,990* for a gasoline-engined 505 S, this is the "S" (for Special) equipment that comes with the car:

- Air conditioning, factory-installed
- Electric sunroof
- Cruise control
- Electric windows
- Digital AM/FM 40-watt stereo radio with cassette player, scan tuning, and four speakers
- Automatic electric antenna
- Central locking system—a twist of the key locks all four doors and fuel-filler door
- Alloy wheels
- Multi-adjustable driver's seat.

So fully equipped is the car that the only other ways to spend your money are by ordering a 3-speed automatic ($370) in place of the 5-speed overdrive manual gearbox, and by specifying metallic paint ($295).

New York is noted for its punishingly rough streets. Cabdrivers are switching in droves to Peugeot 505 diesels, despite their higher cost.

An aversion to silly doodads

The Peugeot 505 S is how much car you should expect for $14,000 today. It positively bristles with *useful* equipment rather than gee-whiz gimmickry.

The engine is fitted with silver-tipped spark plugs. And magnetic-pulse ignition. And precisely metered fuel injection. It is designed to shrug off tune-ups for 30,000 miles.

The brakes are disc brakes—*on all four wheels*—because discs shed water and dissipate heat faster than drum brakes.

The shock absorbers contain pistons with four times as many valves as normal shock absorbers. These awesome damping devices have the almost eerie ability to reduce bumps in the road to little more than ripples.

A rare and expensive test

Most cars are *presumed* to be in proper working order when they roll off the assembly line. This is not Peugeot's way; it takes nothing for granted.

Every single Peugeot is put through a road test on a special track before it is permitted to leave Europe for the U.S. One of Peugeot's 65-member team of test drivers will personally drive the car.

The driver will test the steering, brakes, lights, engine performance, transmission, and suspension. Listen for rattles and vibrations. Check the instruments and controls. Try the heating and air-conditioning systems.

A rarity today, this final road testing costs a small fortune. Peugeot considers it indispensable in maintaining a high level of quality.

Peugeot has little tolerance for poorly made cars.

See your dealer for details of Peugeot's limited warranties.

*Based on P.O.E. manufacturer's suggested retail prices. Actual prices may vary according to local dealer. Destination charges ($255), state and local taxes, dealer preparation, if any, and license fees are extra.

For name of nearest Peugeot dealer, call **1-800-447-2882** toll free in the continental U.S. In Illinois, 1-800-322-4400.

Please send me the Peugeot Facts Kit, complete with brochures, road tests by leading car magazines, plus list of my nearest dealers.

☐ Please include details on Overseas Delivery.

Name__________

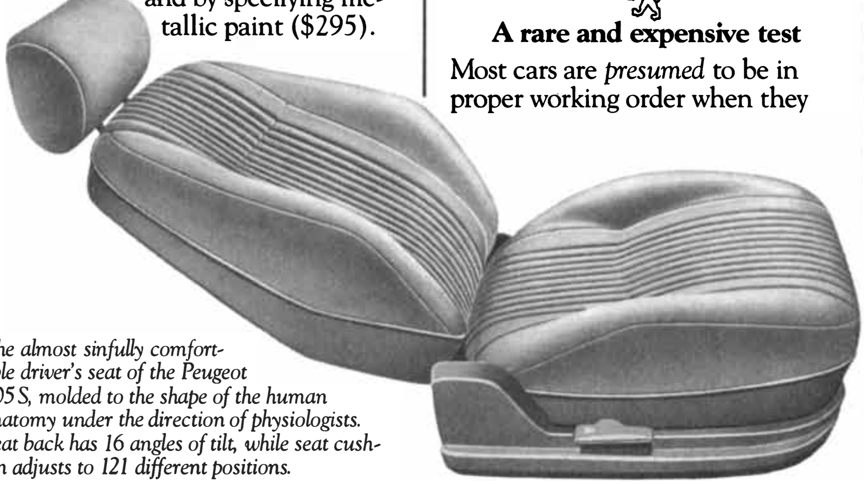
Address__________

City__________

State__________ Zip__________

Peugeot Motors of America, Inc., One Peugeot Plaza, Lyndhurst, New Jersey 07071

69-000SA-0010



The almost sinfully comfortable driver's seat of the Peugeot 505 S, molded to the shape of the human anatomy under the direction of physiologists. Seat back has 16 angles of tilt, while seat cushion adjusts to 121 different positions.

PEUGEOT

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

OCTOBER, 1932: "When an independent individual makes a scientific discovery, he is free to make a fortune, if he can, from its proceeds, but when a scientist attached to a university makes one, there is a tradition that he must exert no effort to profit personally by it but must publish it for the benefit of humanity. Too often this means that humanity receives the benefit only after an astute person who has never contributed the shred of an idea toward the original discovery reaps a munificent profit from something he did not create. Professor Winterton C. Curtis of the University of Missouri recommends that this tradition be broken in the following manner: university scientists will go through the formality of taking out patents on their discoveries wherever practicable and will assign these to permanent holding companies organized by the universities. The scientist would not receive personal gain but he would have the satisfaction of seeing a fair share of the ultimate profit returned to his institution."

"The subject of lift-increase devices for aircraft has been popular with inventors and designers for many years. Now the National Advisory Committee for Aeronautics has become interested in the subject, and its studies have been more thorough and systematic than heretofore. What is a lift-increase device? It is a method not of increasing the lift throughout the flight range but of so changing the characteristics of a wing that it has its normal characteristics in normal flight, experiencing an increase in lift when the aircraft is landing or gliding down toward a landing. Thus it is possible to have a low-lift and high-speed wing, and when we want to land, make it a high-lift, slow-landing wing. The maximum lift of an ordinary wing can be increased from 50 to 150 per cent by the aid of such devices. The landing speed can be decreased correspondingly. Perhaps the history of such devices will be that of the four-wheel brake on the automobile. At one time no manufacturer would consider the four-wheel brake seriously. Then one employed it and made a special feature of it. Then every car adopted it."

"It has been realized since the work of Lockyer and Young, more than 50 years ago, that different sunspots show spectra that, though notably different from the spectrum of the sun's normal surface, are practically identical among themselves, and so we can speak of the sunspot spectrum as a definite entity. A detailed study of the entire sunspot spectrum has recently been made by Miss Moore, working with the tower telescope and the great spectroscope at Mount Wilson. Methods for estimating the number of atoms involved in producing solar lines of various intensities were already available, so that the various influences affecting the sunspot spectrum could be expressed numerically. For example, the diminution in the number of excited atoms due to the lower temperature of a spot was found to correspond to a factor of .65 for each volt of excitation. Taking the temperature of the sun's disk as 5,740 degrees centigrade, the generally accepted value, this leads to a temperature of 4,720 degrees for the spots."

SCIENTIFIC AMERICAN

OCTOBER, 1882: "M. Stéphane Tarnier, the surgeon of the Maternity Hospital in Paris, struck by the great mortality among infants prematurely born, and those that are very sickly after birth, has conceived the ingenious idea of constructing a box that is almost exactly similar to the incubators used for poultry. The box is divided into two compartments, the lower one being used as a reservoir for hot water. The infant is placed in the upper one, which is fitted with a sliding glass cover. The temperature is maintained at 86 degrees Fahr., and M. Tarnier has found that by keeping infants in the incubator for a period varying from two days to six weeks their vitality is enormously improved. He has made experiments upon five six-month children, six seven-month children and 13 eight-month children, and he has lost only two of them. According to his statement three-fourths of them would have died but for this adventitious aid to their vitality."

"Clemens Zimmerman has prepared metallic uranium by Peligot's method and has determined some of its properties. Its specific gravity is 18.685, and its specific heat between 99° and 0° is 0.02765. The latter number multiplied by 240 gives a product of 6.64, which agrees with the mean atomic heat indicated by Dulong and Petit's law. The controversy between the values 120, 180 and 240 for the atomic weight of uranium is thus settled in favor of the highest number and in accordance with Mendelejeff's classification."

"Sir William Thomson bids us think of the surface of the earth as sovereigns placed on a surface of stiff jelly, and to picture the basin-like smoothly sloping dimple the coins would cause. It is plain that the stiffer the jelly is, the shallower will be the dimple and the less will be the strain. An idea of the strength and stiffness of the jelly may be gained by noting the amount and kind of yielding that results from known weights resting on it. Now, the great continents and mountain chains that rise above the general level of the earth must strain it in just this way. The earth must have a certain strength and stiffness that allows it to yield under them but not to break. We can, in fact, calculate what stresses the actual mountain chains call forth, and so get a notion of the general strength of the materials of the globe that supports them. Mr. George H. Darwin, in a paper titled 'On the Stresses Caused in the Interior of the Earth by the Weight of Continents and Mountains,' has carried out this calculation with noteworthy patience and skill."

"Submarine mines, like all other offshoots of that mighty agent, electricity, have of late developed very rapidly. They may be classed under the following heads: (1) those fired at will from the shore when a vessel is seen to be in their vicinity, generally known as observation mines; (2) those fired from the shore on the contact of the vessel with the mine, termed electro-contact mines; (3) those having no cables or connection with the shore, the firing arrangements being self-contained. Observation mines, holding as much as 500 lb. of gun-cotton, are placed at a considerable depth and can be fired with such accuracy as to explode under a vessel going at full speed. The other types of mines hold about 100 lb. of gun-cotton and are placed about 13 feet under water."

"The inaction of the committee having in charge the work of soliciting subscriptions for the foundation and pedestal for Bartholdi's Statue of Liberty has led to some impatience on the part of the French Committee of Presentation. The rumor that the latter committee were contemplating an offer of the statue to Boston has stirred the New York committee a little, but they are still debating whether to attempt to raise the money by popular subscription or to solicit a few large contributions from the wealthier citizens of the city. A member of the committee has said that various engineers have estimated the cost of a suitable base for the study at from $150,000 to $1,000,000. It is thought that an acceptable base can be made for $200,000. This sum is a mere trifle for the wealth of New York, and if the committee cannot raise it promptly, they ought in courtesy to the artist and his friends decline the proffered gift in favor of Boston or some other more deserving city."

Creating an 'eye' that will see

Lockheed knows how.

It's a project with promise that defies comprehension.

It's the NASA Space Telescope, managed by the Marshall Space Flight Center and being assembled and tested by Lockheed. It's scheduled for launch into 320-mile-high orbit by the Space Shuttle in 1985. And it will extend the observation reach of astronomers far out into the violent turmoil of space.

They'll look back about 14 billion light-years to where many think the universe edges lie. That's seven times farther than Earth-bound telescopes, hindered by atmospheric murk, have yet been able to see. And the telescope will find, lock onto, and study faint objects and masses only 1/50th as bright as any so far beheld.

Jupiter as seen from Earth (left) and through the clarity of outer space.

A masterpiece gets ready for action.

This astronomical observatory, the finest galactic telescope system ever built, will be a masterpiece of advanced technology in optics, instrumentation, communications, and pointing control.

It will be large—43 feet long, 14 feet in diameter, and weighing about 24,000 pounds. When carried aloft, it will occupy the entire cargo bay of NASA's big-airliner-size Space Shuttle.

Once the telescope is deployed, its solar panels (being built by the European Space Agency) will unroll like long window shades, facing the sun to convert light into battery-stored electric power. High-gain radio antennas also will stretch out into position to work via relay satellites in communication to and from Earth.

And then the great instrument will go to work, its on-board computing and control systems already intricately programmed for the first viewing events.

The magnificent mirrors.

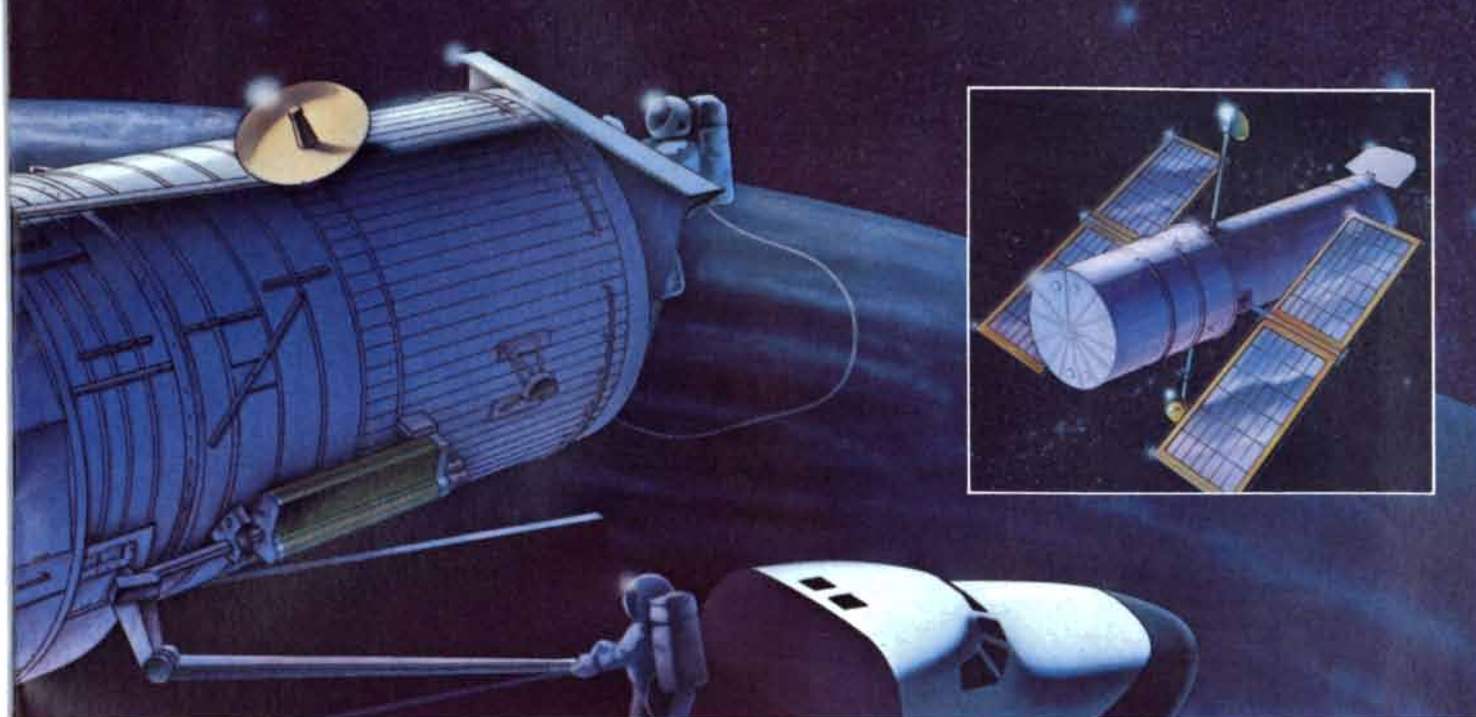
Capturing a wide spectrum of faint light (visible, red, blue, infrared, and ultraviolet) that has traveled billions of years is the job of the optical assembly (being built by Perkin-Elmer).

The extremely smooth, precisely shaped primary mirror surface will result from two full years of grinding and polishing.

A concave, foot-thick primary mirror, eight feet in diameter, will receive light coming in through the telescope's sun-shielded aperture. It will reflect and focus the light forward onto a smaller, facing mirror, which in turn will reflect the light back through a hole in the primary mirror's center. The final focal plane will be shared by a complex of guidance sensors and five scientific instruments . . . high-resolution cameras, spectrographs, and a photometer (being built by several institutions).

In their interference-free environment, the mirrors and instruments will help produce and transmit images with

the edges of the universe.



The great Space Telescope will be tucked into orbit by the Space Shuttle, which will also serve as an occasional on-orbit service center. INSET: The telescope deployed.

resolution 10 times greater than now possible with ground-based telescopes.

Fixing mankind's steadiest gaze.

Those mirrors and instruments, however, couldn't do their best without a unique Lockheed pointing control system (PCS) that positions and holds them precisely on target. And that is incorporated in the Lockheed Support Systems Module, containing the interfacing, communications, thermal control, data management, and electric power systems.

Using sensors and gyroscopic devices, the PCS will move and adjust the huge telescope into a programmed viewing attitude. The Perkin-Elmer fine-guidance sensors will fix upon a pair of known, bright 'guide stars,' references for finding the target.

When the target object is dead in the telescope's sight line, designated instruments will be turned on and other functions shut down. That's because the tiniest torque or motion in a small motor or electric relay could jar the 12-ton vehicle off its lock-on point.

And how precise is that lock-on? In angular measurement, there are 1,296,000 arc-seconds in a circle. For perhaps as long as 20 hours at a time, the telescope will hold onto a target with a stability accuracy of just *seven thousandths of one arc-second* in all directions.

That stability is 100 times greater than any ever achieved so far. It about equals standing in Boston and holding rifle sights steady on a dime suspended over Washington, D.C.!

Keeping a beauty in shape.

Lockheed's innovative design will also let the telescope have periodic 'house calls' for checkups and maintenance

In a neutral buoyancy tank, an astronaut goes through special training simulating the service tasks to be performed in space.

during its planned 15-year life. Armed with nothing more than a 7/16″ ratchet wrench and a Lockheed-devised foot-anchoring tool, a Space Shuttle astronaut will be able to open access doors, make inspections, replace modular units, and see to the telescope's well-being. When necessary, the Shuttle will retrieve the telescope and return it temporarily to Earth for general servicing.

The world of science is waiting expectantly. The wonders yet to be discovered are waiting, too. The Space Telescope will arrange a meeting.

When it comes to bringing all this together—testing, refining, integrating all the critical elements of a miraculous scientific breakthrough—Lockheed knows how.

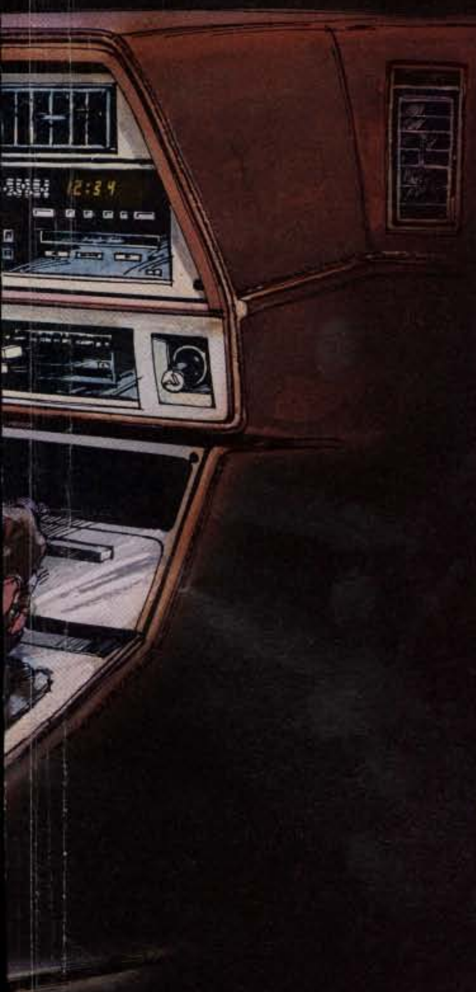
For more information about Lockheed, write for the 1981 Annual Report Highlights to A.F. Melrose, Lockheed Corporation, P.O. Box 551, Burbank, CA 91520.

Lockheed


CIMAR
CUTS THROUGH THE
BEST OF ALL...IT'S A CADILLAC.

RON '83

IMPORT MYSTIQUE.



HOW CIMARRON '83 BEATS THE IMPORTS AT THEIR OWN GAME

	CIMARRON '83	AUDI 5000S	BMW 320i	VOLVO GL	SAAB 900S SEDAN
FRONT-WHEEL DRIVE	STANDARD	STANDARD	NOT AVAILABLE	NOT AVAILABLE	STANDARD
ELECTRONICALLY FUEL-INJECTED ENGINE*	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD
FIVE-SPEED MANUAL INCLUDING OVERDRIVE	STANDARD	STANDARD	STANDARD	STANDARD 4-SPEED	STANDARD
TACHOMETER WITH FULL INSTRUMENTATION	STANDARD	EXTRA COST	STANDARD	STANDARD	STANDARD
POWER-ASSISTED RACK AND PINION STEERING	STANDARD	STANDARD	RACK AND PINION ONLY	STANDARD	STANDARD
EPA PASSENGER COMPARTMENT VOLUME	89 CU. FT.	90 CU. FT.	82 CU. FT.	89 CU. FT.	89 CU. FT.
TUNGSTEN HALOGEN FOG LAMPS	STANDARD	NOT AVAILABLE	EXTRA COST	NOT AVAILABLE	DEALER INSTALLED OPTION
ALUMINUM ALLOY WHEELS	STANDARD	STANDARD	EXTRA COST	STANDARD	STANDARD
AIR CONDITIONING	STANDARD	STANDARD	EXTRA COST	STANDARD	STANDARD
ELECTRONICALLY TUNED AM/FM STEREO RADIO	STANDARD (may be deleted for credit)	EXTRA COST	EXTRA COST	DEALER INSTALLED OPTION	DEALER INSTALLED OPTION
LEATHER-WRAPPED STEERING WHEEL	STANDARD	STANDARD	EXTRA COST	DEALER INSTALLED OPTION	NOT AVAILABLE
LEATHER SEATING AREAS	STANDARD	EXTRA COST	NOT AVAILABLE	STANDARD	NOT AVAILABLE

*Some Cadillacs are equipped with engines produced by other GM divisions, subsidiaries or affiliated companies worldwide. See your dealer for details. All imports shown are 1982 models. Level of standard equipment varies.

Bold new performance. In '83 Cimarron sports a snappy new 2.0 liter electronically fuel-injected engine with a 5-speed manual transmission. (3-speed automatic is available.) Plus an exclusively tuned Touring Suspension, front-wheel drive and more.

Bold new look. There's a strong new grille treatment and tungsten halogen fog lamps. A proud new hood medallion. And crisp, new aluminum alloy wheels.

Plus Cadillac touches... Like body-contoured, leather-faced front bucket seats with lumbar support.

Air conditioning. Full instrumentation including tach. All standard. And available features like a new electronically tuned AM/FM stereo cassette radio with 5-band graphic equalizer.

Extraordinary value. Compare Cimarron '83 to any import in its class. Feature for feature. Sticker for sticker. Most of all, compare performance with a demonstration drive. See a Cadillac dealer soon.

THE AUTHORS

LYNN R. SYKES and JACK F. EVERNDEN ("The Verification of a Comprehensive Nuclear Test Ban") are geophysicists who in addition to their scientific work have spent considerable time on means of distinguishing underground nuclear explosions from earthquakes. Sykes is Higgins Professor of Geological Sciences at Columbia University and head of the earthquake studies group at the university's Lamont-Doherty Geological Observatory. His B.S. and M.S. were both awarded in 1960 by the Massachusetts Institute of Technology. His Ph.D. in geology was granted in 1964 by Columbia and he has remained at the university since, taking up his present jobs in 1971. Sykes's scientific work has centered on plate tectonics, including the plate-tectonic structure of Asia, which may be significant in the verification of a test-ban treaty. Evernden is program manager of the U.S. Geological Survey's National Center for Earthquake Research. His B.S. in mining geology (1948) and his Ph.D. in geophysics (1951) are from the University of California at Berkeley. He writes: "I, of course, agree with all conclusions presented in this article, but that agreement should not be construed as acceptance of those conclusions by either the U.S. Geological Survey or the United States Government."

RODOLFO R. LLINÁS ("Calcium in Synaptic Transmission") is professor and chairman of the department of physiology and biophysics at the New York University Medical Center in New York. Born in Colombia, he received his M.D. from the Pontifical University of Javeriana in 1959. He then moved to Australia as a research scholar at the Australian National University, from which he got his Ph.D. in 1965. He came to the U.S. in the same year as associate professor at the University of Minnesota. From 1966 to 1970 he was on the staff of the Institute for Biomedical Research of the American Medical Association Education and Research Foundation. In 1970 he became professor of physiology and biophysics and head of the division of neurobiology at the University of Iowa; he left Iowa in 1976 to take up his present jobs at New York University. In addition to synaptic transmission Llinás has worked on the evolution of the central nervous system.

JEROME KRISTIAN and MORLEY BLOUKE ("Charge-coupled Devices in Astronomy") are respectively an astrophysicist and a solid-state physicist who share an interest in the subject of their article. Kristian is astronomer at the Mount Wilson and Las Campanas Observatories of the Carnegie Institution of Washington and the California Institute of Technology. He received his training in physics at the University of Chicago, from which he obtained his master's degree in 1956 and his doctorate in 1962. From 1962 to 1964 he was visiting lecturer and research scientist at the Center for Relativity Theory at the University of Texas at Austin. From 1964 to 1967 he was assistant professor of astronomy at the University of Wisconsin at Madison. He is a member of the group that is building the wide-field camera for the Space Telescope, which is scheduled to be launched in 1985. Kristian's main scientific interests are extragalactic astronomy and cosmology. Blouke was graduated with a B.S. from Union College in 1963. His M.S. (1965) and Ph.D. (1969) are both from the University of Illinois. Since receiving his doctorate he has worked for Texas Instruments.

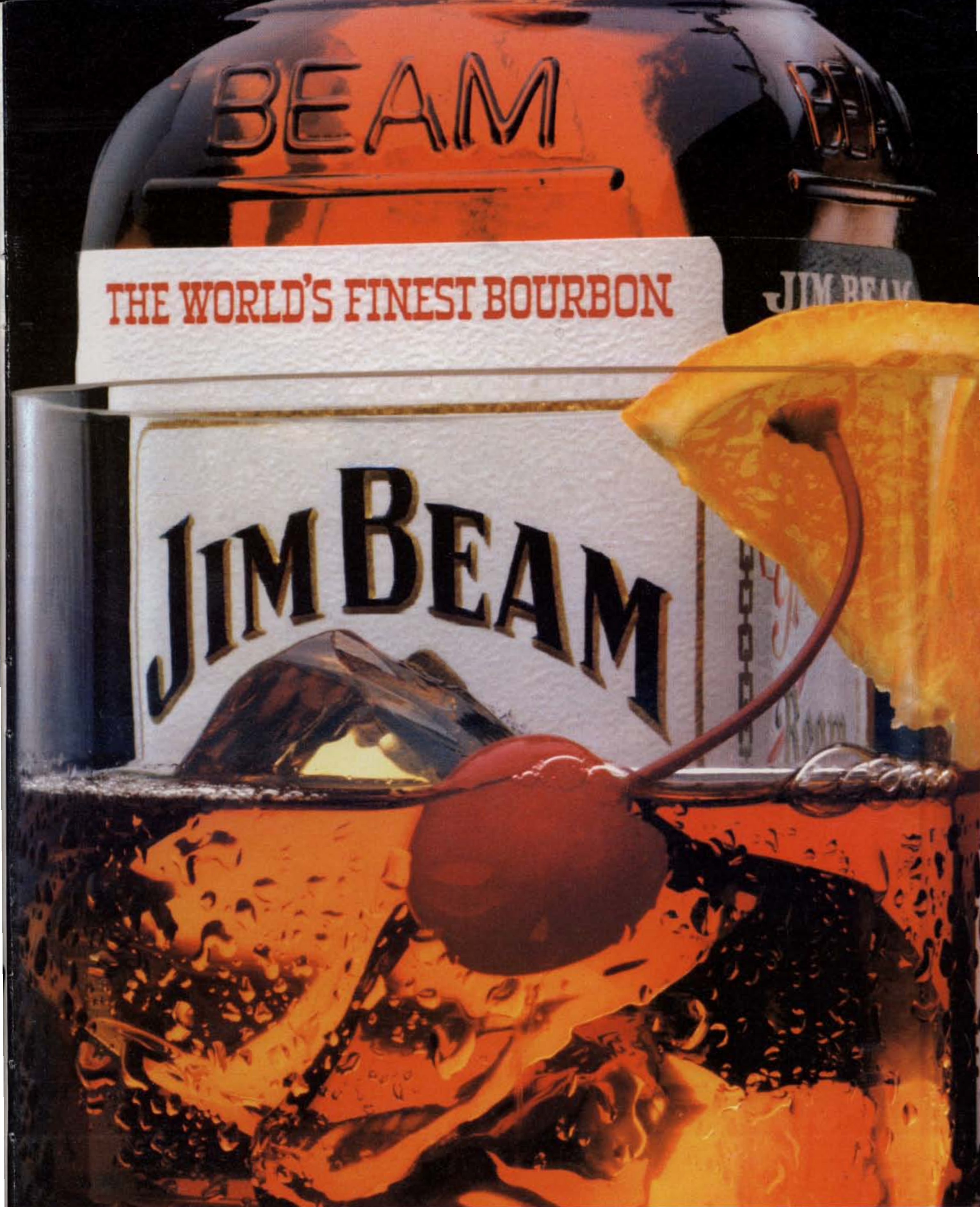
DAVID L. WALTZ ("Artificial Intelligence") is professor of electrical engineering at the University of Illinois at Urbana-Champaign and research professor at the Coordinated Science Laboratory there. He earned three degrees in electrical engineering from the Massachusetts Institute of Technology: a B.S. in 1965, an M.S. in 1968 and a Ph.D. in 1972. After a year of postdoctoral work at M.I.T. he moved to Illinois. He has been interested in the possibility of mimicking the action of the human nervous system since his days at M.I.T. He writes that his "early interests at the M.I.T. Artificial Intelligence Laboratory included the design of a touch-sensing system for robot hands." When he went to Illinois, he became interested in the utilization of existing human languages as a means of instructing computers. He writes that his "group designed and constructed an experimental English-language question-answering system called PLANES that served as a front end for a large base of Navy aircraft flight and maintenance data." Waltz adds that outside the laboratory his interests include "playing baritone saxophone (moderately well) in a concert band."

RONALD E. ROSENSWEIG ("Magnetic Fluids") is senior research associate and group leader in fluid physics at the Exxon Research and Engineering Company. His degrees are in chemical engineering; they include a bachelor's degree from the University of Cincinnati, given in 1955, and a master's degree and a doctorate, awarded in 1956 and 1959 respectively, from the Massachusetts Institute of Technology. In 1959 he joined the M.I.T. faculty as an assistant professor and remained there until 1962, when he moved to the Avco Corporation. In 1969 he became president and technical director of the Ferrofluidics Corporation. He served the company in several capacities before he went to Exxon in 1973; he remains a member of the board of directors of Ferrofluidics. Rosensweig's scientific interests include colloids, turbulent mixing and hydrodynamic stability.

DAVID K. LYNCH ("Tidal Bores") is a member of the staff of the Hughes Research Laboratories in Malibu, Calif., where he does work on optics and holography. He was graduated from Indiana University with a B.S. in 1969 and went on to obtain his Ph.D. in astronomy in 1975 from the University of Texas at Austin. Before moving to Hughes he held research positions at several institutions, including the California Institute of Technology, the University of California at Berkeley and the Sacramento Peak Observatory.

THOMAS D. SEELEY ("How Honeybees Find a Home") is assistant professor of biology at Yale University. He earned his A.B. in 1974 from Dartmouth College. His Ph.D. in biology was awarded by Harvard University in 1978. After he had received his doctorate election to the Society of Fellows at Harvard enabled him to spend a year living in the mountain jungles of Thailand, where he studied the ecology of honeybees native to the area. Since joining the Yale faculty he has concentrated on the overall strategy of foraging by honeybee colonies. Seeley writes that "the focus of my research on the highly evolved honeybee societies undoubtedly reflects my deep pleasure in observing individuals (albeit insects) living together in harmonious societies."

F. RICHARD STEPHENSON ("Historical Eclipses") is senior research fellow in physics at the University of Durham. Born in England, he got his undergraduate education at Durham. His master's degree and his doctorate in geophysics are from the University of Newcastle upon Tyne. From 1972 to 1979 he was a research fellow at Newcastle. In 1979 he moved to the University of Liverpool, where he was until this month university research fellow. The focus of his work is historical astronomical observations and their application to current problems in astronomy and geophysics. With David H. Clark he is the author of *The Historical Supernovae* (Pergamon Press, 1977) and *Applications of Early Astronomical Records* (Adam Hilger Ltd., 1978). Stephenson writes that he has a "deep sense of gratitude to the astronomers of the ancient and medieval world for the observational records they have bequeathed to us."


BEAM
THE WORLD'S FINEST BOURBON
JIM BEAM

METAMAGICAL THEMAS

Variations on a theme as the essence of imagination

by Douglas R. Hofstadter

George Bernard Shaw once wrote (in *Back to Methuselah*): "You see things; and you say 'Why?' But I dream things that never were; and I say 'Why not?'" When I first heard this aphorism, it made a lasting impression on me. To "dream things that never were"—this is not just a poetic phrase but a truth about human nature. Even the dullest of us is endowed with this strange ability to construct counterfactual worlds and to dream. Why do we have it? What sense does it make? How can one dream, or even "see," what is visibly not there?

On my table sits a Rubik's Cube. I look at it and see a $3 \times 3 \times 3$ cube whose faces turn. I see—so it seems to me—what is there. Some people, however, looked at the cube and saw things that *weren't* there. They saw cubes with shaved edges, spherical "cubes," differently colored cubes, Magic Dominoes, $2 \times 2 \times 2$ cubes, $4 \times 4 \times 4$ and higher-order cubes, skew-twisting cubes, pyramids, octahedrons, dodecahedrons, icosahedrons, four-dimensional polyhedrons. And the list is not complete yet! Indeed, it is impossible to imagine closing the book on this rich idea.

How did it come about? How is it that, in looking directly at something solid and real on a table, people can see far beyond that solidity and reality, can see an "essence," a "core," a "theme" on which to devise variations? I must stress that the solid cube itself is not the theme (although it is convenient and easy to speak as if it were). In the mind of each person who perceives a Rubik's Cube there arises a *concept* we could call Rubik's cubicity. It is not the same concept in each mind, just as not everyone has the same concept of asparagus or Beethoven. The variations that are spun off by a given cube inventor are variations on that concept. In a discussion of perception and invention the distinction between an object and the concept of the object in someone's mind is crucial.

Now, when Sally Cubelover comes up with a new variation, let us say the $4 \times 4 \times 4$, is it as a result of racking her brain, trying as hard as she can to "go against the grain" in order to come up with something original? Does she think to herself, "Golly, Rubik must have really exerted himself to come up with this totally new idea; therefore I too must strain my mind to its limits in order to invent something original"? Surely not. Einstein didn't go around racking his brain, muttering to himself, "How can I come up with a great idea?" Like Einstein (although on a lesser scale), Sally never needs to ask herself, "H'm, let's see, shall I try to figure out some way to spin off a variation on this object sitting here in front of me?" No, she just does what comes naturally.

The bottom line is that invention is much more like falling off a log than like sawing one in two. In spite of Thomas Edison's memorable remark, "Genius is 1 percent inspiration and 99 percent perspiration," we are not all going to become geniuses simply by sweating more or resolving to try harder. A mind follows the path of least resistance, and it is when it feels easiest that it is probably being its most creative. Or, as Mozart used to say, things should "flow like oil"—and Mozart ought to know. Trying harder is not the name of the game; the trick is getting the right concept to begin with, so that making variations on it is like taking candy from a baby.

Uh-oh—I've let the cat out of the bag. Let me, then, straightforwardly state the thesis I shall now elaborate: Making variations on a theme is really the crux of creativity.

On the face of it the thesis is crazy. How can it possibly be true? Aren't variations simply derivative notions, never truly original creations? Isn't the notion of a $4 \times 4 \times 4$ cube simply a result of "twiddling a knob" on the concept of Rubik's cubicity? You merely twist the knob from its "factory setting" of 3 to the new setting of 4, and presto—you've got it! An inner voice protests. That's just too easy. That's certainly not where relativity or Rubik's Cube came from, is it? Isn't there a "magic spark" that leaps across a gap when an Einstein or even a Rubik comes up with a great idea, something that is patently lacking when Sally Cubelover twiddles a knob on the already existing notion of Rubik's Cube?

To be sure, inventing the notion of a $4 \times 4 \times 4$ cube is far less deep than coming up with special or general relativity. This does not mean, however, that the underlying mental processes are necessarily based on totally different principles. Of course, there is an obvious sense in which the underlying mental processes in your brain, my brain, Sally's brain and Einstein's brain are all "the same": they all depend on the neural hardware. But it is not this microscopic, biological level I mean when I suggest that the underlying mental processes in different brains are somehow the same. What I mean is that there are mechanisms, processes, call them what you will, that can be described functionally, without reference to the neural substrate enabling them to take place in brains.

Hence a notion such as twiddling a knob on a concept bears no relation to the activities of neurons in the brain, or at least no obvious relation. Well then, is there any reality to it, or is it just a metaphor? If someday we come to understand the brain, will we then be confident that we are on solid ground when we speak of a brain literally *containing concepts?* Or will such statements forever remain shaky and metaphorical *façons de parler* compared with such hard science facts as "At the back of each human brain there is a cerebellum"? Well, until words such as "concept" have become as scientifically legitimate as, say, "temperature," we will not have come anywhere close to understanding the brain—at least not in my book.

It must be admitted that at present words such as "concept" are only metaphorical. They are protoscientific terms awaiting explication. This, however, is an excellent reason to try to flesh them out as much as possible, to try to see what the metaphor of twiddling knobs on a concept involves. Pinning down the meaning of such a metaphor will help us to know much more clearly what we would ideally want from a "hard science" explanation of the brain.

This metaphor makes your imagination conjure up a vision of a tangible thing called a concept that literally has some kind of knobs on it, waiting to be twiddled. What I picture in my mind's eye is something that, instead of being built out of millions of neurons, is more like a metallic "black box" with a panel on it bearing a row of plastic knobs whose little pointers tell you what each one's setting is.

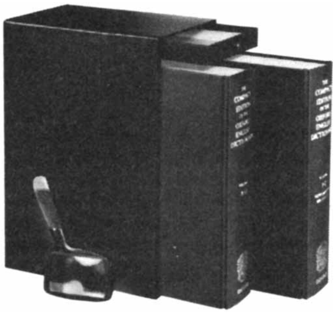
To make this image more concrete, let me describe a genuine example of such a black box with knobs. Back in the days of player pianos good pianists

Variations on a theme: 56 variations on the letter "A" from a recent catalogue of Letraset USA Inc.

Or.

Choose one of the 5 valuable sets below as your introduction to membership in the Book-of-the-Month Club.

You simply agree to buy 4 books within the next two years.



The Compact Edition of The Oxford English Dictionary

for **$24.95** (Publisher's price $125)

The "most complete, most scholarly dictionary of the English language"—*The Christian Science Monitor.* Through photo-reduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition.* Magnifying glass included.

The Story of Civilization by Will and Ariel Durant

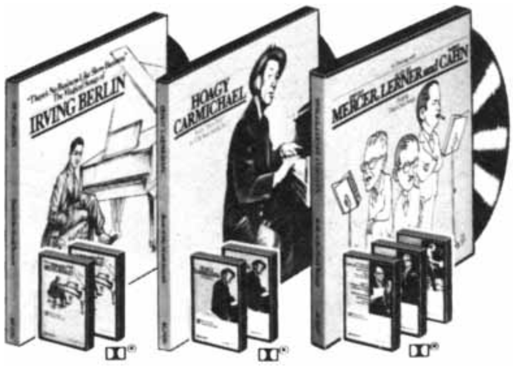
for **$29.95** (Publisher's prices total $307.45)

For almost half a century Will and Ariel Durant traced the continuity of world history to show the foundations of society today. The Durants' 11-volume illustrated masterwork is history come alive in all its dimensions. The enormous scope of the work covers ancient and recent civilization, including Oriental as well as Western history. A Book-of-the-Month Club exclusive for close to 50 years, it is both a reference work of undisputed authority and a lively portrait of the rulers and leaders who have shaped the course of human events.

The Encyclopedia of Philosophy

for **$24.95** (Publisher's price $150)

The most comprehensive encyclopedia of philosophy ever published, this set traces ancient, medieval, modern, Eastern and Western thought. An essential and rewarding reference source for home libraries.



Great American Songwriters

for **$14.95** (Nonmember's prices total $74.85)

The best of American musical theater and popular music in one unique collection. The genius of Irving Berlin, Hoagy Carmichael and Johnny Mercer, Alan Jay Lerner, Sammy Cahn. In ten records, seven cassettes or seven 8-tracks.

The Decline and Fall of the Roman Empire by Edward Gibbon. Edited by J. B. Bury

for **$24.95,** 7 Vols. (Pub price $300)

The definitive Bury edition of the most acclaimed history of all. Illustrated 7-volume set, newly available on long-lasting, acid-free paper, quarter-bound in leather.

Book-of-the-Month Club, Inc., Camp Hill, Pennsylvania 17012 **A170-10**

Please enroll me as a member of Book-of-the Month Club and send me *either* the four books I've listed in the "Either" boxes at right, billing me $2, plus shipping and handling charges, *or* the one set I've checked in the "Or" column, billing me for the appropriate amount, plus shipping and handling charges. In either case, I agree to buy four additional books from the Club over the next two years. A shipping and handling charge is added to each shipment.

Mr.
Mrs.__________
Miss (Please print plainly)

Address__________ Apt._____

City__________

State__________ Zip_____

Prices shown are U.S. prices. Outside the U.S., prices are generally higher.

BOOK-OF-THE-MONTH CLUB®

America's Bookstore since 1926.

Either.

Indicate by number the four books you want

☐
☐
☐
☐

2-04

Or.

Check one box only

☐ 912. Compact OED **$24.95**
☐ 913. The Story of Civ. **$29.95**
☐ 917. Ency. of Philosophy **$24.95**
☐ 948. Great American Songwriters (10 Records) **$14.95**
☐ 949. Great American Songwriters (7 8-Track Tapes) **$14.95**
☐ 950. Great American Songwriters (7 Cassettes) **$14.95**
☐ 951. Decline and Fall of the Roman Empire **$24.95**

2-64

made piano rolls of all kinds of wonderful music. Nowadays you can buy phonograph records of those rolls being played back on player pianos, but you can do better than that. Many of the best rolls (made on a special kind of piano called a Vorsetzer) have been converted into digital cassette tapes, not tapes to be put into a tape recorder but tapes to be played on a piano equipped with a device called a Pianocorder. The Pianocorder "reads" the magnetic tape and converts it into instructions to the keyboard and pedals, so that your own piano then plays the piece. Each Pianocorder has a black box on the front of which is a control panel with a row of three knobs ("Tempo," "Pianissimo" and "Fortissimo") and one switch ("Soft pedal"). By twisting the "Tempo" knob you can make Rachmaninoff speed up; by twiddling the "Pianissimo" and "Fortissimo" knobs you can make Horowitz play more softly or Rubinstein more loudly. It is too bad there is not a knob labeled "Pianist" so that you could select who plays. It would be interesting to change in midstream from one pianist to another.

This device takes us one step toward realizing a dream of the Canadian pianist Glenn Gould. Gould is very much at one with the electronic age, and for years he has been advocating the use of computers to enable people to control the music they hear. You begin with an ordinary recording of, say, Gould himself playing a concerto by Mozart. This is merely raw data for you to manipulate. On your space-age record player you have a bunch of knobs that enable you to slow the music down or to speed it up ad libitum, to control the volume of each separate section of the orchestra, even to correct for high notes played flat by the violinists. In effect you become the conductor, with knobs to control every aspect of the performance dynamically. The fact that it was originally Gould at the piano is, by the time you are done with it, irrelevant. By now you have taken over and made the performance your own. Presumably such systems would eventually evolve to the point where you could start with the written score, dispensing entirely with the acoustic recording stage.

Why not carry this farther, then? If we are allowing ourselves to fantasize, why not go as far as we can imagine? Why should our "raw data" be limited to the finite universe of existing pieces? Why should there not be a knob to control the mood of the composition and another to control the composer whose style it is to be written in? This way we could get a new piece by our favorite composer in any desired mood. But that is too conservative. Why should we be limited to the finite universe of composers already born? Why could there not be a knob to enable us to interpolate between composers, making it possible for us to tune our music-making machine to an even mixture of Johann Sebastian Bach, Giuseppe Verdi and John Philip Sousa (ugh!), or to a position halfway between Franz Schubert and the Sex Pistols (super-ugh!)? And why stop at interpolation? Why not extrapolate beyond a given composer? For instance, I might want to hear a piece by "the composer who is to Ravel as Ravel is to Chopin." The machine would merely need to calculate the ratios of its knob settings for Ravel and Chopin and then multiply the Ravel settings by those same ratios to come up with a super-Ravel.

The LORD is my shepherd;
I shall not want.
He maketh me to lie down
in green pastures:
he leadeth me
beside the still waters.
He restoreth my soul:
he leadeth me
in the paths of righteousness
for his name's sake.
Yea, though I walk through the valley
of the shadow of death,
I will fear no evil:
for thou art with me;
thy rod and thy staff
they comfort me.
Thou preparest a table before me
in the presence of mine enemies:
thou anointest my head with oil,
my cup runneth over.
Surely goodness and mercy
shall follow me
all the days of my life:
and I will dwell
in the house of the LORD
for ever.

Every letter in Donald E. Knuth's Metafont rendition of Psalm 23 is in a different font

It is really no trickier than solving any trivial analogy problem: "What is to a triangle as a triangle is to a square?" "What is to Greece as the Falkland Islands are to Britain?" "What is to a water bed as ice is to water?" and other "easy" problems like that. The truth is, of course, quite the contrary: analogy problems are extremely tricky to mechanize. The knobs on most concepts are not so apparent that we can just read their settings right off. The examples above simply carried a thought to a ludicrous extreme. It is nonetheless worth while to look seriously at the idea that a concept can be considered as a machine whose knobs can be twiddled to yield a fabulous array of variations.

The Rubik's Cube concept with its "order" knob set at 3 gives rise to an ordinary $3 \times 3 \times 3$ cube, and with that knob set at 4 a $4 \times 4 \times 4$ cube. Come to think of it, doesn't there have to be a separate knob for each dimension so that you can twiddle each one independently of the others? After all, not all variations have to be cubical. The Magic Domino is $3 \times 3 \times 2$. Hence if we agree that there are *three* knobs defining the shape, then in the original cube they all just accidentally happened to have the same setting. Now, given these three knobs we can use our concept—our knobbed machine—to generate such mental objects as a $7 \times 7 \times 7$ Rubik's Cube, a $2 \times 2 \times 8$ Magic Domino, even a $3 \times 5 \times 9$ Rubik's Magic Brick (or, if you will excuse me, a Rubrick).

But wait a minute. If there really are only three knobs, we are locked into three dimensions. Obviously we do not want *that*. Then let us add a fourth knob to control the length in the fourth dimension. With this knob we can now make a four-dimensional $2 \times 3 \times 5 \times 7$ Rubrick and in addition any Rubik's Tesseract we might want. But needless to say, once we have gone through the gate from three dimensions to four we should certainly expect to be able to go farther. For any n we could imagine n-dimensional Rubik's objects, for example a $2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ Hyper-Rubrick. But something peculiar has happened. We must now conceive of our machine—our concept—as having a potentially *unlimited* number of knobs on it (one for each dimension in n-dimensional space). If n is set at 3, there need be only three more knobs. But if n is 100, we need 100 extra knobs!

No real machine has a variable number of knobs. This may sound like a somewhat trivial observation, but it leads into some tricky waters. The point is that if we want to keep on using the metaphor of a concept as a machine with knobs on it, we have to stretch the very concept of "knob." New knobs must be able to materialize, depending on the settings of other knobs. Or you can think of it this way, if you like: On each concept there are potentially an infinite number of knobs, and at any moment some new knobs may be revealed as a result of the settings of other knobs.

I am not sure I like that view, however. It is too cut and dried, too closed and predetermined for my taste. I am more in favor of a view holding that the "knobs" on any one concept depend on the set of concepts that happen to be active simultaneously in the mind of the person. This way new knobs can spring into existence seemingly out of nowhere; they do not all have to be present from the beginning in the isolated concept. If we go back to Rubik, it would mean that *his* concept of Rubik's Cube did not (and still does not) explicitly—or even implicitly—incorporate all the possible variations people may come up with. Rubik anticipated, and even designed, many of the objects that have subsequently appeared and that we perceive as variations on a theme, but his mind did not exhaust that fertile theme. Once the concept entered the public domain it began to develop in ways Rubik never could have anticipated.

There is a way concepts have of slipping from one into another, following a quite unpredictable path. This slippage affords us perhaps our deepest visions into the hidden nature of our conceptual networks. Sometimes the slippage is totally accidental, as it is when we make a typographical error or a grammatical mistake, choose the wrong word, create a malapropism or concoct a phrase out of other phrases. Sometimes it is nonaccidental but comes straight out of our unconscious mind. By "nonaccidental" I do not mean to imply it is deliberate. It is not that we say to ourselves, "I think I shall now slip from one concept into a variation of it," since this kind of deliberate conscious slippage is most often quite uninspired and infertile. "How to think" and "how to be creative" books (even thoughtful ones such as George Pólya's *How to Solve It*) are, for that reason, of little use to the would-be genius.

Strange though it may sound, nondeliberate yet nonaccidental slippage permeates our thought processes and is, I believe, the very core of thinking. This subconscious manufacture of subjunctive variations on a theme is something that goes on day and night in each of us, usually without our slightest awareness of it. It is one of those things that like air or gravity or three-dimensionality tend to elude our perception because they define the fabric of our lives.

To make this concrete let me contrast an example of "deliberate" slippage with an example of "nondeliberate but nonaccidental" slippage. Imagine that one summer evening you and Sally Cubelover have just walked into a surprisingly crowded coffeehouse. Now go ahead and manufacture a few variants on that scene, in whatever ways you want. What kinds of things do you come up with when you deliberately "slip" this scene into variants of itself?

If you are like most people, you will come up with some pretty obvious slippages, made by moving along what seem to be the most obvious "axes of slippability." Typical examples are:

It could have been a winter evening instead of a summer one.

You could have come with Adam Spherehater instead of Sally Cubelover.

You could have gone to a Chinese restaurant instead of a coffeehouse.

The coffeehouse could have been almost empty.

Now contrast your variations with one I overheard one summer evening in a crowded coffeehouse when a man walked in with a woman. He said to her: "I'm glad I'm not a waitress here tonight." This is a perfect example of a subjunctive variation on the given theme, but unlike yours it emerged spontaneously and for the purpose of communication. The list above looks positively mundane next to this casual remark. And the remark was not considered to be particularly clever or ingenious by his companion. She merely agreed with the thought by saying "Yeah." It caught my attention less because I thought it was clever than because I am always on the lookout for interesting examples of slippability.

I found this example not just mildly interesting but highly provocative. If you try to analyze it, it would appear at first to force you as a listener to imagine a sex-change operation done in record time. But when you simply *understand* the remark, you see that in reality there was no intention in the speaker's mind of bringing up such a bizarre image. His remark was much more figurative, much more abstract. It was based on an instantaneous perception of the situation, a kind of there-but-for-the-grace-of-God-go-I reaction, that induces a quick flash to the effect of "Simply because I am human I can place myself in the shoes of that waitress; therefore I *could have been* that waitress." Logical or not, this is the way our thoughts go.

Thus when you look carefully, you see that this particular thought has practically nothing to do with the speaker, or even with the waitresses he sees. It is just his flip way of saying, "Boy, it sure is crowded here tonight." And that, of course, is why nobody really is thrown for a loop by such a remark. Yet the remark was made in such a way that it invites you to lightly "map" him onto a waitress, just barely noticing (if you notice it at all) that there is a sex difference. What an amazingly subtle thought process is involved here! And what is even more amazing (and frustrating) to me is how hard it is to point out to people how amazing it is. People find it hard indeed to see what is amazing about the ordinary behavior of people. They cannot quite imagine how it might have been otherwise. It is hard to slip mentally into a world in which people would *not* think by slipping mentally into other worlds, very hard to create a counterfactual world in which counterfactuals were not a key ingredient of thought.

Here, briefly, is another example: I was having a conversation with someone who told me he came from a town in Indiana named Whiting. Since I did not know where the town was, he said it was near Chicago and then added, "Whiting would be in Illinois if it weren't for the state line." Again the remark was made casually; it was certainly not an attempt to be witty. He didn't chuckle, nor did I. I simply smiled, signaling my understanding of his meaning, and then we went on. But try to analyze what this remark means! On a logical level it is somewhat like a tautology. Whiting would undoubtedly be in Illinois if the Illinois state line made it so, but then if we are letting the border of Illinois slip, what is to prevent it from enclosing Toronto or even Peking? But psychologically the remark is quite sensible, relying implicitly on some shared intuition about the impermanence and arbitrariness of geographic boundary lines, an intuition about how state lines could indeed slip. It wouldn't ever occur to us—except in discourse such as the present one—to slip the Illinois state line around Peking. Yet it *did* occur spontaneously in that man's mind. He was thus revealing some deep qualities of his mental representation of Whiting.

Remarks such as this one betray the hidden "fault lines" of the mind; they show which things can slip and which cannot. And yet they also reveal that nothing is reliably unslippable. Context contributes an unexpected quality to the knobs that are perceived on a given concept. The knobs are not displayed on a neat little control panel, forever unchangeable. Instead changing the context is like taking a tour around the concept, and as you get to see it from various angles more and more of its knobs are revealed. Some people get to be good at perceiving fresh knobs on concepts where others thought there were none, just as some people get to be good at seeing mushrooms in a forest where others see none.

It may still be tempting to think that for each well-defined concept there must be an "ultimate" or "definitive" set of knobs such that the abstract space traced out by all possible combinations of the knobs yields all possible instantiations of the concept. A case in point is the concept of the letter *A*. The typographically naive might think that here there would be only four or five knobs to twiddle. The more you look into letterforms, however, the more elusive any attempt to define them mathematically becomes. One of the most valiant efforts at "knobbifying" the alphabet has been the letterform-defining system called Metafont, developed at Stanford by the computer scientist Donald E. Knuth.

Knuth's purpose is not to arrive at an ultimate mathematical definition of the letters of the alphabet (I suspect he would laugh at the very notion) but to allow a user to create "knobbed" letters; we could call them letter schemas. This means you can choose for yourself what the variable aspects of a letter are, and then, with the aid of Metafont, you can easily devise knobs allowing those aspects to vary. That would include just about anything you could think of: the length of a line in a letter, widenings or taperings of lines, the shape of a curve, the presence or absence of serifs and so on. The full power of the computer is then at your disposal; you can twiddle away to your heart's content, and the computer will generate all the products your knob settings define.

Going further than dealing with letters in isolation, Knuth allowed letters to *share* parameters. That is, a single "master knob" can control a feature common to a group of related letters. Then, although there may be hundreds of knobs when you count the knobs on all the control panels of all the letters of the alphabet, there will be far fewer master knobs that have a deep and pervasive influence on the entire alphabet. What happens, in effect, is that by twiddling the master knobs alone you have a way of drifting smoothly through a "space" of typefaces.

Perhaps Knuth's greatest virtuoso feat with Metafont is what he did with Psalm 23, which in English consists of 593 characters (including spaces). Knuth had defined a full set of characters that shared 28 master knobs. He began his printed version of the psalm with all 28 knobs at their leftmost settings. Then, character by character, he inched his way toward the rightmost settings, turning each knob 1/592 of the way, so that by the time he reached the final character the extreme opposite end of the spectrum had been reached. In one sense every letter in the psalm is in a different typeface. And yet the transition is so smooth that it is locally undetectable. This example is drawn from Knuth's inspiring article "The Concept of a Meta-Font" (*Visible Language*, Vol. 16, No. 1, pages 3–27; Winter, 1982).

One of Knuth's main theses is that with computers we are in the position of being able to describe not just a thing in itself but *how that thing would vary*. Metafont epitomizes this thesis. In a sense the computer, rather than simply blindly reproducing fixed letterforms, has a crude "understanding" of what it is drawing, created by the designer who "knobbified" the letters. And yet one should be careful not to fall under the illusion, easily created by Metafont's extraordinary power, that these 28 master knobs—or any finite set of knobs—actually span the entire space of all possible typefaces. This is about as far from the truth as would be the assertion that the set of all possible types of human faces could be captured in a computer program with 28 knobs.

Even the space of all versions of the letter *A* is only barely explored when you twiddle *all* the knobs in Knuth's representation of *A*, not just the 28 master knobs it shares with other letters but the many "private" knobs it has as well. Even 1,000 knobs would not suffice to cover the variety of *A*'s people can recognize easily. Some examples of the richness of the full space of *A*'s are given in the illustration on page 21.

There is a crucial distinction to be made here. A machine with one off-on switch (the most trivial kind of knob) for each square in a 200×200 grid will certainly define any of the *A*'s shown, but it will not exclude *B*'s, *Z*'s or pictures of your grandmother. It is another matter entirely to define a set of knobs whose twiddling covers all the *A*'s shown, all the interpolations between them (as well as extrapolations in all possible directions), yet never leads you out of the space of recognizable *A*'s. This is far trickier! Similarly, it is a nearly trivial project to write a computer program that in principle writes all possible sequences and combinations of tones in all possible rhythmic patterns, but it is a far cry from writing a program that produces only pieces in the style of Bach. Putting on the constraints makes the program unutterably more complex.

What Metafont gives you, rather than the full space of the *A*'s in all type faces, is a *subspace*, and such a tightly related one that it is perhaps best to call it a *family*. Nobody would be able to predict the existence of butterflies from having studied only ants, wasps and beetles. Likewise nobody would be able to predict the full magnitude of the *concept* of *A* from seeing the family traced out by the finite number of knobs in any realistic Metafont program for *A*.

The next stage beyond Metafont will be a program that on its own can extract a set of knobs from a set of given input letters. This, however, is a program for the distant future. At present it takes a highly trained and perceptive type designer months to convert a set of letterforms into Metafont programs with knobs flexible enough to warrant the trouble taken. It would be relatively easy to do it in some crude mechanical way, but what one wants is for stylistic unity to be preserved even as the master knobs are twiddled. Therefore the task of mechanizing the production of Metafont programs amounts to the mechanization of artistic perception. It is hardly around the corner.

There is a curious work called *One Book Five Ways*, published in 1978 by William Kaufmann, Inc. It came about as follows. As an educational experiment in comparative publishing procedures, a manuscript on indoor gardening was sent to five different university presses. The presses all cooperated in coming up with full publication versions of the book, which turned out to be stunningly different at all conceivable levels. William Kaufmann had the bright idea of publishing pieces of the various versions side by side; what resulted was this elegant "metabook." It brings home the meaning of the old saying that there is more than one way to skin a cat.

Making the book was an extravagant foray into "possible worlds," the kind of thing that seems very hard to do. One of Knuth's points, however, is that as computers become commoner and more sophisticated the notion of skinning a cat in 10 different ways will gradually become less extravagant. Once your "cat" has been represented inside a powerful computer program it is no longer just one cat; it is a "cat schema," a mold for many cats at once, and you can skin them all differently (or at least until the cat schema runs out of lives).

Text formatting and computer typesetting present us easily with many alternative versions of a piece of text. Metafont shows us how letterforms can glide into alternative versions of themselves. It is now up to us to continue this trend of extending our abilities to see farther into the space of possibilities surrounding what *is*. We should use the power of computers to aid us in seeing the full concept—the implicit "sphere of hypothetical variations"—surrounding any static, frozen perception.

I have concocted a name for this imaginary sphere: I call it the "implicosphere," which stands for "implicit counterfactual sphere," referring to things that never were but that we cannot help seeing anyway. (The word can also be taken as referring to the sphere of implications surrounding any given idea.) If we want to enlist computers as our partners in this venture of inventing variations on a theme, which is to say turning implicospheres into explicospheres, we must give them the ability to spot knobs themselves, not just to accept knobs we human beings have spotted. To do this we will have to look deeply into the nature of slippability, into the fine-grained structure of those networks of concepts in human minds.

One way to imagine how slippability might be realized in the mind is to suppose that each new concept begins life as a compound of previous concepts and that from the slippability of those concepts it inherits a certain amount of slippability. That is, since any of its constituents can slip in various ways, modes of slippage are induced in the whole. Generally letting a constituent concept slip in its simplest ways is enough, since when more than one slippage comes at a time, it can already create many unexpected effects. Gradually, as the space of possibilities of the new concept—the implicosphere—is traced out, the commonest and most useful of the slippages become more closely and directly associated with the new concept itself rather than having to be derived repeatedly from its constituents. In this way the new concept's implicosphere becomes more and more explicitly explored. Eventually the new concept becomes old, and it reaches the point where it too can be used as a constituent of a fresh, new, young concept.

Some examples of this kind of thing were presented in "Metamagical Themas" for September, 1981. Although September is almost October and 1981 is almost 1982, you may not have those examples at your mind's fingertips, or on the tip of your mind's tongue. Let me give a few more examples of slippage of a new notion based on slipping some of its parts in their simplest ways. The notion I have chosen is the one of you yourself sitting there reading this very column at this very moment. Here are some elements of the implicosphere of that concept:

You are almost reading the September 1981 issue of *Scientific American*.

You are almost reading a piece by Richard Hofstadter, the historian.

You are almost reading a column by Martin Gardner.

You are almost reading this column in French.

You are almost reading my book *Gödel, Escher, Bach: an Eternal Golden Braid.*

You are almost writing this column.

I am almost talking to you.

By now the original concept is almost lost in a sea of "almost" variations, but it has been enriched by the exploration, and when you come back to it, it will have been that much more reified as a stand-alone concept, a single entity rather than a compound one. After a while, under the proper triggering circumstances, this very example may be retrieved from memory as naturally and effortlessly as the concept of "fish" is.

This is an important idea: the test of whether a concept has really come into its own, the test of its genuine mental existence, is its retrievability by that process of unconscious recall. That is what lets you know the concept has been firmly planted in the soil of your mind. It is not whether the concept appears to be "atomic," in the sense that you have a single word to express it by. That is far too superficial.

Here is an example to illustrate why. A friend told me recently that the first edition of *Encyclopaedia Britannica* consisted of three volumes: Volume 1 was *A* through *B*, Volume 2 was *C* through *L* and Volume 3 covered the rest of the alphabet. *A* was given 511 pages and *M* through *Z* were given 753 pages altogether! This amusing fact instantaneously triggered the retrieval of another memory, implanted in my mind years ago under totally unremembered circumstances, of how records used to be made back in the days when there was no magnetic tape and the master disk was actually cut during the live performance. The performers would be singing or playing along and all of a sudden the recording engineer would notice that there was not much room left on the disk, and so the performers would be given a signal to hurry up. As a result the tempo would be faster the closer to the center of the disk the needle got. I think it is obvious why the one concept triggered the retrieval of the other. But then again, is it really obvious?

On the surface the two concepts are completely unrelated. One concerns printed matter, books, the alphabet and so on; the other concerns wax disks, sounds, performers, recording techniques and so on. At some deeper conceptual level, however, these really *are* the same idea. There is just one idea here, and this idea I call a conceptual skeleton. Try to verbalize it. It is cer-

Why an ear, nose and throat man went into radishes, lettuce and peppers.

It began small, a quarter century ago.

In a back room of his house, Dr. Maynard Murray, a Chicago physician, started a hydroponic garden.

Instead of growing vegetables in soil, he grew them in a solution of water and minerals. And produced enough to feed a family of four.

Encouraged, Dr. Murray moved to Florida and went into hydroponic farming commercially.

Today, he has 178 beds, each 100 feet long and a yard across, growing vegetables in water.

Pumps developed by the people of ITT Bell & Gossett flood the beds twice a day using minerals, like "sea solids," dissolved in the water to feed the vegetables.

This new kind of farming is virtually weed-free.

It's also economical—and oblivious to weather.

Of course, hydroponic farming requires dependable, trouble-free pumps. (Dr. Murray is still using the same ITT pumps he started with.)

At ITT, our pumps are used in homes and industry, for everything from air conditioning to etching computer chips.

But we feel a particular pride in helping a business like Dr. Murray's—something that's just beginning to grow.

The best ideas are the ideas that help people. **ITT**

tainly not just one word. It will take you a while. And when you do come up with a phrase, the chances are it will be awkward and stilted—and still not quite right!

Both of the cited instances of this conceptual skeleton—in itself nameless, majestically nonverbalizable—are floating about in the implicosphere that surrounds it, along with numerous other examples I am unaware of, not yet having twiddled enough knobs on that concept. I do not, of course, even know which knobs it has, but I may eventually find out. The point is that the concept itself has been reified; this much is proved by the fact that it acts as a point of immediate reference, that under the proper circumstances my memory mechanisms are capable of accessing it directly. The vast majority of our concepts are wordless in this way, although we can certainly make a stab at verbalizing them when we need to.

Early in this column I stated a thesis: that the crux of creativity lies in the ability to manufacture variations on a theme. I hope now to have sufficiently fleshed out this thesis for you to understand the full richness of what I meant when I said "variations on a theme." The notion encompasses knobs, parameters, slippability, counterfactual conditionals, subjunctives, "almost" situations, implicospheres, conceptual skeletons, mental reification, memory retrieval—and more.

The question may persist in your mind: Aren't variations on a theme somehow trivial compared with the invention of the theme itself? This leads one back to the seductive notion that Einstein and other mighty creators are cut from a different cloth from the one used to make ordinary mortals, or at least that certain cognitive acts of such creators involve principles transcending the everyday ones. This is something I do not believe at all. If you look into the history of science, for instance, you will see that every idea is built on a thousand related ideas. Careful analysis leads one to see that what we choose to call a new theme is itself always some kind of variation, on a deep level, of earlier themes.

Newton said that if he saw farther than others, it was only because he stood on the shoulders of giants. Too often, however, we simply indulge in wishful thinking when we imagine that a clever or beautiful idea was somehow due to unanalyzable, magical, transcendent insight rather than to any mechanisms, as if all mechanisms by their very nature are necessarily shallow and mundane.

My own mental image of the creative process involves viewing the organization of a mind as consisting of thousands, if not millions, of overlapping and intermingling implicospheres, at the center of each of which is a conceptual skeleton. The implicosphere is a flickering, ephemeral thing, rather like the electron cloud, with its quantum-mechanical elusiveness, around an atomic nucleus. If you have studied quantum chemistry, you know that the fluid nature of chemical bonds can best be understood as a consequence of the curious quantum-mechanical overlap of electronic wave functions in space, wave functions belonging to electrons orbiting neighboring nuclei. In a metaphorically similar way, it seems to me, the crazy and unexpected associations allowing creative insights to pop seemingly out of nowhere may well be consequences of a similar chemistry of concepts with its own special types of "bonds" that emerge out of an underlying "neuron mechanics."

The novelist Arthur Koestler has long been a champion of a mystical view of human creativity, advocating occult views of the mind while at the same time eloquently and objectively describing its workings. In his book *The Act of Creation* he presents a theory of creativity whose key concept he calls bisociation: the simultaneous activation and interaction of two previously unconnected concepts. That view emphasizes the coming together of *two* concepts, bypassing discussion of the internal structure of a *single* concept. This is in keeping with Koestler's philosophy of believing wholes are somehow greater than the sum of their parts.

In contrast, I have been emphasizing the idea of the internal structure of one concept. In my view the way concepts can bond together and form conceptual molecules at all levels of complexity is a consequence of their internal structure. The crux of the matter is the internal structure of a single concept and how the concept "reaches out" toward things it is not. I am not one to believe wholes elude description in terms of their parts. If we come to understand the "physics of concepts," then perhaps we can derive from it a "chemistry of creativity," just as we can derive the principles of chemistry from those of physics. But again it is not just around the corner. Alan Turing's words of cautious enthusiasm about artificial intelligence remain as apt as they were in 1950, when he wrote them in concluding his famous article "Computing Machinery and Intelligence": "We can only see a short distance ahead, but we can see plenty there that needs to be done."

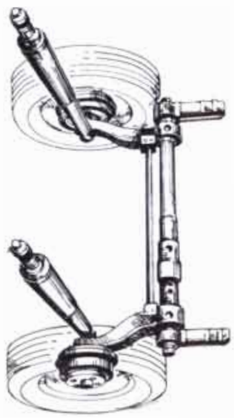
Recently I happened to read a headline on the cover of a popular electronics magazine that blared something to the effect of CHIPS THAT SEE. Bosh! I'll start believing in chips that see as soon as they start seeing things that never were and asking "Why not?"

RENAULT ALLIANCE.

NOW THERE'S AN AFFORDABLE EUROPEAN SEDAN WITH ELECTRONIC FUEL INJECTION, TWIN AXIAL REAR TORSION BARS, FOUR-WHEEL INDEPENDENT SUSPENSION AND...

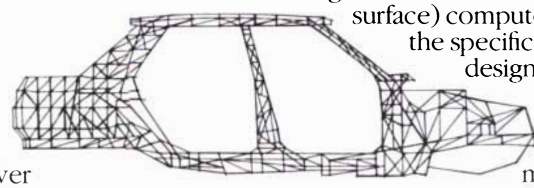
European performance and handling.
The new Renault Alliance DL. Powered by an aluminum head 1.4 litre engine with Bendix single-point fuel injection. An on-board computer monitors fuel feed to provide precise response through five superbly ratioed gears.

Alliance DL. Front wheel drive from the world's all-time leader in front wheel drive. And an integrated group of road-holding components. Of course there's independent McPherson front suspension and rack and pinion steering. But also an entirely new twin axial torsion bar/trailing arm rear suspension that's more efficient than leaf springs or coils. Fore and aft stabilizer bars and steel-belted radials are also standard.



European engineering.
Alliance DL. Nearly one and a half million hours of development and testing and over $200 million invested in American Motors' Kenosha assembly plant have produced a sophisticated small sedan of European breeding and American manufacture.

Renault has employed computers extensively throughout the design process. First, finite element analysis improves the structural integrity and helps pare weight. Then a UNISURF (uniform surface) computer probe "reads" the specifications of the design model and translates them—within a tenth of a millimeter—directly to the die that produces the Alliance body parts. Finally, the computer-controlled, fully automated Gilman assembly process assures precise and consistent body fit



from the first Alliance to the 100,000th.

The result of all this technology is a spirited car that looks as good in the wind tunnel (.39Cx) as it does on the road. How good is that? Good enough for the European version of the Alliance to win Europe's 1982 Car of the Year award.

Driver appeal and room for five.
Alliance DL. Proof-positive of Renault's passion for ergonomics. With bio-formed bucket seats on curved tracks to tilt and recline in totally new ways. An instrument group whose every element is in easy reach. And laser-calibrated lines of sight for improved visibility.

A driver's car, yes. But sensitive to passengers as well. The Alliance comfortably seats five, with extra hip and leg room provided by a stretched wheelbase and the added room under the pedestal front seats. Surprisingly,

there's separate rear seat heating, so the people in back can enjoy the Alliance as much as you do.

Sophisticated options.
Alliance DL. With standard equipment typical of fine European sedans, and a provocative array of options: computer-controlled 3-speed automatic transmission. Power windows and door locks. Clearcoat

metallic finishes and halogen lights. 5.5″ x 13″ cast aluminum wheels. Even a patented Infrawave Keyless Entry System, to lock and unlock the doors from outside the car.

One thing that's never an option: the exclusive American Motors Buyer Protection Plan.® With the only full 12-month, 12,000-mile warranty that covers every part except tires, even if it just wears out. Plus a five-year No Rust-Thru Limited Warranty.™

52 est. hwy.

*Compare 1983 EPA estimates with estimated MPG for other cars. Your actual mileage depends on speed, trip length and weather. Actual highway mileage and CA figures will probably be lower.

37 EPA est. MPG

A surprisingly low price.
Such European engineering, precision manufacturing and fuel economy* is remarkable. But the real achievement is this: Renault Alliance starts at **$5,595****

Renault Alliance. The Alliance of quality and affordability is here. At Renault and American Motors dealers.

Vehicle type:	5-passenger front wheel drive 2-door and 4-door sedan
Engine type:	Transverse-mounted, 4-in-line, Bendix single-point electronic fuel injection (Bosch multi-point in CA)
Steering	Rack and Pinion with Opt. Power Assist
Brakes	Power-Assisted Front Disc, Rear Drum
Suspension	4-Wheel Independent
Displacement	1397 cc
Wheelbase	97.8 in.
Length	163.8 in.
Curb Weight	1945 lbs. 2-dr. 1980 lbs. 4-dr.

RENAULT

American Motors

**Manufacturer's suggested retail price for the Alliance 2-door. Price does not include tax, license, destination charges and other optional or regionally required equipment.

BOOKS

Legend and reality in volcanic eruptions, and a magician of scientific illustration

by Philip Morrison

The Time of Darkness: Local Legends and Volcanic Reality in Papua New Guinea, by R. J. Blong. University of Washington Press ($25). The 1980 Eruptions of Mount St. Helens, Washington, edited by Peter W. Lipman and Donal R. Mullineaux. U.S. Geological Survey Professional Paper 1250, U.S. Government Printing Office ($35). The trackless mountain plateau of Papua New Guinea is home to a million people or more, their well-tended green gardens scattered among luxuriant rain forest and tall-grass lands dissected by valley and peak. They have lived there for a long time, long enough for the land itself to have become a garden of languages and cultures. Under the shadow of Mount Hagen, central to the region, there is a prehistoric site, the Kuk swamp, where the spade testifies to nearly 10,000 years of pig keeping and at least 6,000 years of ditched gardens. The stratigraphy, however, is not at all obvious; rainfall is more than 100 inches per year, and the topsoil has been much worked by human hands. Ten thin, irregular layers of recognizable volcanic ash provide the relative chronology.

The Time of Darkness records the surprising way of an Australian investigator whose acute vision retains a wide field of view. He sought to extend the utility of that telltale ash. Could it be recognized as a time marker across the entire plateau? R. J. Blong's study began with the topmost layer, an olive gray, silty fine sand called the Tibito Tephra, lying an inch or so thick a foot down in the peat of the swamp. For comparison he studied the second layer too, the Olgaboli Tephra, a grayer deposit typically found a couple of feet lower. More than 200 field samples were taken at sites scattered across a couple of hundred miles of the highlands and down to the coast.

All the ash samples were carefully analyzed, mainly by X-ray spectrometry. The analysis of the major elements is slow and disappointing; it does not discriminate the two layers. On the other hand, trace analysis, say for 10 or 100 parts per million of strontium and rubidium, is revealing. Points for the two ash layers fill two satisfyingly well-separated patches on the plot of rubidium-v.-strontium content. The recent Tibito Tephra is a single fall, to be recognized right across the land in any small sample. Whence did it originate? The eroded volcanoes everywhere in the highlands are long dormant; the few small steaming vents here and there can be excluded. Terrible Krakatau comes at once to mind. It is 2,000 miles west, but such far-flung ash is always rich in silica; the heavier particles of rocky stuff fall out faster than the frothy siliceous pumice as the clouds drift downwind. Yet no highland ash is as rich in silica as the ash Krakatau sent over nearby Java.

The volcanologists had already determined that there is a chemical gradient along the arc of active island volcanoes offshore from New Guinea in the Bismarck Sea. Probably reflecting the rate of plate convergence, the efflux from these big craters differs systematically in composition from west to east. Both tephra layers fit well into the series; only one or two island volcanoes might be their source. The lake-filled caldera on Long Island, five miles wide, is the best candidate. The island fringe has been carefully studied. For three periods over the past 20,000 years Long Island has shown powerful activity, once quite recently. Its ashes resemble the Tibito Tephra closely in both minor- and major-element composition. Mapping the samples strengthens the indictment. The plot of highland ash-layer thickness against sample distances from Long Island shows a crude but statistically satisfactory pattern; the layers thin out and the particle size dwindles steadily downwind from Long Island. A complete map cannot be prepared from the spotty samples, but the trend is compelling. Long Island, 300 miles upwind from Kuk swamp, was the source of the Tibito Tephra (although not of the older layer).

That was a big event. The total volume of ash can be estimated: it is tens of cubic kilometers. Long Island's latest eruption was rather bigger than that of Krakatau and more than an order of magnitude bigger than that of Mount St. Helens. It stands as a major eruption on a world scale not merely over a decade or two but over a couple of centuries. Its ash fall would surely have brought darkness at noon to a sizable area, and the Papuan highlanders bear witness.

The second half of this geologist-turned-ethnographer's admirable detective work is the recovery of the human record. It is carried in oral tradition—legend—as told in scores of languages across the highlands. A survey, made by postal questionnaire and printed appeal, brought in a sampling of the relevant tales. Eight are retold here. "In olden days, men saw to the south the whole land was covered with dark clouds.... The ash-storm reached them and they had to stay inside their houses for four or five nights.... Gradually over a two-day period it grew light again." Or: "On a certain morning long ago... the people of our village awakened to find that the day did not lighten." Or again: "I am going to tell the story of darkness.... I did not see it. People told me and so I know it." Again: "When they went outside, they found everything covered with sand—like the sand in a stream."

The inquiry drew some 100 replies, mostly through linguists and anthropologists. The names and places of the informants are given in full. Of course, translation is a source of noise. There are nonetheless about 56 responses that probably refer to the fall of tephra, and only a dozen in which the informant knew of no such legend; the rest are perhaps accounts of hail or snow or eclipses, or simply remain ambiguous. Most of the respondents believed the event was historical and had happened only once. Six physical samples came in, collected in response to a question in the legend and offered as the material that fell during the time of darkness. Four of the samples are the Tibito Tephra itself; the other two seem to be volcanic ash but from some other fall. The people still recognize the kind of stuff that fell from the sky to darken the day. In some places they remember the ash as evil; in others it was the basis of a period of plenty. Ethnovolcanology is manifestly alive in the highlands.

The legends are carefully sifted for their detailed implications on the ash fall. The question left most uncertain is the epoch of that ancient time of darkness. Carbon-14 dating of partly burned trees on Long Island, which may well have been killed by the latest eruption there, is rather uncertain; the best calibrations suggest a time around A.D.

Finally you can afford to satisfy your lust for power.

It's been said that the future belongs to those who understand computers.

Well, this is your chance.

What you see here is the Sinclair ZX81 personal computer for only $99.95.

It's a personal computer that you can order right now. That you can learn to use. And that will open the door to understanding the technology that is rapidly changing the very way you live and work.

The computer for everyone.

The ZX81 is a true personal computer—not a business computer or an elaborate games computer. Everyone can afford it and learn to use it.

It hooks up to your black and white or color TV for video display. And it comes with a comprehensive instruction book that takes you step-by-step into the world of computers.

You and your children will learn how computers work. How to write your own programs. Even how to create computer games.

And Sinclair offers dozens of ready-to-use software packages. Try your skills at a game of chess. Learn famous composers and musicians. Computerize your personal budget. These programs and more are described in our software catalog which you'll receive with your ZX81. The programs come on cassette tapes, and work with a standard cassette recorder or player.

The $99.95 personal computer.

For less than $100, the Sinclair ZX81 will get you started in personal computing right now. Your children will gain an understanding of computers that will benefit them the rest of their lives. And you will be prepared to make informed decisions about using and buying computers, both in your career and in your home.

Free guide to programming

There's never been a better time to buy a personal computer.

And for only $99.95, there's never been a better price.

How to order today.

Call our toll-free number and use your MasterCard or VISA. Or send the coupon along with a check or money order.

Then try out the ZX81 for 10 days. If you're not satisfied, just return it to us and we'll refund your money.

Call toll free: 800-543-3000.

Ask for operator #509. In Ohio call: 800-582-1364; in Canada call: 513-729-4300. Ask for operator #509. Have your MasterCard or VISA ready.

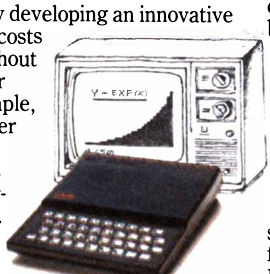
Phones open 24 hours a day, 7 days a week. These numbers are for orders only.

If you just want information, please write: Sinclair Research Ltd., 2 Sinclair Plaza, Nashua, NH 03061.

How did we do it?

Other personal computers cost hundreds, even thousands of dollars more. How did Sinclair create a personal computer for less than $100?

We did it by developing an innovative design that cuts costs dramatically without cutting computer power. For example, our unique Master Chip replaces as many as 18 chips used in other personal computers.



Sinclair programs are available

The success of the ZX81 speaks for itself. It is now the fastest-selling personal computer in the world. And we stand behind our product. If anything goes wrong in the first 90 days, we'll repair or replace it free of charge. Even after that, our national service-by-mail facilities are as close as your post office.

Why now?

Why should you take the time right now to order your Sinclair?

Because computers have become an important part of everyone's life. In business, you'll find them on the desks of both secretaries and presidents. In schools, you'll find them in a growing number of classrooms. Even in grade schools.

To order call toll free 800-543-3000 Operator #509

AD CODE: A0SA

Sinclair Research Ltd., One Sinclair Plaza, Nashua, NH 03061.

Yes! I'd like to try the Sinclair ZX81 personal computer. I understand that if I'm not completely satisfied, I can return it in 10 days for a full refund.

	PRICE†	QTY.	AMOUNT
ZX81 personal computer	$99.95		
Add shipping	$4.95	–	$4.95
		TOTAL	

†U.S. Dollars

☐ Check or money order enclosed.

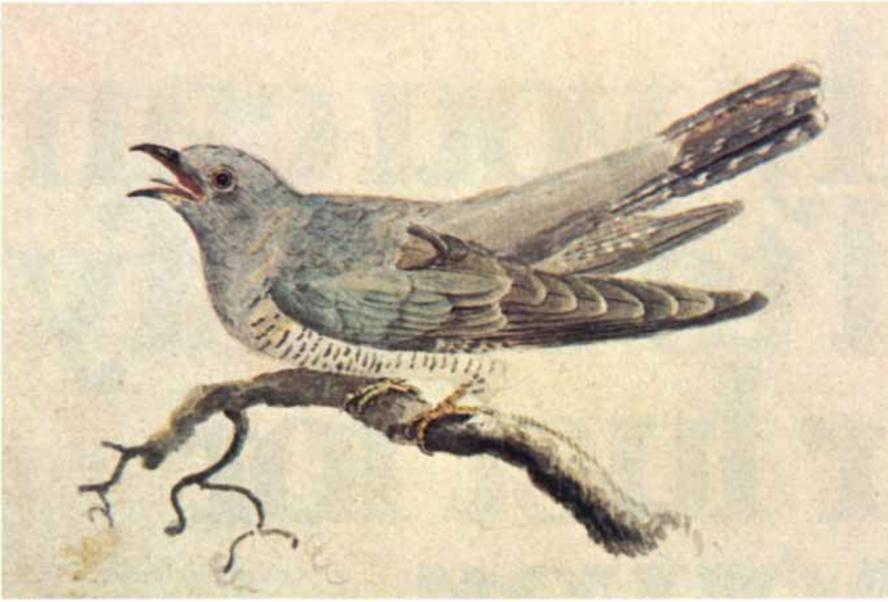
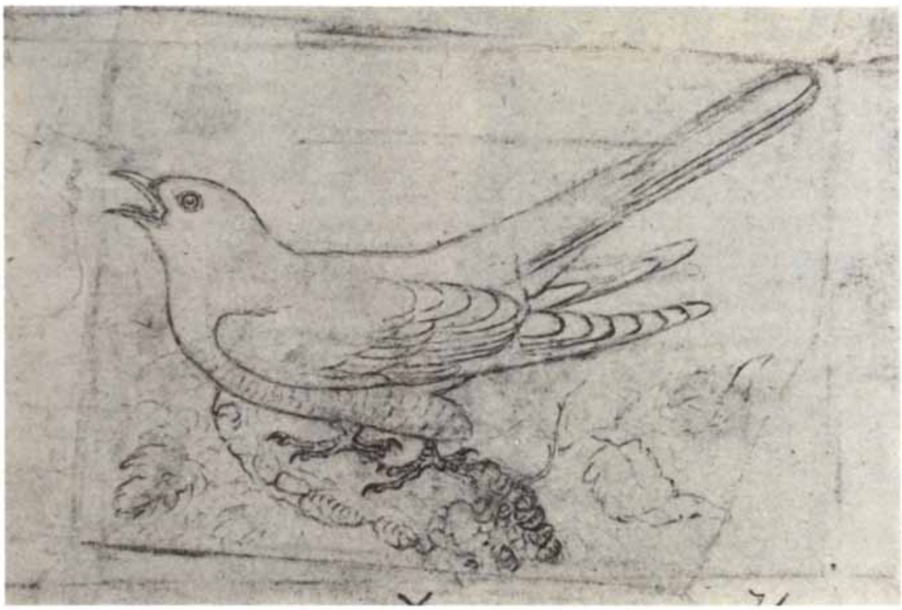
Name ______

Street ______

City ______ State ______ Zip ______

sinclair

Sinclair technology is also available in Timex/Sinclair computers under a license from Sinclair Research Ltd.

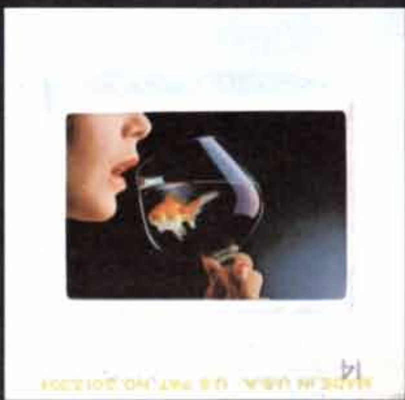

A Bewick cuckoo: watercolor study **(top),** *pencil drawing* **(middle)** *and wood engraving*

1650. The people now on Long Island seem to agree that they are relative newcomers. Travelers' reports help a little but are inconclusive. The English buccaneer-naturalist who named the "very green and flourishing" Long Island drew a profile like today's as he sailed by in 1700, but that cannot quite exclude a later explosion. A Russian nobleman who lived on the mainland shore of the Bismarck Sea opposite the island through the 1870's offers observations that exclude any time much later than 1820.

The preponderance of Papuan taletellers put the darkening only three or four generations ago. Only a few counted seven generations back or more, and none gave the clear estimate of eight or 10 generations the carbon-14 dating seems to imply. Such quantitative information appears the least reliable in the unwritten history of the Papuans. There are hints from the coast of two darkenings at one place; possibly even the older Olgaboli fall is dimly remembered there, along with an island that disappeared in a case of caldera collapse still unknown to the geologists, a millennium ago.

The 1980 Eruptions of Mount St. Helens, a hefty volume strikingly illustrated with many action photographs in color, presents the definitive professional study of our recent Oregon "time of darkness." There is even a good deal of eyewitness account among the 62 papers that treat the events at the volcano through 1980, its most active year so far. The editors are expert indeed; Donal Mullineaux is one of the senior volcanologists who recognized that St. Helens is the youngest and most active of all the Cascades and who in 1978 forecast its vigorous eruption within the century. He and Dwight R. Crandell review its eruptive history over Holocene time in the first paper. The volume describes the 1980 events in full detail, with the geophysical record, the analyses of all deposits, the widespread effects and assays of potential hazard and prediction.

Suffice it to admire this big and handsome book by giving only the following impressions of its rich and detailed contents. Gary Rosenquist was camping with a few friends about 10 miles northeast of the summit. On that Sunday morning one of the party noticed some steam emerging from the mountain. Rosenquist put his camera on a tripod and took one picture. He left the camera in place. His companion watched the mountain through binoculars, until he saw motion. He shouted that the mountain was going, and Rosenquist began taking pictures as fast as he could wind the film. He managed two dozen, of which 10 are reproduced here in color.




Vivitar

By arrangement with government officials throughout the world, a complete collection of all the circulating coins — in mint-fresh condition — from every coin-issuing country of the world.

Please mail your application by October 31, 1982.

Limit of one collection per person.

You have the unprecedented opportunity to acquire a collection of world coinage unlike any that has ever been issued before. A collection of *complete sets* of official monetary coins from the coin-issuing nations of the world—with each set sealed in an individual cachet, date-canceled and postmarked in the country of issue.

This comprehensive collection will provide every member of your family with an intriguing way to learn about countries and peoples in every part of the world—through their official coinages. It will be an enjoyable and enriching experience, and an educational adventure as well. For you will gain a better understanding of each country—while you are building a comprehensive collection of world coins that is certain to become a prized family possession.

Assembled by special arrangement with government officials in more than 100 countries

To put together the Coin Sets of All Nations collection, The Franklin Mint made special arrangements with the central banks or monetary authorities of more than one-hundred coin-issuing nations—and with the individual postal authority of each of those nations. Every country that regularly mints and issues coins will be represented in this collection, except where government regulations or restrictions on availability prohibit.

The result will be a comprehensive and fascinating collection of *mint-fresh* coins that would be extremely difficult to assemble even if you were to travel to every one of the coin-issuing countries of the world—because coins in mint-fresh condition usually are obtainable only for a short period of time after they're struck.

Sealed in specially postmarked cachets

Here are the features that make this collection unique:

- All of the circulating coins from each of the coin-issuing nations will be included in the collection.
- Each coin in every set will be in mint-fresh condition.
- Every coin set will be sealed in its own individual cachet, which has been designed especially for this collection.
- Each cachet will be officially stamped and date-canceled, and will be postmarked in the country that issued the coins.
- The complete collection will be available only by subscription.

Coin Sets of All Nations will not be sold through any coin dealers, stores or even national banks in this country or abroad. The collection can only be acquired directly from The Franklin Mint and solely by subscription.

Educational and enjoyable for every member of the family

Many of the coins in this series are outstanding for their beauty, their historical significance and the themes they represent. Thus the collection provides a pleasurable learning experience for young and old alike. The coinage of Greece, for example, evokes its rich classical heritage: the 20 drachmai portraying the great statesman Pericles and the 10 drachmai, the philosopher Democritus. The coinage of Japan features the 100 yen coin with its lovely design of cherry blossoms and the 10 yen with its portrayal of the Byodo-in Temple near Kyoto, capital of ancient Japan.

The coinage of Austria consists of 8 different coins, with the 5 schilling piece showing a Lippizaner stallion of the renowned Spanish Riding School in Vienna. Each coin in Australia's issue portrays a different native animal: the kangaroo, emu, platypus, lyrebird, spiny anteater. The coins of Indonesia depict the exotic birds of that land, and the coins of Fiji ceremonial objects unique to the culture of this island people. By contrast, the coinage of Sweden is very formal: the 1 Krona bearing a classic portrait of King Carl Gustaf and the 50 Ore featuring the royal monogram and the Swedish Crown. More than 100 complete sets in all—each one an education in itself.

These official coinage sets form a collection of infinite variety—and include coins of many different shapes: round, octagonal, twelve-sided, scalloped. And they are minted in a variety of metals: bronze, copper, brass, nickel and cupro-nickel.

Storage cases and reference folders provided

To enable you to store and protect your coinage cachets—and locate the issues quickly and easily—a set of four handsome hardbound cases will be included as part of the collection. In addition, a specially written folder will be sent with each coinage set. It will give important facts about the country represented and background information on its coinage—information you won't find in most books or magazines.

As a subscriber, you will receive your collection at the rate of two issues per month. The price for each issue is $13.95 including the mint-fresh coins, cachets, stamp, foreign postmarking and all customs charges. There is *no added charge* for the storage cases or reference folders.

Subscription applications for Coin Sets of All Nations are being accepted at this time. But because of the difficulty of obtaining a sufficient quantity of *mint-fresh* coins for every nation represented, it is important that your application be mailed promptly. (A future opportunity to subscribe may be offered but cannot be guaranteed.)

To subscribe for this complete series of coinage sets from around the world—and to bring your family a fascinating and educational experience that will enrich their lives—be sure to mail the accompanying postpaid application to The Franklin Mint, Franklin Center, PA, by October 31st.

Coins in special date-canceled cachets, shown smaller than actual size.

An amateur's prize, they form one basis for the careful quantitative study of the initial event.

A volume of magma some 600 meters across had slowly intruded into and visibly bulged the upper north flank of the mountain cone. Less than a couple of kilometers underground lay the seat of a fierce swarm of small earthquakes, the seismic sign of steady fracturing of rock by the magmatic pressure below. The last of these shocks, at 0832 : 11 PDT on May 18, 1980, so shook the stressed and thinned rock wall above that it failed under gravity. Two or three cubic kilometers slid slowly down the north slope of rock and ice. The heavy lid was suddenly off, and the superheated water and dissolved gases in the intruded melt exploded northward. Thirty seconds later the main magma conduit was itself exposed; its sudden depressurizing gave rise to an eruption column high into the stratosphere, a froth of rock, the widespread tephra.

The darkness was quite local and transient, but then the cavity of the Long Island caldera is very much larger than the missing north wall of Mount St. Helens. That wall, its ice chunks melted, lies now as rubble 50 meters thick, choking the valley of the North Fork of the Toutle River, its contours and rock types colorfully mapped in the Geological Survey quadrangle that of course accompanies and summarizes this many-sided account.

The Watercolours and Drawings of Thomas Bewick and His Workshop Apprentices, introduced and with editorial notes by Iain Bain. The MIT Press ($225). Thomas Bewick's Birds: Watercolours and Engravings. The MIT Press ($8.95). "I have seen him draw a striking likeness on his thumbnail, in one moment; wipe it off with his tongue, and instantly draw another." So wrote a friend calling up memories of Thomas Bewick. Once a country schoolboy crazy about drawing, Bewick himself recollected that "as soon as my question was done upon my Slate, I spent as much time as I could find, in filling, with my pencil, all the other spaces of it, with representations of such objects as struck my fancy.... At that time I had never heard of the word 'drawing,'... nor did I know of any other paintings, besides the Kings Arms in the Church" and the four pub signs in the village nearby.

In 1767, at 14, Bewick was wisely apprenticed to a talented general engraver in Newcastle upon Tyne, to pass most of his life in or around that "smoky town." Working over the years with many apprentices, including his brother and his son, he became the most admired of English wood engravers. The topics of his extended work were natural history; his volumes of engravings—on quadrupeds ("every Animal of that Kind"), on British birds of land and water and finally a *Fables of Aesop*—all went through many editions. He died with his last project incomplete; it was a *History of British Fishes,* jointly prepared along with his diffident and talented son Robert.

His genius both of conception and of technique is manifest at small scale. Wood engraving was done on end-grain slices sawed to a thickness matching the height of printer's type. The imported "Sticks of Turkey box," a hard, close-grained wood, were usually not much wider than the palm of the hand. Direct comparison of Bewick's engravings with the copies made by the remarkably skillful reproductive engravers so conspicuous in early Victorian times makes plain his masterly and expressive technique. He did not work as they did, cutting regularly down around the drawn lines, but rather by releasing most black lines rather freely from the white areas treated to yield contrast and texture. A complete sequence of his mature work, say in the *Birds,* began with a splendid small watercolor. Then he made a careful drawing in pencil. From that drawing he transferred the main outlines to a prepared block by scribing over the drawing, whose back he blackened in soft pencil. The block was probably coated with chalk or brick dust to take the transfer. He worked best from life (a traveling menagerie was for him a windfall), but he also used books and the stuffed specimens of the museums.

The big two-volume set of *The Watercolours and Drawings of Thomas Bewick* is a fine example both of bookmaking and of scholarship, historical and visual. Its presentation of Bewick's remarkably detailed and lively watercolors is something of a surprise; most know of him only through the published engravings, which of course exist in a great many copies, widely reproduced to this day. The watercolors are actually no recent find, but they have not before been so carefully associated with the familiar engraved outcome.

The smaller *Thomas Bewick's Birds* is a happy dividend from the larger book; it is a satisfying selection from the work for *British Birds*—jackdaw, owl, night heron—and a bargain. The text paragraphs that describe the birds of these 50 hand-size paintings and engravings are the work of Iain Bain, who is evidently both an art historian and an up-to-date birder. A group of partridge feathers rendered in convincing fluffiness and color delightfully fill the endpapers of the smaller book; they were understandably never translated into line for reproduction.

Color lithography became the dominant ornament for bookmaking in large editions only after Bewick's time. It is remarkable how devotedly this artist, long the private polychrome master of watercolor in miniature, wrought in black ink, the only technique for mass reproduction at high resolution exploited in his day. The critic muses on what Bewick might have done with a color process fit to publish his painstaking genius. Perhaps the austere constraints that were set by printer's ink and boxwood were essential to that compact brilliance?

Agricultural Plants, by R. H. M. Langer and G. D. Hill. Cambridge University Press ($39.50). Human life rises from grass roots. That truism is given strong foundation in this interesting and useful book, which lies between an introductory text and a work of reference, between undergraduate botany and specialized agronomy. About a third of its pages treat the family of the grasses, the source not only of our daily-bread grains but also of the savory roasts of the well-to-do; cattle and sheep depend largely on forage crops or on natural or sown pasture, again chiefly grassy plants. (Rice is a little slighted, naturally enough in a work by New Zealanders that centers on Temperate Zone field crops.)

The book opens with a brief account of agricultural production over space and time. A fresh chapter then gives the essentials of plant tissues and structures. The next 13 chapters contain summary accounts of many Temperate Zone crops, arranged by families. The view is wide, omitting only tree crops and gardening; besides the great family of the grasses we find the full range from cotton, tobacco and hemp to eggplant, radish and winged bean. These chapters, say the authors, define a kind of vocabulary for the language of agriculture; its grammar and use are foreshadowed in one final chapter on the physiological basis of plant yield.

Grasses evolved in a world of hungry grazers. Their survival is a function of their structure and capabilities: the vital growing apex that sends out tillers—the new shoots—is kept safe at soil level, out of the way of the grazer. Leaves still in growth can continue even after their tops are taken off. The trampling of hoofs is also mitigated: tillers partly separated from the rest of the plant can strike more roots of their own. In fact, a grass plant is a little population of rooted tillers, whose statistics determine the plant's overall life cycle. If the majority flower and die in open season, the grass is an annual. If only some flower and some persist in forming leaves, the plant continues indefinitely. How diverse maize and barley and the rest have become in their symbiosis with humankind is an old story; the freshest part of this account concerns the pasture grasses. For each important grass genus there is a small narrative of structure, history and use.

The grasses, however, are not all. A fine grassland pasture will be sown with a mixture of a grass with some forage legume, for example rye grass with white clover. That clover is a prostrate, branching plant in easy symbiosis with the nodular nitrogen-fixing bacteria at its roots. Such pasturage is palatable to all livestock, and it long sustains a careful balance of grazing and mineral fertilization. Nitrogen fertilizer overencourages the grass; insufficient grazing denies light and hence growth to the overshadowed low clover. The system has more components still. If bees are few, only a poet's prairie can flourish. Then there is another legume: alfalfa, or lucerne. It spread from Anatolia along with the nomads and their cavalry, the jet fuel of ancient warfare, the authors say. It is a highly productive source of nutritious hay, particularly in dry climates, but no one has found any companion grass to fit its growth rhythm. Honeybees do not pollinate alfalfa well; bumblebees, so unsocial a population as to be beyond human management, are reliable only where they are abundant.

We read a little of bloat, the formation of a stabilized foam in the fermenting grassy rumen of cattle that have overeaten rich alfalfa or clover. There is also a biochemical hazard. Glucosides are nitrogenous sugars, some of which can hydrolyze to release hydrogen cyanide; ruminants can detoxify that poison to a degree, but some strains of clover, such as the grassy sorghums, lead to grazer's trouble. The sense of system is well conveyed.

It is sunlight that supplies the free energy to pasture and field. Here is a brief tale of the two distinct routes of photosynthesis; they seem to be adaptations to two climates rather than two open options. The C_4 plants are tropical grasses such as corn, good crops for hot summer weather. It is the strong correlation between leaf area and yield, however, that is the key. "The art of grazing management has its scientific basis in the concept of leaf area index": the ratio of the leaf area to the area of ground covered, integrated over the time of growth.

This text is not simple for the nonbotanical reader, but it is a valuable path around the bigger specialized volumes, a guide to the extensive literature and a city reader's eye-opener. The drawings in line and stipple are numerous and attractive.

An Album of Fluid Motion, assembled by Milton Van Dyke. The Parabolic Press, P.O. Box 3032, Stanford, Calif. 94305 ($20). "We who work in fluid mechanics are fortunate," the author disarmingly begins, "that our subject is easily visualized." Most texts and many articles on fluid mechanics present a few flow photographs. Over the decades the scientific literature has grown into a treasure house of flows visualized. The methods are many. Some are as simple as the use of aluminum dust, smoke, ink or fluorescent dye to mark flow paths in a transparent fluid. An immersed wire pulsed with electric current can generate precise trains of tiny bubbles by the electrolysis of water. Various shadowgraph and optical-interference schemes render the transparent more subtly visible, taking advantage, for instance, of minute gradations of density.

The three-dimensional form of some phenomena is often confusing, with all meaning self-shielded; a cunning optical sample made by a slit or a line of illumination can simplify a great deal. The manipulation of time is central. Time exposures integrate motion; fast flashes stop it. Ingenious experimenters have prepared special setups to display to the eye just the point they choose to make, sometimes realizing an abstract mathematical construction or displaying a sequence of events to demonstrate the critical influence of some single flow parameter.

All these images and more are loosely ordered by topic in this extensive compilation. Each picture has a clear explanatory caption; an intrigued innocent can enjoy the images for their manifest meaning and beauty. But the more you know, the more you gain. The explanations are terse and precise, if informal. An expert or a serious student will profit most; he who runs will overlook what she who counts rings or compares distances can delight in. The center of the work is at the scale of the laboratory and the test wind tunnel; few of the flow patterns shown are tiny, and only a couple, such as a jet aircraft's contrail or the bore of the River Severn, were made outdoors. Professor Van Dyke has stayed with the prepared model; no tornado, galaxy or bacterium is to be found. The volume therefore has a somewhat austere interest. Mainly in the sections on supersonic flow and shock waves do we see practical forms, real projectiles or model aircraft; most of the geometry is the sphere, the cylinder, the slot, the plate or the ring, all dear to the theorists. Let the chosen start be simple, soon enough the continuous flowing fluid expresses its inherent complexity.

A few examples will serve to indicate what is here. These are images meant to be examined by the eye in full two-dimensional spread; they are only meagerly mapped onto a short string of words. Smoke rings are shown in a dozen shots,


commodore
THE COMMODORE 64. ONLY $595.
What nobody else can give you at twice the price.

leapfrogging as one ring passes through the leader and is in turn overtaken and penetrated: interaction with the flow field. Another splendid view in the book, repeated on the back cover, makes clear what we all should have seen but missed. The real vortex ring is not a single closed doughnut surface; rather within it there is a tight toroidal spiral, the outer sheet curling in on itself a score of times. (It works with dye in water too.)

The flow of a jet past a smoke wire displays with dazzling virtuosity a slow passage from bland order through surface oscillations breaking into rolls and vortex rings, ending in wispy chaos. A layer of heated oil marked with aluminum powder forms a striking honeycomb of hexagonal convection cells, within which a single sixfold set of aberrant diamonds signals a tiny dent in the hot plate below. A page shows a sequence of simple waves in water, but the familiar watery motion is traced by small white particles. Below the pure running wave the water elements execute their small perfect circles; within a standing wave the water streams in smooth arcs back and forth from trough to crest.

Finally, in the Army Ballistic Research Laboratory they sent a small sphere flying at Mach 3 past a plate with a line of regularly spaced holes. As the bow shock that travels with the projectile passes each opening its presence initiates an acoustic wave, a half sphere visibly expanding from each small hole. The envelope of these waves defines the Mach number graphically in a single photograph.

There is even a sense of history. The frontispiece is a print of a brass bullet in flight enlarged from a tiny negative made almost a century ago in Prague by Ernst Mach himself, in the first year of the visualization of the supersonic bow shock. In 1883, when Osborne Reynolds carefully fed a line of dye along the axis of a glass tube carrying a flow of water, he reported by sketches rather than photographs. His apparatus remains in Manchester, and his successors there have repeated his work with photographs to display the sharp onset of turbulent mixing as the flow speed increases. The ceaseless tremor of "modern traffic ... made the critical Reynolds number lower than the value 13,000 found by Reynolds."

No color photographs are found here; although they can be striking, they would have raised the price beyond the student's resources, a point much on the mind of Professor Van Dyke, who devotedly assembled this set of prints from many colleagues around that world where all things flow. Self-published, designed handsomely within the family, this collection needs and will enduringly reward many owners. What technical library can be without it?


ROOM WITH A VIEW
'83 Ford Club Wagons. Here's hotel luxury — to go.
But a luxury hotel isn't rated to handle up to 10,000 pounds* of loaded trailer. Ford Club Wagons are — they can outtow and outcarry any station wagon on the road.
Ford Club Wagons accommodate up to 12 passengers in comfort that would make a hotel proud. Or 15 folks in Ford's Super Wagon! With the bench seat removed, a Club Wagon has more cargo length than Chevy Suburban. Plus more seating arrangements (like Captain's Chairs) than Suburban.
And all Club Wagons are Built Ford Tough with rugged body-on-frame construction and Twin-I-Beam front suspension.
A luxury hotel never had it so good.
*Available December, 1982.
FORD CLUB WAGON
FORD DIVISION
Ford
Get it together — Buckle up.
AMERICA'S TRUCK
BUILT FORD TOUGH

Established 1845 SCIENTIFIC AMERICAN October 1982 Volume 247 Number 4

The Verification of a Comprehensive Nuclear Test Ban

Networks of seismic instruments could monitor a total test ban with high reliability. Even small clandestine explosions could be identified even if extreme measures were taken to evade detection

by Lynn R. Sykes and Jack F. Evernden

Two treaties put into effect over the past 20 years have set limits on the testing of nuclear weapons. The Limited Test Ban Treaty of 1963, which has been signed by more than 120 nations, prohibits nuclear explosions in the atmosphere, the oceans and space, allowing them only underground. The Threshold Test Ban Treaty of 1976, a bilateral agreement between the U.S. and the U.S.S.R., prohibits underground tests of nuclear weapons with a yield greater than 150 kilotons. In the present climate of widespread pressure for more effective control of nuclear arms the idea of a comprehensive ban on all nuclear testing is receiving renewed attention. Such an agreement would be an important measure. It might inhibit the development of new weapons by the major nuclear powers, and it might also help to prevent the spread of nuclear-weapons technology to other countries.

A halt to all testing was the original goal of the negotiations that led to the 1963 Limited Test Ban. New talks with the aim of achieving a total ban were begun in 1977 by the U.S., the U.S.S.R. and Britain, but the talks were suspended in 1980. In both cases the main impediment to a comprehensive treaty was the contention by the U.S. and Britain that compliance with the treaty could not be verified because sufficiently small underground nuclear explosions could not be reliably detected and identified. In July the Reagan Administration announced that the test-ban negotiations with the U.S.S.R. and Britain will not be resumed. Once again the primary reason given was a lack of confidence in methods of verifying compliance.

In 1963 the reliability of measures for the verification of a treaty banning explosions larger than about one kiloton may have been arguable, but it no longer is. We address this question as seismologists who have been concerned for many years with the detection of underground explosions by seismic methods and with means of distinguishing underground explosions from earthquakes. We are certain that the state of knowledge of seismology and the techniques for monitoring seismic waves are sufficient to ensure that a feasible seismic network could soon detect a clandestine underground testing program involving explosions as small as one kiloton. In short, the technical capabilities needed to police a comprehensive test ban down to explosions of very small size unquestionably exist; the issues to be resolved are political.

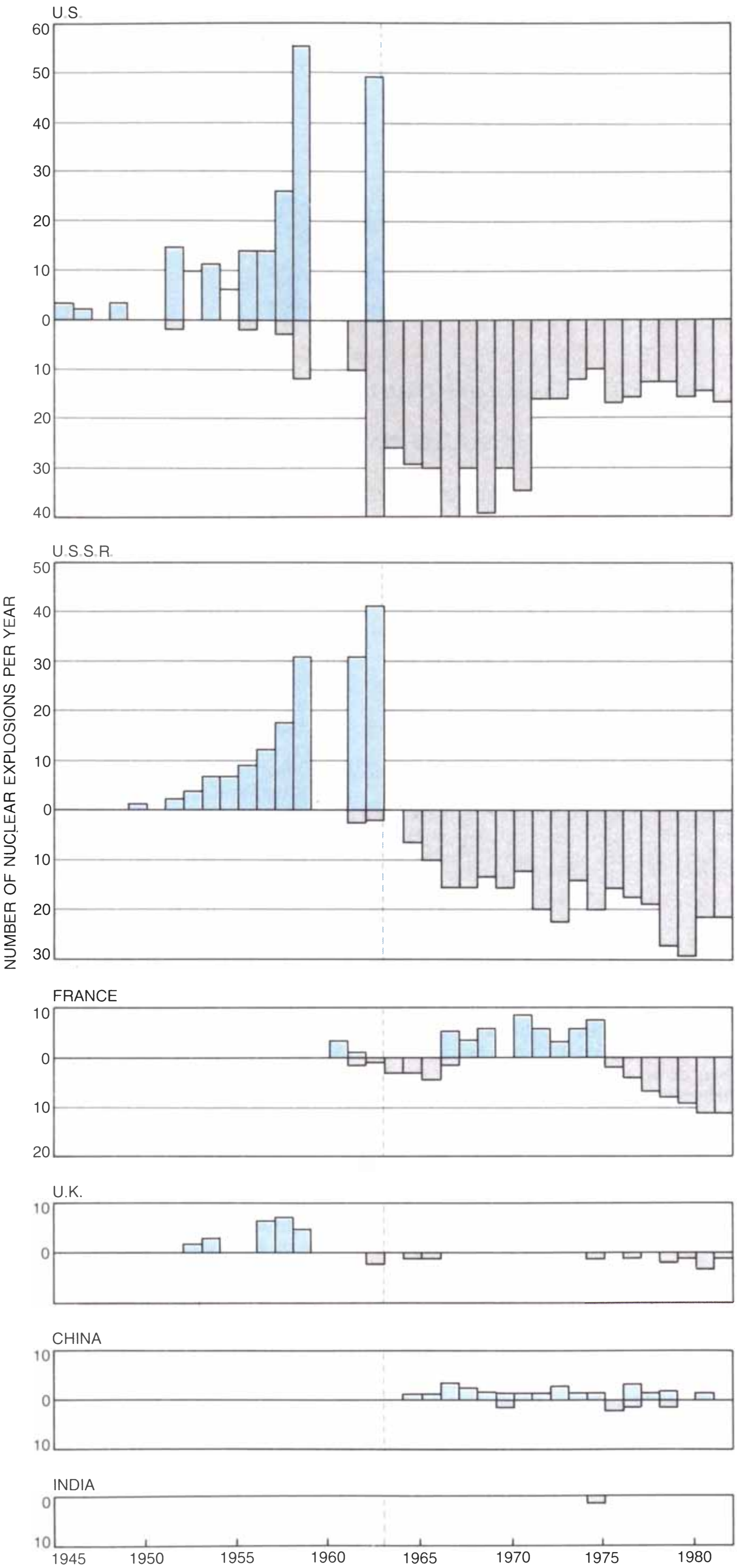
An underground explosion sets up elastic vibrations that propagate as seismic waves through the earth and along its surface. The waves travel great distances, and seismic monitoring instruments in common use are sensitive enough to record even those generated by very small explosions. Once the waves have been detected the main task is to distinguish the seismic signals of explosions from those of earthquakes. This can be done with a network of several widely separated seismometers.

Two types of elastic vibrations can propagate through the solid body of the earth, that is, through the crust and the mantle. The first waves to arrive at a seismometer are compressional waves, which are similar to sound waves in air or water; the seismological name for them is *P* (for primary) waves. The slower body vibrations are shear waves, which are similar to the waves on a vibrating string; they are called *S* (for shear or secondary) waves. An underground explosion is a source of nearly pure *P* waves because it applies a uniform pressure to the walls of the cavity it creates. An earthquake, on the other hand, is generated when two blocks of the earth's crust rapidly slide past each other along the plane of a fault. Because of this shearing motion an earthquake radiates predominantly *S* waves.

A result of the spherical symmetry of the explosion source is that all the seismic waves it generates have a nearly radial symmetry around the focus of the explosion. In contrast, the highly directional character of an earthquake source gives rise to seismic waves with strongly asymmetric patterns. The asymmetry in the amplitude of the waves received at seismometers throughout the world provides the means whereby seismologists can determine the faulting mechanism of a given earthquake.

In addition to the *P* and *S* body waves there are also two types of seismic waves that propagate only over the surface of the earth. They are called Rayleigh waves and Love waves, and they result from complex reflections of part of the body-wave energy in the upper layers of the earth's crust. A simple explosion can generate Rayleigh waves but not Love waves, whereas an earthquake generates waves of both types.

Seismologists characterize the size of a seismic event by means of magnitudes. A given event can be assigned several



magnitudes, each one based on a different kind of seismic wave. A magnitude is the logarithm of the amplitude of a particular type of wave normalized for distance and depth of focus. Of the numerous magnitudes that can be defined for a single seismic event we shall discuss only two, which in seismological notation are designated M_S and m_b. The former is generally based on Rayleigh waves with a period of 20 seconds, the latter on one-second *P* waves. The magnitude of a seismic signal is ultimately related to the energy released at the site of the event. For a nuclear explosion the customary measure of energy release is the yield in kilotons, where one kiloton is the energy released by detonating 1,000 tons of TNT.

Every year there are numerous earthquakes whose magnitudes are in the range corresponding to the yields of underground explosions. Several methods can be applied to several types of waves to distinguish the seismic waves of explosions from those of earthquakes. The location of a seismic event and its depth below the surface are important criteria; indeed, the great majority of routinely detected events can be classified as earthquakes simply because they are either too deep or not at a plausible site for an explosion. The remaining events can be reliably classified by the amount of energy radiated in the several kinds of waves at various frequencies.

The location of an event in latitude and longitude is a powerful tool for classification. The position is determined by recording the arrival time of short-period *P* waves at several seismographic stations in various parts of the world. The travel time of the *P* waves to each station is a function of distance and depth of focus. From the arrival times it is possible to determine the location of the source with an absolute error of less than 10 to 25 kilometers if the seismic data are of high quality.

The identification of seismic events at sea is quite simple. It is assumed that the network monitoring a test-ban treaty would include a small number of simple hydroacoustic stations around the shores of the oceans and on a few critical islands to measure pressure waves in seawater. The hydroacoustic signal of an underwater explosion is so different

NUCLEAR TESTS CONTINUE to be carried out at a rate of about 50 per year, principally by the two leading nuclear-weapons powers: the U.S. and the U.S.S.R. As this bar chart shows, the main effect of the Limited Test Ban Treaty of 1963 (*broken vertical line*) was not to reduce the number of test explosions but merely to drive most of them underground. Nuclear test explosions in the atmosphere and underwater are represented by colored bars, those underground by gray bars.

from that of an earthquake and can be detected at such long range that the identification of a seismic event at sea as an explosion or an earthquake is simple and positive. Hence any event whose calculated position is at least 25 kilometers at sea (a margin allowing for errors) can be classified as an earthquake on the basis of its location and the character of its hydroacoustic signal.

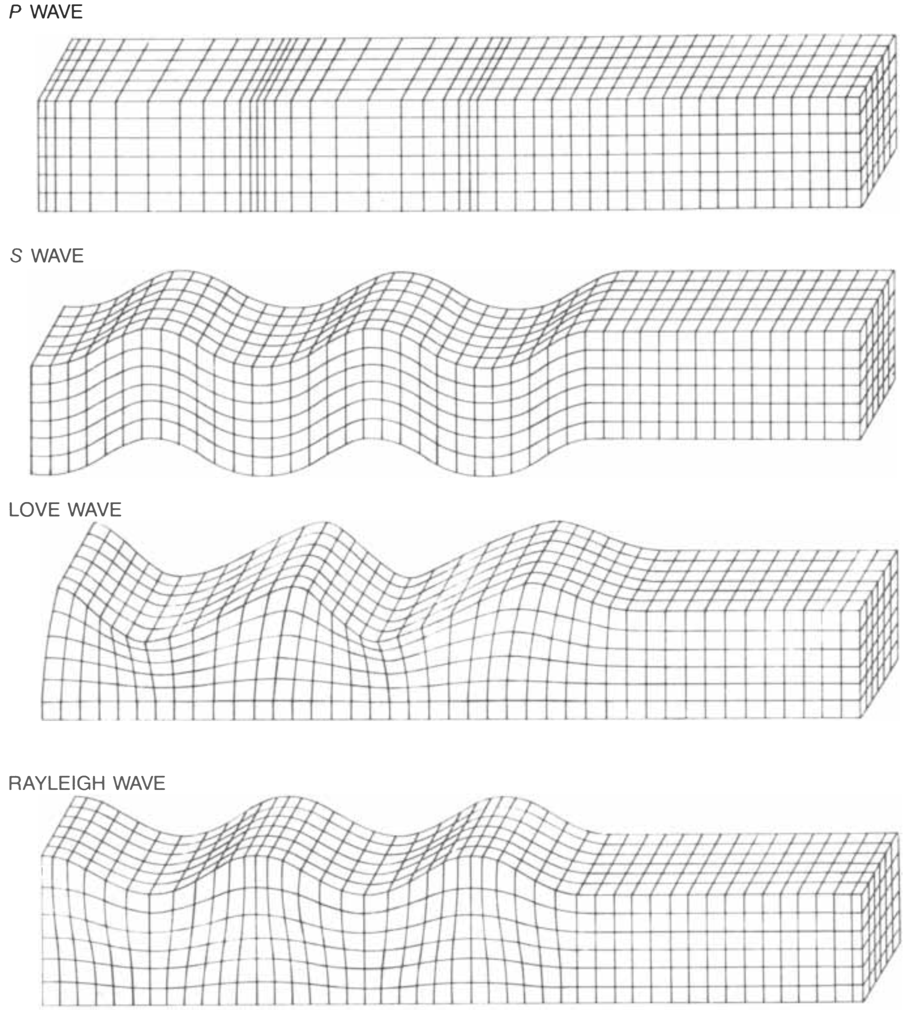
The accuracy with which the position of a seismic event can be determined in an area offshore of an island arc has been tested with an array of ocean-bottom seismometers off the Kamchatka Peninsula and the Kurile Islands in the U.S.S.R. The tests indicate that the accuracy of a seismic network under these circumstances is much better than 25 kilometers. Holding to that standard, however, one finds that well over half of the world's seismic events are definitely at sea and are therefore easily identified as earthquakes.

Another large group of detected events have their epicenters on land but in regions where no nuclear explosions are to be expected; these events too can be safely classified as earthquakes. Indeed, almost all the world's seismic activity is in regions that are of no concern for monitoring compliance with a comprehensive test ban. Thus the simple act of locating seismic events classifies most of them as earthquakes.

Calculating the depth of focus provides a means of identifying a large fraction of the remaining earthquakes. From 55 to 60 percent of the world's earthquakes are at depths of more than 30 kilometers; at least 90 percent are more than 10 kilometers deep. Any seismic event as deep as 15 kilometers is certainly an earthquake. No one has yet drilled into the earth's crust as far down as 10 kilometers, and the deepest nuclear explosions have been at a depth of about two kilometers.

Several seismological procedures can be employed to determine an event's depth of focus. In most cases the depth is calculated at the same time as the location. When a seismic event is detected at 20 stations or more, a routinely calculated depth of 30 kilometers or more ensures with a 95 percent degree of confidence that the event was at least 15 kilometers below the surface.

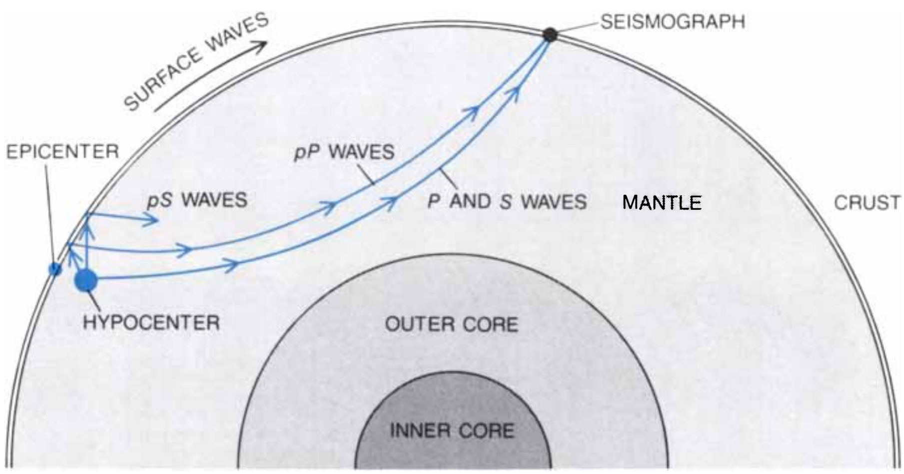
A powerful technique for estimating depth can be applied if at least one seismological station is within a few hundred kilometers of the detected event. (A monitoring network for a comprehensive test ban would be quite likely to meet this condition in areas where nuclear testing might be expected.) A pair of *P* and *S* waves generated at the same instant and recorded by a station near the event follow identical paths but propagate at different speeds. The difference in their times of arrival, or in other words the difference in their phases, therefore serves to determine the time of origin of the event. With experience the seismograms of a station near the event can be successfully analyzed to detect at least one pair of such *P* and *S* phases. Given the time of origin determined in this way and the arrival times of the *P* waves at only a few distant receivers, an accurate estimate of the depth of focus can be made.



FOUR TYPES OF SEISMIC WAVE are illustrated. The two waves at the top propagate through the solid body of the earth; the two at the bottom propagate only near the surface. The compressional body waves called *P* (for primary) waves travel fastest and are the first ones to arrive at a seismometer; they are the predominant type of body wave produced by an underground explosion. The slower body waves called *S* (for shear or secondary) waves vibrate in a plane transverse to their direction of propagation; they are the predominant type of body wave produced by an earthquake. The surface waves called Rayleigh waves and Love waves result from complex reflections of the *P* and *S* body waves in the upper layers of the earth's crust.

There may remain critical seismic regions where nearby stations do not exist. Data from large events can then be employed to refine the calculated depth and location of smaller events. The essence of the technique is to correct the observed times of small events by noting the differences between the observed and the calculated times for a large event in the same area. The procedure is in routine use by several networks.

The combined effectiveness of location and depth in distinguishing earthquakes from explosions is impressive. More than 90 percent of all earthquakes either are under oceans or are at least 30 kilometers deep (or both). Most of the remaining earthquakes are of little interest because they are in countries that are unlikely to be testing nuclear weapons or in countries where clandestine testing would be impossible. For the U.S., of course, the U.S.S.R. is the country of prime interest. About 75 percent of the earthquakes in and near the U.S.S.R. are in the eastern part of the country near the Kamchatka Peninsula and the Kurile Islands. Almost all of the shocks in these areas either have a focal depth greater than 50 kilometers or are well offshore. It turns out that seismic events whose calculated position is on land in the U.S.S.R. or less than 25 kilometers at sea and whose calculated depth is less than 50 kilometers constitute only about .5 percent of the world's earthquakes. This amounts to about 100 earthquakes



PATHS OF SEISMIC WAVES are traced on a cross section of the earth. Body waves from an earthquake or an explosion travel through the crust and mantle along the curved paths labeled *P, S, pP* and *pS*. A *pP* wave is a compressional wave that is produced by the reflection of a *P* wave from the surface of the earth just above an earthquake or an explosion; a *pS* wave is a shear wave that results from the conversion of part of the compressional energy of an upward *P* wave into transverse energy as the *P* wave is reflected from the surface. Surface waves such as Rayleigh waves and Love waves diminish rapidly in amplitude with increasing depth. The hypocenter is the focal point of an earthquake or an underground explosion from which the waves radiate. The epicenter is the point on the earth's surface directly over the hypocenter.

per year with an m_b magnitude greater than 3.8 for which other seismic discriminants must be employed.

None of the measures we have discussed so far relies on the detailed characteristics of the waves radiated by earthquakes and explosions. Several powerful discriminants are based on those characteristics, in particular on the relative amounts of energy in waves of different types and periods. For example, a shallow earthquake generates 20-second Rayleigh waves with amplitudes at least several times greater than those of an explosion that releases the same amount of energy. In the notational practice of seismology the comparison of the two magnitudes is referred to as the $M_S : m_b$ ratio, that is, the ratio of long-period to short-period waves.

A second spectral discriminant is based on the observation that long-period *P* and *S* waves are rarely or never seen in association with explosions but one type or the other is routinely detected today by simple seismometers for most earthquakes that have a one-second *P*-wave magnitude of at least 4.5. More sophisticated seismic stations and more sophisticated analysis of the signals could lower the magnitude at which such waves can be detected.

A third distinction is that surface waves of the Love type are generated far more strongly by shallow earthquakes than they are by underground explosions, including even abnormal explosions. Still another characteristic feature of the seismic signal from explosions is that the first motion of the earth stimulated by *P* waves is always upward because the explosion itself is directed outward; the first *P*-wave motion in an earthquake can be either upward or downward.

An important factor contributing to the separation of earthquakes from explosions on an $M_S : m_b$ diagram is that *P* waves from the two kinds of events have different radiation patterns. Explosions radiate short-period *P* waves equally in all directions, whereas earthquakes have very asymmetric patterns. Hence most earthquake sources show a decrease of from .4 to one magnitude unit from the peak values when the *P*-wave amplitudes are averaged over pertinent radiation angles. A simple explosion does not initially radiate any shear waves; earthquakes typically generate large shear waves. As a result Rayleigh waves generated by many types of earthquakes have a larger amplitude than the corresponding waves generated by underground explosions of the same m_b.

There is a characteristic time for the formation of the source of a seismic event; the time is equal to the maximum source dimension divided by the velocity of source formation. The source dimension for earthquakes is the length of the break where most of the short-period energy is released; it is from three to 20 times greater, depending on the state of stress in the rocks, than the radius of the cavity and shatter zone of a comparable explosion. The velocity of source formation for earthquakes is from somewhat less to much less than the velocity of shear waves in the rocks surrounding the fault, whereas the relevant velocity for explosions is the velocity of shock waves in the rock, which is essentially the velocity of compressional waves. As a result of these differences in the size of the source and the velocity of source formation the characteristic times for earthquakes and explosions differ by a factor of from six to 40. It is therefore not surprising that differences are observed between the short-period *P*-wave spectra of earthquakes and explosions.

Observations of several U.S. explosions have demonstrated the existence of a phenomenon called overshoot. It is related to shock waves in strong rock, but it can be thought of as the equivalent of cavity pressure rising to high values followed by a decrease in pressure by a factor of four or five; the lower pressure is then maintained for many tens of seconds. Overshoot, when it occurs, provides additional *P*-wave spectral discrimination and augments discrimination by means of the $M_S : m_b$ ratio for larger events.

It was once thought that an explosion could not give rise to any Love waves at all. A phenomenon that was of great significance in thwarting President Kennedy's effort to achieve a comprehensive test-ban treaty in 1963 was the observation that many underground nuclear explosions at the U.S. testing site in Nevada, particularly those in hard rock, generated unmistakable Love waves. The failure of the qualitative criterion "No Love waves from explosions" (at a time when such quantitative criteria as the comparison of the magnitudes of long-period and short-period waves were not adequately established) left seismologists unable to guarantee their ability to distinguish the seismic waves of underground explosions from those of earthquakes.

The presence of Love waves in the Nevada tests has since been explained. What was not considered in the earlier analyses was the influence of the natural stressed state of the earth on the waves generated by an explosion. The creation of a cavity and its surrounding shatter cone by an underground explosion leads to the release of some of the natural stress, which in turn generates seismic waves equivalent to those of a small earthquake, including Love waves. The observed waves are a superposition of the waves from the explosion and from the release of the stress.

The release of natural stress also alters the amplitude of Rayleigh waves. The perturbation has never been large enough, however, to put in doubt the nature of an event identified by the ratio of long-period to short-period waves. Only rarely does the perturbation significantly affect the amplitude of *P* waves; it is not known ever to have changed the direction of their first motion. Moreover, if the magnitude M_S is determined from Love waves rather than Rayleigh waves, the ratio method ($M_S : m_b$) provides an excellent discriminant.

In short, if seismologists had done their homework thoroughly by 1963, the nations of the world might well have achieved a comprehensive test-ban treaty then. Today the release of natural stresses in the earth is significant only as a perturbing factor that must be taken into account when the yield of an explosion is estimated from Rayleigh waves.

Reports that earthquakes occasionally have $M_S : m_b$ values like those of explosions have been cited as a factor that might impede the effective monitoring of a comprehensive test ban. In analyzing a large set of earthquakes in all parts of the world and of underground explosions in the U.S. and the U.S.S.R. we found only one example of this kind of ambiguity. The focus of the event was far from the area in which the seismometer network gave its best results.

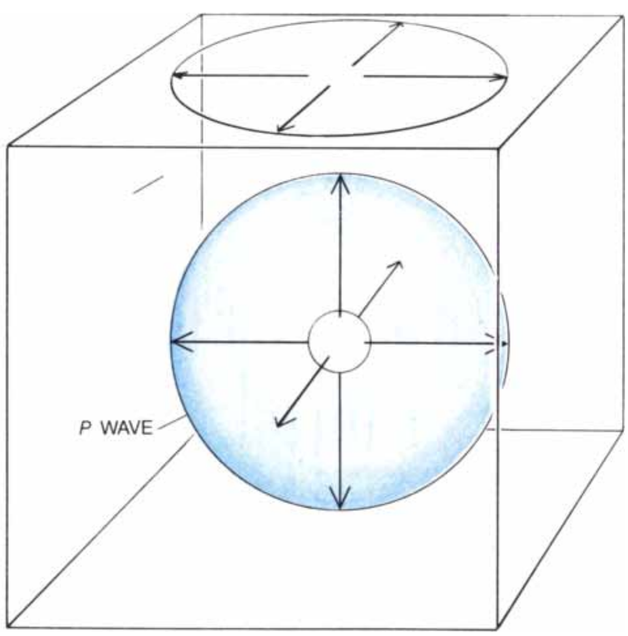
In 1972, at a meeting of the UN Committee on Disarmament, the U.S. submitted a list of 25 "anomalous" events that were said to be indicative of a problem in discrimination. In 1976 the 25 events were reanalyzed by one of us (Sykes) and two other seismologists, Robert Tathum and Donald Forsythe. It was established that about half of the events had $M_S : m_b$ values that put them clearly in the earthquake population. Most of the original magnitudes had been determined from only one or two stations, and much existing information had not even been consulted. When the records of other available stations were examined, the events ceased to be "anomalous."

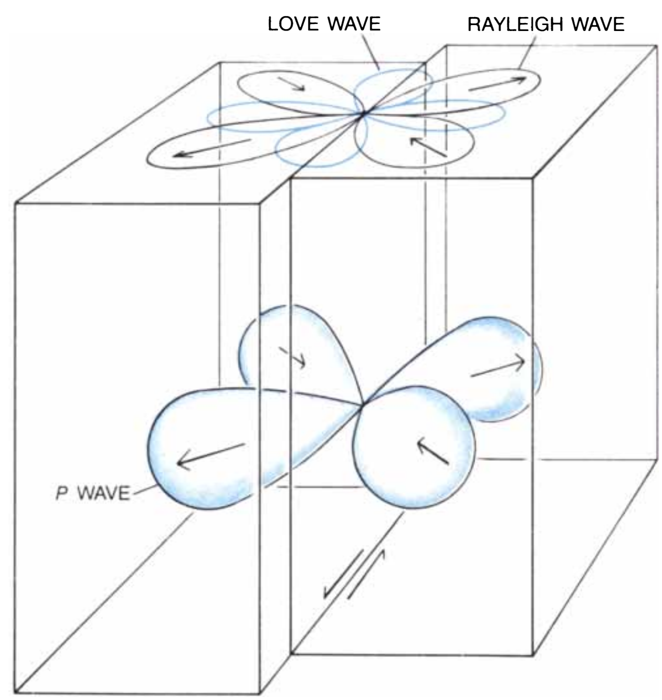
For the remaining problem events $M_S : m_b$ measurements based on 20-second Rayleigh waves gave values in the range characteristic of explosions. Several of these events were at depths of from 25 to 50 kilometers, where the possibility of nuclear testing can be excluded in any case, but the magnitude ratio nonetheless demanded explanation. It is known from seismological theory that certain types of earthquakes at these depths excite long-period Rayleigh waves poorly. The theory predicts, however, that Love waves and vibrations called higher-mode Rayleigh waves are in many instances vigorously generated in these circumstances. An analysis of recordings of the Love waves and the higher-mode Rayleigh waves identified several more of the problem events as earthquakes.

Only a single sequence of events at one place in Tibet remained as a problem. In that region underground nuclear testing is unlikely, but the nature of the events could not be determined with certainty from the magnitude ratios. We think the reason is that with the seismographic networks of the 1960's, when the events were recorded, Love waves could not be detected for small-magnitude events because they were obscured by background earth noise. New installations and new modes of data processing have greatly reduced the problem. If the same series of events or a similar series were to take place today, we think they would be identified unambiguously. Long-period seismographs in boreholes and routine digital processing of seismograms lead to a suppression of background noise and increase the detectability of many types of waves, including Love waves.

As it happened, the nature of the Tibetan problem sequence was resolved in spite of the inadequacies of the long-period data of the time. At several stations the first motion of the *P* waves was downward, which is not possible for an explosion. Hence the events must have been small earthquakes.

It seems reasonable to say that for the networks we shall describe below there should no longer be any problem events at m_b 4 or more. We know of no Eurasian earthquake with a one-second *P*-wave magnitude of 4 or more in the past 20 years whose waves are classified as those of an explosion. (Of course, numerous smaller Eurasian earthquakes during that period went unidentified because of inadequate data.) Furthermore, to our knowledge not one out of several hundred underground nuclear explosions set off in the same period radiated seismic waves that could be mistaken for those of an earthquake. Our experience indicates an extremely low proba-





RADIATION PATTERNS of the *P* waves resulting from an underground explosion (*left*) and an earthquake (*right*) are compared. The first motion of the *P* waves from an explosion is uniformly outward and hence is generally observed as an upward displacement at all seismic stations. The first motion of the *P* waves from an earthquake is outward in some directions and inward in others; the pattern of the waves at the surface depends on the orientation of the plane of the earthquake fault. In the comparatively simple case of a vertical strike-slip fault, shown here, the four-lobed radiation pattern observed at the surface for both *P* waves and Rayleigh waves is a simple projection of the three-dimensional *P*-wave configuration emanating from the hypocenter of the earthquake. The radiation pattern of the Love waves emitted by the same source is rotated by 45 degrees with respect to the surface pattern of the *P* waves and the Rayleigh waves.

bility that an event will remain unidentified when all the available techniques of discrimination are brought to bear.

No monitoring technology can offer an absolute assurance that even the smallest illicit explosion would be detected. We presume that an ability to detect and identify events whose seismic magnitude is equivalent to an explosive yield of about one kiloton would be adequate. It is often assumed that for the U.S. to subscribe to a comprehensive test ban it would require 90 percent confidence of detecting any violation by another party to the treaty. Developing a new nuclear weapon, however, generally requires a series of tests, and the probability that at least one explosion will be detected rises sharply as the number of the tests is increased. Moreover, a 90 percent level of confidence for the detection of even a single explosion probably is not needed. For a country seeking to evade the treaty the expected probability of detection would certainly have to be less than 30 percent, and perhaps much less, even if only one illicit test were planned. The test-ban agreements that have been considered over the years all include an "escape clause" through which a country could renounce its treaty obligations. Unless the probability of detection were very low, a country whose national interest seemed to demand a resumption of testing would presumably invoke the escape clause rather than risk being caught cheating.

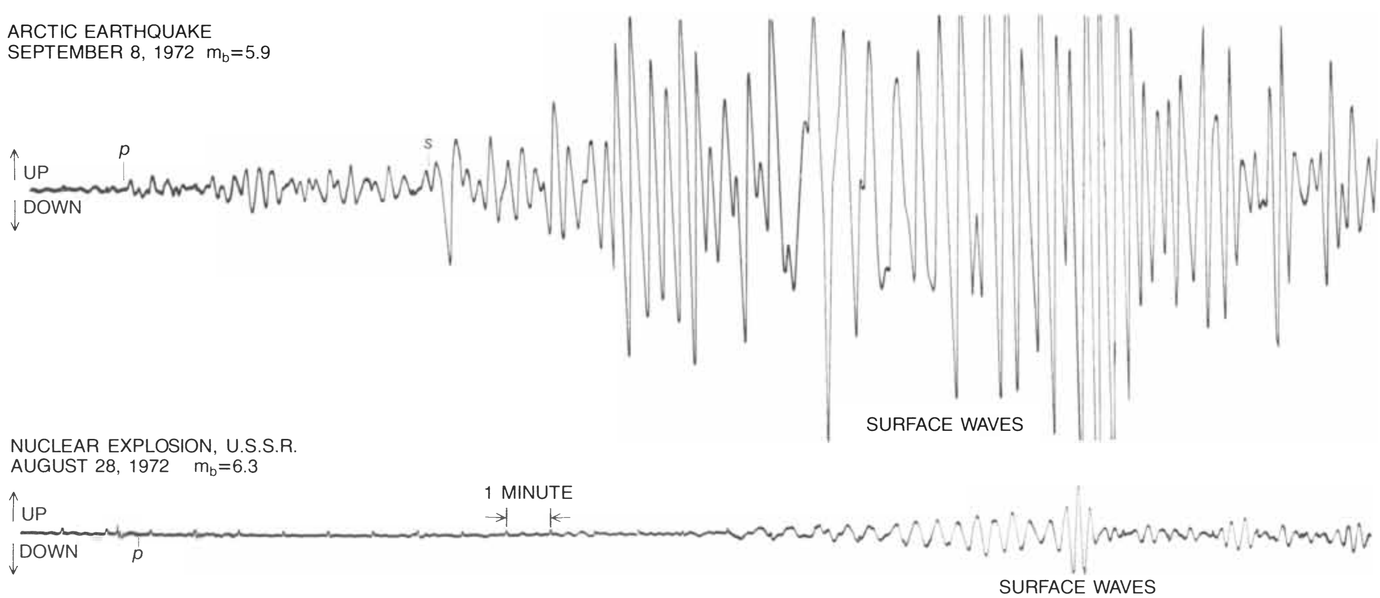
Given these standards of reliability for a monitoring system, it is possible to specify the size and the sensitivity of the seismic network that would be needed to verify compliance with a comprehensive test ban. Two kinds of network can be considered for maintaining seismic surveillance of the U.S.S.R. One network consists of 15 stations outside the borders of the U.S.S.R. In the second network the 15 external stations are supplemented by 15 internal ones.

The ultimate limit on the detection of seismic signals is imposed by microseisms, or random vibrations of the earth's surface. Most microseisms are induced by the earth's atmosphere and oceans. In order to detect a one-kiloton explosion in much of the U.S.S.R. a monitoring network would have to be able to recognize above the background noise any event with a short-period *P*-wave magnitude of 3.8 or more. In order to distinguish an explosion from an earthquake by comparing the long-period magnitude with the short-period one, the network would also have to be able to detect surface waves with an M_S magnitude of 2.5 or more. The network of 15 external stations could achieve these goals. Indeed, since almost all the seismic areas of the U.S.S.R. are along its borders, the external network would be sensitive to events of even smaller magnitude there. The mere detection of a seismic event in most areas of the interior would constitute identification of the event as an explosion.

The lower limit of one kiloton on the yield of an explosion that could be detected by an external network is based on the assumption that the coupling between the explosion and the seismic radiation is efficient and that the explosion was not set off during or soon after a large earthquake. If one must consider the possibility that a country would try to evade a test-ban treaty by decoupling, or muffling, an explosion and thereby reducing the amplitude of the emitted seismic signals, an improved network would be required. In principle such muffling could be done by detonating the explosion in a large cavity or by using energy-absorbing material in a smaller cavity. The former stratagem might reduce the seismic signal of an explosion by 1.9 magnitude units as measured by one-second *P* waves (that is, by m_b). The latter stratagem might bring a reduction of one unit.

The use of an oversize cavity is clearly the more worrisome possibility, but it could be attempted only in certain geologic formations: a salt dome or a thick sequence of bedded salt deposits. Few areas of the U.S.S.R. have deposits of salt in which the construction of a cavity large enough for decoupling a several-kiloton explosion would be possible. The maximum size of a cavity that could reasonably be constructed and maintained sets a limit of two kilotons on explosions that might be muffled in this way and escape detection by the 15-station external network.

Another way to reduce the amplitude of radiated seismic waves is by detonating an explosion in a low-coupling medium such as dry alluvium. The maximum thickness of dry alluvium in the U.S.S.R. sets a limit of 10 kilotons on explosions that might be concealed by this means, again assuming that only the 15 external stations were installed.



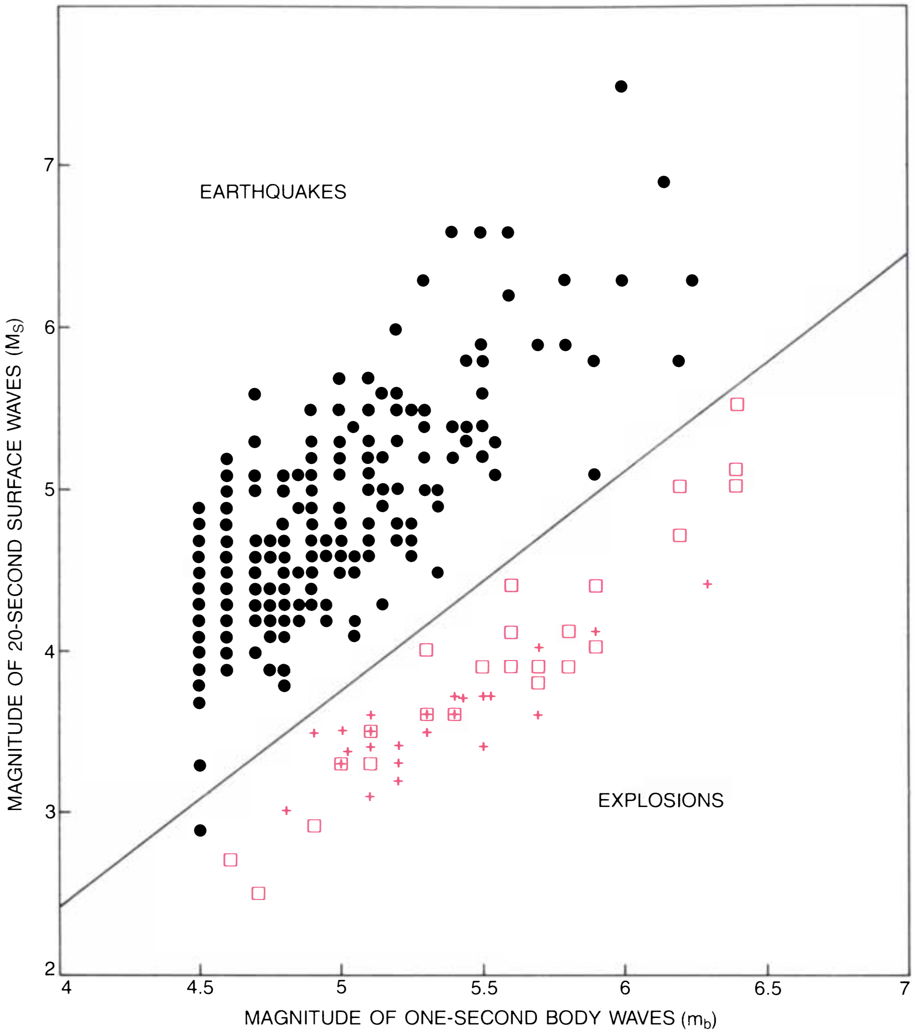
SEISMOGRAMS OF LONG-PERIOD WAVES from an earthquake in the Arctic near the U.S.S.R. (*top*) and an underground nuclear explosion in the U.S.S.R. (*bottom*) were recorded at a seismic station in Elath, Israel, roughly equidistant from the two events. The short-period body waves generated by the two shocks were observed to have almost the same magnitude. The magnitude of the long-period Rayleigh waves recorded in these traces, in contrast, is clearly much greater for the earthquake than it is for the explosion. The ratio of long-period surface waves to short-period body waves has been shown to be a reliable criterion for distinguishing the seismic waves of earthquakes from those of explosions. In addition the *P* wave of the explosion has more high-frequency energy than the *P* wave of the earthquake. The *S* wave of the earthquake is large, whereas that of the explosion is small and not easily identified in the seismogram.

Another possible drawback of an exclusively external network should be mentioned. Confusion could arise when signals from two or more earthquakes reached a station simultaneously. The effect would be most troublesome when the long-period waves from a small event in the U.S.S.R. arrived at the same time as similar waves from a much larger earthquake elsewhere in the world. Under these circumstances it might be difficult to establish with certainty by comparing M_S with m_b the identity of the event in the U.S.S.R. With a network of 15 external stations there would be a few events per year in which the smaller earthquake was in the territory of the U.S.S.R. or within 25 kilometers of its borders and at a depth of less than 50 kilometers.

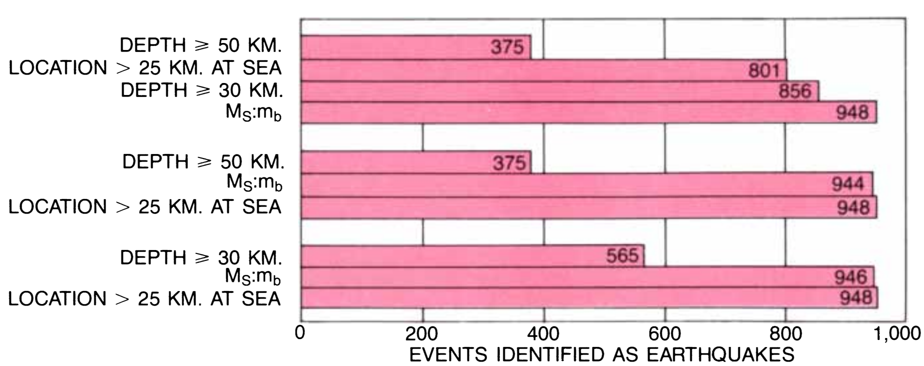
A monitoring network made up of 15 seismographic stations outside the U.S.S.R. and 15 inside it would largely eliminate the problem of coincident earthquake signals and would greatly reduce the maximum yield of an explosion that might escape detection, even if decoupling were attempted. The internal monitoring stations would be simple unattended ones, with the capability of measuring vertical ground motion and two orthogonal components of horizontal motion, so that the distance and direction of a nearby event could be estimated from the data of a single short-range station. With such a network in place, and assuming that muffling was attempted in the presence of normal earth noise, the largest explosion that would have a 30 percent chance of escaping detection in any setting except a salt dome would be .5 kiloton.

For salt domes the main area of concern in the U.S.S.R. is the region north of the Caspian Sea. Our hypothetical network has three stations there. Even a small explosion in a large salt-dome cavity would emit certain *P* and *S* waves with an amplitude large enough to be detected by nearby stations. Furthermore, detection by even one of the stations would immediately identify the event as an explosion because the area has no natural seismic activity. As a result evasion would not be likely to be attempted at a yield greater than one kiloton even in the salt-dome area.

A possible strategy for evasion that has been mentioned from time to time is the one of hiding the seismic signal of a nuclear explosion in the signal of a large earthquake, which might be near the site of the explosion or far from it. For the U.S.S.R. the only credible possibility is a distant earthquake because the only possible testing sites where earthquakes are frequent enough to make the effort worth while are on the Kamchatka Peninsula and in the Kurile Islands. Clandestine testing there is not likely because seismic activity in the area can be moni-



CLEAR DISTINCTION between earthquakes and explosions is evident in this plot of the magnitude of long-period surface waves (M_S) against that of short-period body waves (m_b). The 383 earthquakes represented by the black dots were compiled from a set of all the earthquakes recorded worldwide in a six-month period that had an m_b value of 4.5 or more and a focal depth of less than 30 kilometers. (There are fewer dots than earthquakes because the magnitudes of some of them coincided.) The colored squares designate underground explosions in the U.S. and the colored crosses underground explosions in the U.S.S.R. Only one earthquake falls within the explosion population, as defined by the straight line separating the two groups of events. This single event, which had the smallest magnitude of any of the earthquakes in the survey, took place in the southwest Pacific Ocean, a region where the sensitivity of the network of seismic stations is poorer than it is in most of the Northern Hemisphere. The m_b values were adjusted to take into account regional variations in the amplitudes of short-period waves.



METHODS OF DISTINGUISHING earthquakes from explosions were tested by applying the methods to all the earthquakes with a magnitude of 4.5 or more recorded during a 162-day period in 1972. The sample consisted of 948 events. Many of them could be classified as earthquakes (rather than explosions) by their location or their depth. The remaining events could be classified by comparing the magnitude of long-period surface waves with the magnitude of short-period body waves (the ratio $M_S : m_b$). The sequence in which the tests were applied affected the efficiency of the procedure, but all events were identified regardless of the sequence.

tored in detail from stations in Japan and the Aleutian Islands. Indeed, ocean-bottom seismometers and hydroacoustic sensors could be placed just offshore.

The first defense against evasion by the masking of a test in a large earthquake is the questionable feasibility of the subterfuge. Unless the evader maintained several testing sites the number of opportunities per year for clandestine testing would be quite limited. In addition the evader would have to maintain his weapons in constant readiness for firing. To attain the evasion capability given below he would have to set off an explosion within 100 seconds of the time of arrival of the short-period waves of the earthquake. He would have to estimate the maximum amplitude and the decay rate of the earthquake waves with high accuracy, and he would have to be certain of the amplitude of the *P* waves generated by the explosion to within .1 magnitude unit. Even after taking these precautions the evader would have to accept a high probability that the event would be detected by at least one monitoring station and a small probability that three stations would detect it. He would also have to install and operate his testing site (including a large cavity) and his own seismological network in total secrecy over a period of years.

In contrast to these daunting requirements for successful evasion, the only requirements for a monitoring nation are to operate a network of high-quality seismic stations and to process the data with determination. Against a network of 15 external stations and 15 internal ones the only effective evasion schemes at yields of one kiloton or more would require both decoupling and hiding the explosion signal in an earthquake.

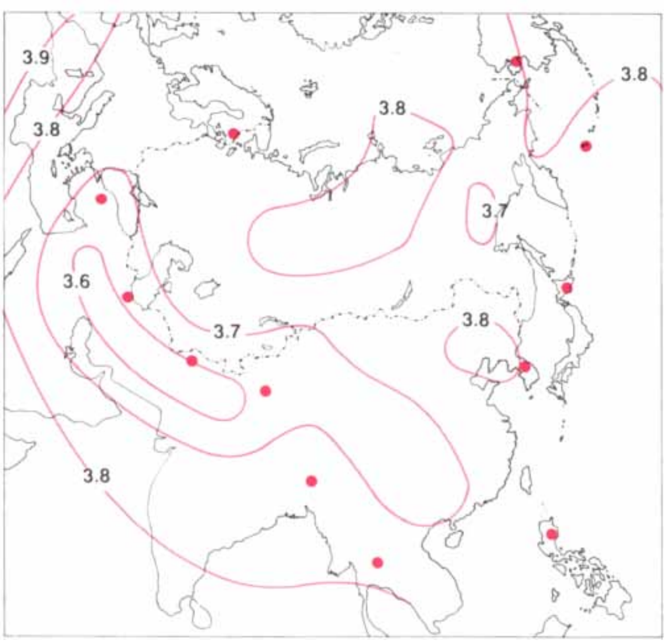
The issues relating to the monitoring of a comprehensive test ban can be summarized as follows. The understanding of seismology and the testing of seismometer networks are sufficiently complete to ensure that compliance with a treaty could be verified with a high level of confidence. The only explosions with a significant likelihood of escaping detection would be those of very small yield: less than one kiloton provided the monitoring system includes stations in the U.S.S.R.

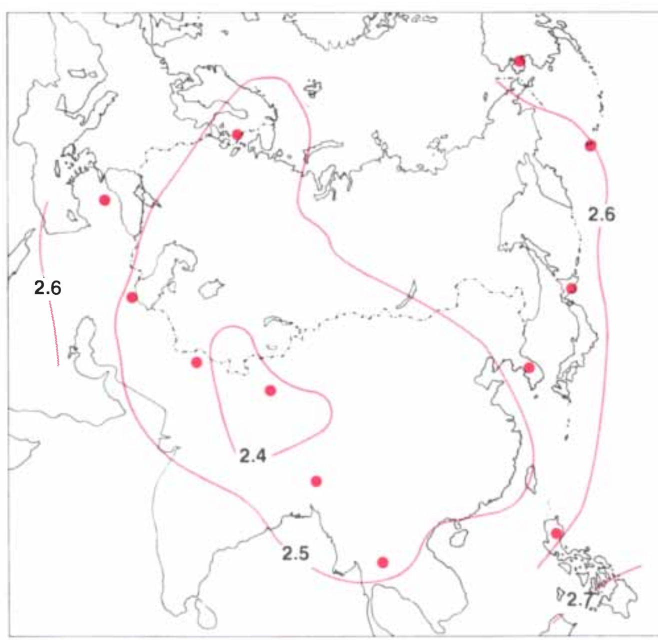
It is important to view the question of yield in the context of the nuclear weapons that have been tested up to now. The ones that ushered in the nuclear age in 1945 had a yield of from 15 to 20 kilotons. Yields increased rapidly to the point where the U.S.S.R. tested a 58,000-kiloton weapon in 1961. The largest underground explosion had a yield of almost 5,000 kilotons. Unclassified reports place the yield of the weapons carried by intercontinental missiles in the range from 40 to 9,000 kilotons. The yields of underground explosions that might go undetected or unidentified under a comprehensive test ban are therefore much smaller than those of the first nuclear weapons. If the threshold of reliable detection and identification is one kiloton, that is only one-150th of the limit specified by the Threshold Test Ban Treaty of 1976.

From the viewpoint of verification a comprehensive test ban would actually establish the equivalent of a very low threshold, since weapons of extremely low yield could be tested underground without the certainty of being detected and identified. A treaty that imposed a threshold near the limit of seismological monitoring capability might therefore be considered an alternative to a comprehensive test ban. Such a treaty might be preferable to the present quite high threshold, but it would have the disadvantage that arguments could arise over the exact yield of tests made near the threshold. Indeed, the judgment of whether or not a test has taken place will always be less equivocal than an exact determination of yield.

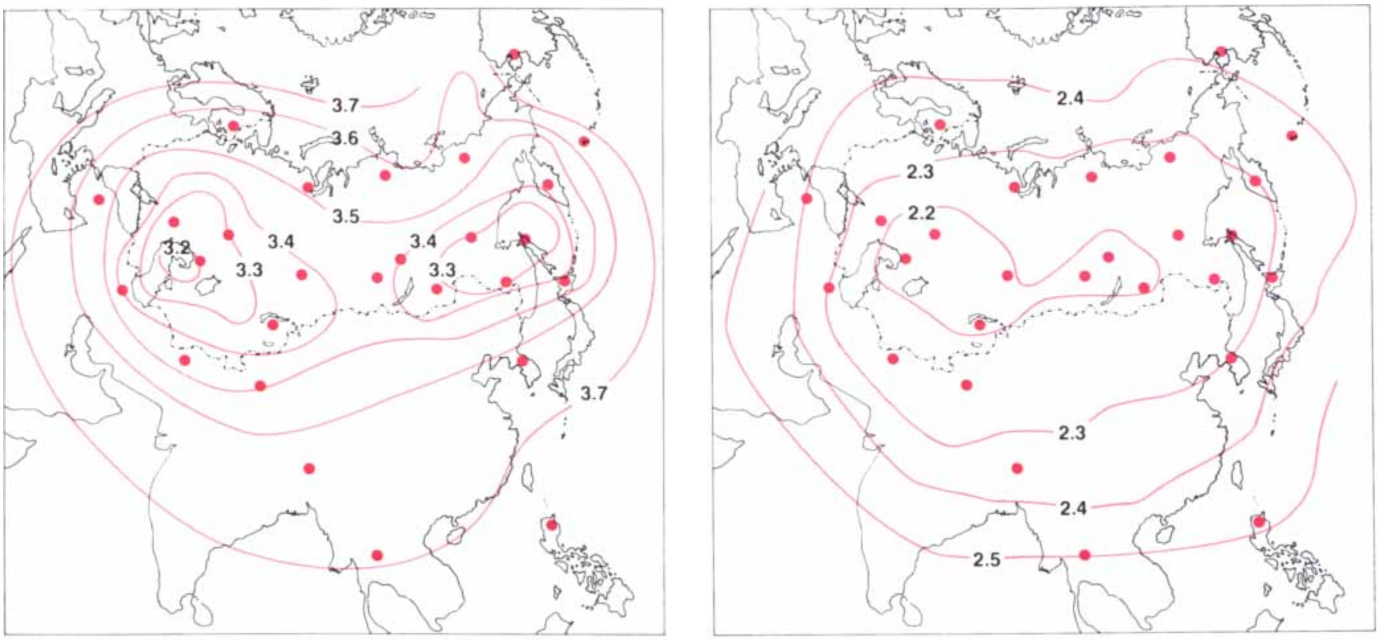
In recent years there have been reports that the U.S.S.R. may have repeatedly violated the 1976 treaty by testing devices with a yield greater than the 150-kiloton limit. Such reputed violations were recently cited as evidence that the threshold treaty, which has not been ratified by the U.S. Senate, is not verifiable and should be renegotiated. On the basis of our analysis we conclude that the reports are erroneous; they are based on a miscalibration of one of the curves that relates measured seismic magnitude to explosive yield. When the correct calibration is employed, it is apparent that none of the Russian weapons tests exceed 150 kilotons, although several come close to it.

Observations at the Nevada Test Site (NTS), where American nuclear-weapons tests are held, indicate there are lin-





THRESHOLDS OF DETECTION for seismic events in the Eastern Hemisphere are delineated by the two sets of contours drawn on this pair of maps for a proposed network of 15 seismic stations established outside the U.S.S.R. Colored dots give the location of 12 of the 15 stations; three others are off the maps. The number on each contour indicates that an event of that magnitude or larger has at least a 90 percent probability of being detected by four or more stations. The contours on the map at the left represent the detection thresholds for short-period body waves and those on the map at the right the detection thresholds for long-period surface waves. On these maps and the ones on the opposite page the only seismic noise taken into account is the microseismic noise generated by the atmosphere and oceans.



IMPROVED DETECTION THRESHOLDS for seismic events in the Eastern Hemisphere are delineated by the two sets of contours drawn on this pair of maps for a proposed network of 30 seismic stations: 15 outside the U.S.S.R. and 15 inside. For most of the U.S.S.R. the effect of adding the 15 interior stations would be to lower the detection threshold for short-period body waves from a magnitude of 3.8 to one of 3.4 (*left*); the corresponding effect for long-period surface waves would be to lower the detection threshold from a magnitude of 2.6 to one of 2.3 (*right*). The interior stations would also provide more accurate information on the focal depth of a seismic event.

ear correlations between the logarithm of the explosive yield and the two magnitude values, M_S and m_b, for explosions with yields greater than 100 kilotons. When the measured M_S and m_b values of explosions at the Russian test site near Semipalatinsk are inserted into the NTS formulas, however, the resulting estimates of yield given by m_b are more than four times as great as those given by M_S. For explosions in hard rock at many test sites estimates of yield based on the NTS M_S formula have invariably agreed with actual yields, whereas estimates based on the NTS m_b formula have sometimes been in drastic disagreement with the actual yield.

A strong correlation has been found between m_b values measured at individual stations and *P*-wave travel times to these stations. The U.S.S.R. routinely publishes seismological bulletins that include *P*-wave arrival times of earthquakes, and it is straightforward to interpret the times for stations in central Asia in terms of the expected pattern of m_b values near Semipalatinsk. From an analysis of the *P*-wave signals it is predicted that the m_b value for an explosion at Semipalatinsk is 40 percent greater than an equivalent explosion at NTS. This is the same correction that must be applied to the curve relating m_b to yield at NTS to make the m_b estimates of the yield of Russian explosions consistent with the M_S estimates. Thus two modes of analysis lead to the conclusion that there is an essentially universal relation between M_S and yield whereas the curve relating m_b to yield must be calibrated for each test site.

A comprehensive treaty would have an additional advantage over a low-threshold treaty: all technological uncertainties would work against the potential evader. A country planning a surreptitious nuclear test could not know the exact seismic-detection capability of other nations or the exact magnitude of the seismic waves that would be generated by his test. A ban on nuclear explosions of all sizes would also have the important conceptual value that nuclear weapons, no matter what their size, would be recognized as inherently different from conventional weapons.

It is sobering to consider how the state of the world would differ if a full test ban had been achieved in 1963. The number of nuclear weapons has grown tremendously since then and is now estimated at from 50,000 to 100,000. The loss of life and the social damage that would be inflicted in a major nuclear exchange are vastly greater than they were in 1963. Furthermore, both the U.S. and the U.S.S.R. are less secure now than ever before, not because of any failure to develop arms but because of the growing stockpiles of weapons and the inability of any nation to defend itself against nuclear attack.

A comprehensive test-ban agreement should not be regarded as a substitute for disarmament. Meaningful reductions in the nuclear threat must include a continuing and serious process of arms control; in this process, however, a comprehensive test-ban treaty could have an important part. The problems of negotiating such a treaty are overwhelmingly political rather than technical and must be recognized as such.

Before the suspension of negotiations between the U.S., Britain and the U.S.S.R. in 1980 tentative agreement had been reached on a number of issues. All three nations agreed that a test-ban treaty would include a prohibition of all tests of nuclear weapons in all environments, a moratorium on peaceful nuclear explosions until arrangements for undertaking them could be worked out, provisions for on-site inspections, a mechanism for the international exchange of seismic data and the installation of tamperproof seismic stations by each country in the territory of the others. The proposed treaty would have a term of three years. The agreements on the long-standing issues of on-site inspection, peaceful explosions and the placement of monitoring stations in each country represented important breakthroughs. It would be a setback for the cause of international security if this hard-won ground were now lost.

For many years the stated policy of the U.S. has emphasized the desirability of a complete test ban if verification could be ensured. The policy was not fundamentally altered by the recent decision of the Reagan Administration to put off further negotiations on the test ban. On the contrary, it was reported that the Administration still supports the ultimate goal of a comprehensive ban on nuclear testing but has doubts about the efficacy and reliability of seismic methods of verification. As we have attempted to show here, there can be no substance to such doubts.

Calcium in Synaptic Transmission

A current of calcium ions triggers the passage of signals from one nerve cell to another. The process is studied in a synapse (a neuronal junction) hundreds of times a synapse's usual size

by Rodolfo R. Llinás

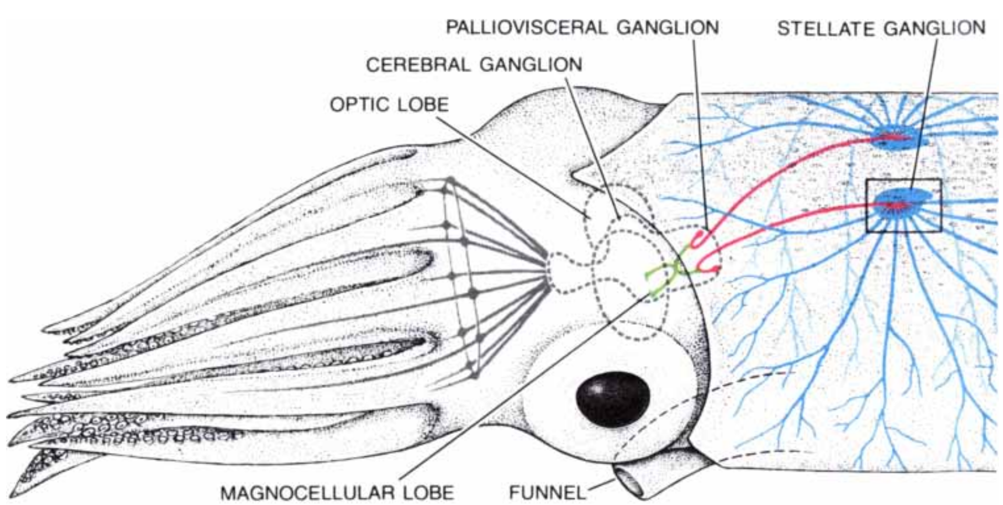
Synapses are the sites at which neurons, or nerve cells, communicate with one another. At an earlier stage in the life of an organism they are important in determining how the nervous system develops. (Synapses form at the tips of the fibers that sprout from the body of a neuron.) It seems certain that much of the brain's ability to regenerate after injury and much of an organism's ability to learn will ultimately be explained in terms of the function of synapses. Furthermore, it is becoming clear that most of the diseases of the brain and many psychiatric disorders result from a disruption of synaptic communication or are associated with such a disruption. The synapse is the weakest link in brain activity: it is the first part of a neuronal chain to be fatigued by a high level of message transmission, and it is the site of action for most of the drugs that affect the brain, including addictive substances as well as therapeutic ones from aspirin to barbiturates. For all these reasons a detailed understanding of every aspect of synaptic transmission is essential if we are to understand how the brain works and how the brain can malfunction.

Here I shall be concerned in particular with the aspect of transmission called depolarization-release coupling. In synaptic transmission one neuron (the presynaptic cell) releases a biologically active substance (a neurotransmitter), which evokes a response, either excitatory or inhibitory, in a second neuron (the postsynaptic cell). The release of the transmitter is in essence a process of secretion, a process shared by cells throughout the evolutionary sequence. In virtually every known instance secretion is accomplished by exocytosis, a mechanism in which vesicles, or membranous sacs, inside the cell fuse with the membrane that surrounds the cell. The fusion everts the contents of the vesicles into the extracellular environment.

A neuron evidently releases neurotransmitter in the same way. The release is known to be stimulated when the membrane of the presynaptic neuron at the synapse loses its electrical polarization. The step in synaptic transmission that concerns me here can be expressed, then, by a question: How does the depolarization of the membrane lead to the release of neurotransmitter, so that the neuron can act on the next cell in the neuronal chain? It turns out that the connection between the electrical activity of the cell and the release of neurotransmitter is not direct; an essential intermediary is the calcium ion.

When a neuron is at rest, there is an electric potential, or voltage difference, of some 70 millivolts between the inside of the cell and the outside. The voltage is negative inside, and so the cell membrane is said to have a polarization of −70 millivolts. When the membrane is depolarized, the voltage difference diminishes. Indeed, in the type of depolarization called the action potential a reverse potential develops with a value of from +10 to +30 millivolts. The reversal persists for only about a millisecond.

The precise mechanism by which depolarization arises is inconsequential as far as the subsequent release of neurotransmitter is concerned. In neurons such as sensory receptors that respond to stimuli (a burst of sound, a touch on the skin) impinging directly on the surface of the cell, the depolarization is caused by the energy of the stimulus itself. In such a neuron the depolarization can be a subtle change in potential that lasts for many seconds. In neurons in the brain the depolarization is generally caused by the signals each cell receives at the synapses it makes with other neurons. Here the depolarization is typically a rapid modulation in voltage that lasts for only a few milliseconds. If it is sufficiently great, it can induce the cell to generate its own action potential. In ev-



CHAIN OF GIANT NERVE CELLS on each side of the midline in the squid *Loligo pealii* includes a giant synapse: a site of signaling from one nerve cell to another that is .7 millimeter long, or several hundred times the size of a typical synapse. It is so large that it can be impaled by several microelectrodes and still transmit signals while its potential (that is, its voltage) is recorded or even varied experimentally. Here the chain of nerve cells that includes the synapse is shown in color. The first cell, the first-order giant neuron (*green*), is in the magnocellular lobe of what amounts to the brain of the animal. The neuron gathers sensory data from organs such

ery case, however, the effect of a wave of depolarization arriving at the membrane of the presynaptic terminal of a neuron is the same: it causes transmitter to be released.

Depolarization is nonetheless not sufficient in itself to cause the release of transmitter. In addition a supply of calcium ions must be present in the extracellular environment. Depolarization seems to be the means by which an inward current of calcium ions is induced to flow through the membrane of the presynaptic terminal. In this respect too synaptic transmission resembles other known secretory processes. In every secretory process in which the test has been made secretion is triggered by an increase in the concentration of calcium inside the secretory cell. The increase is the result of the entry of calcium from the external environment or of the release of calcium from internal stores. It seems likely that in neurons the entering calcium ions promote the fusion of special intracellular vesicles (synaptic vesicles) into the presynaptic membrane. The vesicles are stationed near the inner surface of the membrane at the site of transmitter release and are filled with the transmitter substance.

Among the difficulties that arise in attempts to advance our understanding beyond these basic points, one problem is implacable: synapses are small. So far the only measuring technique fast enough and precise enough to serve in the study of events in a functioning synapse is the recording of its electrical activity. Yet the diameter of a presynaptic terminal in most vertebrate and invertebrate species is typically from .1 micrometer to five micrometers. The microelectrode with which one can probe the synapse's electrical properties has a diameter of about .5 micrometer. Piercing the presynaptic terminal without damaging the synapse is quite difficult, and so is keeping the electrode in place as the experiment proceeds.

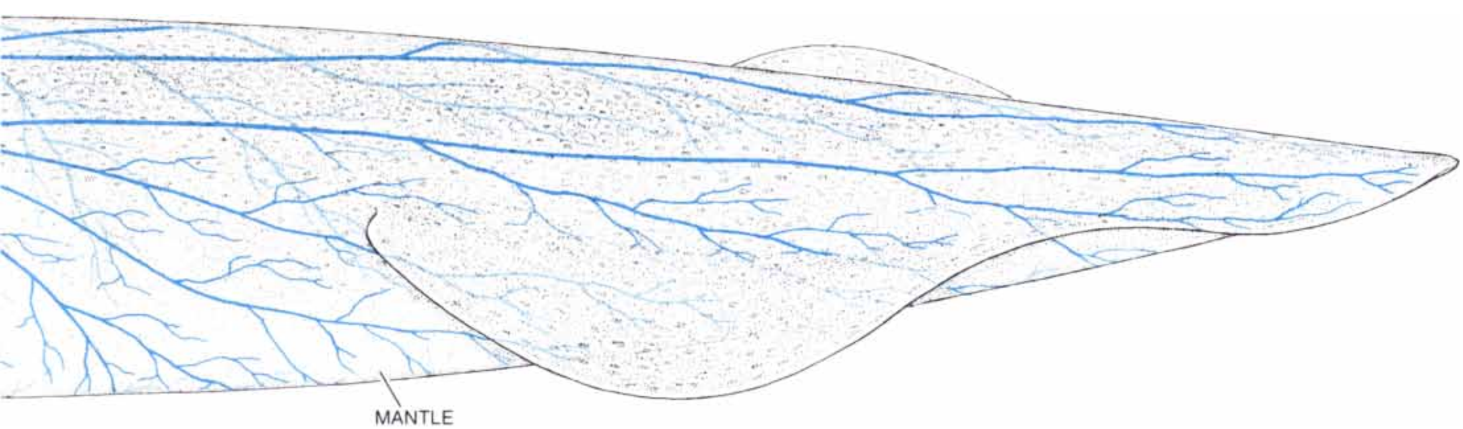
Fortunately for the experimenter some synapses are considerably larger than the average. They include certain synapses found in ganglia (clusters of neurons) in mollusks and crustaceans and the synapse between the neurons called Mauthner cells in the brain stem of fishes and some amphibians. The Mauthner cells govern the movement of the tail. Among large synapses, however, the most notable example is a giant synapse of the squid. It is some 700 micrometers long. Because of its size and accessibility it has yielded most of what is now known directly about the relation of depolarization to the release of neurotransmitter.

The presence of a giant synapse in the squid was first reported in 1935 by J. Z. Young of the Marine Biological Association of the United Kingdom at Plymouth. The synapse was included in his description of a chain of giant nerve cells in the squid; the chain consists of three axons, or nerve fibers, and the neuronal cell bodies from which they arise. The axons' diameters are so great that some of them had been mistaken for blood vessels; Young identified them as nervous tissue on the basis of their ability to conduct an action potential. The chains on each side of the squid's body are now known to govern the animal's flight response and capture of prey.

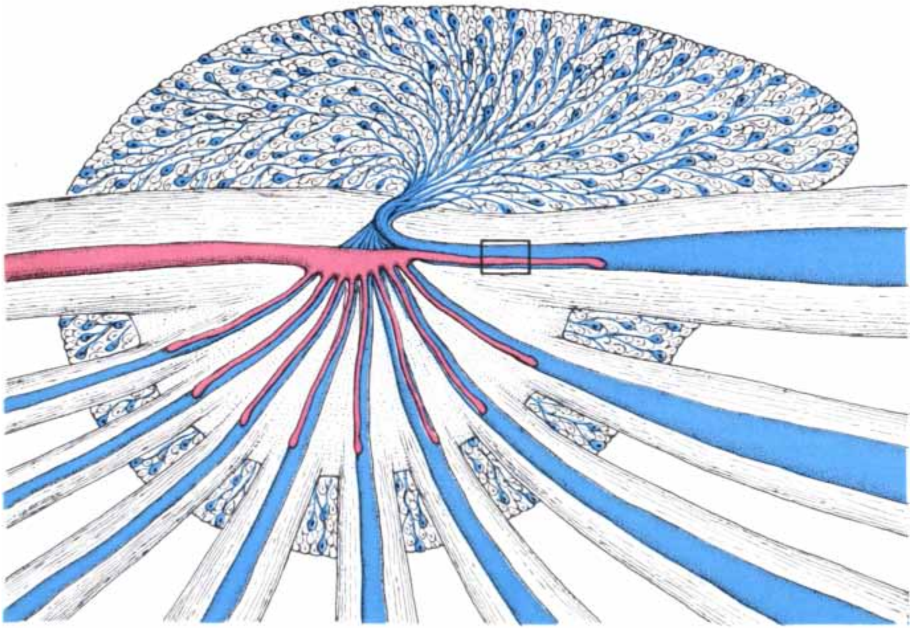
The first nerve cell in the chain is called the first-order giant neuron. It is in the magnocellular lobe, an assemblage of neurons in what amounts to the animal's brain. Its cell body is some 150 micrometers in diameter, which makes it many times larger than the largest nerve cells in the human brain. With its extensive set of dendrites (tubular extensions of the cell body) it is 800 micrometers long. By means of these dendrites it gathers signals that originate in such places as the eyes, the vestibular organs and the tentacles. It is in effect a single-cell computer that assesses danger to the animal. The results of its calculations are transmitted along the membrane of its axon. As if to ensure the synchrony of such signals, the axons of the first-order giant neurons on each side of the body fuse for a short distance at the midline.

The second cell in the chain, the second-order giant neuron, is in the palliovisceral ganglion behind the magnocellular lobe. The second-order cell is 100 micrometers in diameter. Its axon curves through the ganglion on a trajectory that directs it toward the squid's mantle. Along the curve it comes in contact with the axon of the first-order giant cell, which establishes a synapse with it. Then the second-order axon continues on to the center of the stellate ganglion, on the inner surface of the mantle. There it broadens and divides into a set of from eight to 10 terminal branches that look like fingers radiating from the palm of a hand. The branches vary in size: the thinnest are 25 micrometers in diameter and the thickest are twice as big.

The third cell in the chain is actually a set of hundreds of cells, each of which is about 50 micrometers in diameter. The



as the eyes and the tentacles, and it dispatches neural signals along its axon (nerve fiber) to the palliovisceral ganglion. There the axon synaptically contacts the axon emitted by the second cell in the chain, the second-order giant neuron (*red*). In the stellate ganglion at the flank of the animal's mantle the second-order axon synaptically contacts a set of third-order axons (*blue*) that trigger the mantle's contraction. The contraction sends water jetting out of the animal's funnel, propelling the animal from danger. The synapse between the second-order axon and the largest third-order axon, the one passing almost rearward through the muscle tissue of the mantle near the midline, is the giant synapse studied by the author and his colleagues. The squid is shown about 1.5 times life size; the stellate ganglion on the left side of the body (the ganglion enclosed in this drawing by a rectangle) is further enlarged in the top illustration on the next page.



IN THE STELLATE GANGLION the second-order axon (*red*) branches into a series of fingerlike extensions. Each one is a presynaptic terminal: it makes synaptic contact with a third-order axon (*blue*), which arises from the fusion of the axons emitted by several hundred neurons in the ganglion. The largest third-order axon, shown leaving the ganglion toward the right at the center of a bundle of smaller axons, forms the postsynaptic part of the giant synapse. Beyond the synapse the axon widens to .5 millimeter, or a hundred times the diameter of a typical axon. The part of the giant synapse enclosed by the rectangle is enlarged in the illustration below.

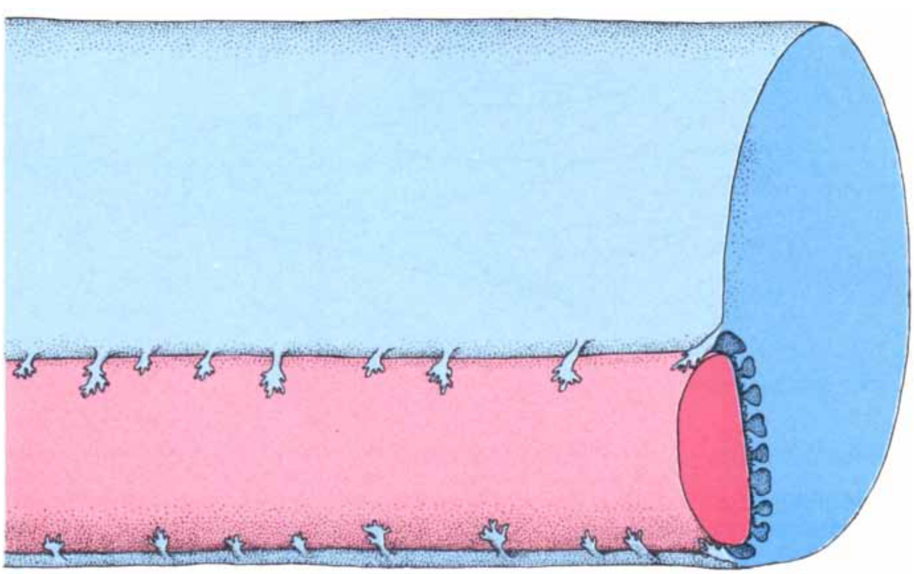
cells are organized into from eight to 10 sets, and in each set all the axons fuse to form a single giant axon. Hence the stellate ganglion gives rise to from eight to 10 giant axons. The thinnest of them is about 50 micrometers in diameter; it receives signals through a synapse with the thinnest of the terminal branches of the second-order axon, then it widens somewhat and proceeds away from the stellate ganglion to innervate muscle tissue in the mantle near the ganglion. The thickest of the third-order axons is about 200 micrometers in diameter; it receives signals through a synapse with the thickest of the terminal branches of the second-order axon, then it widens to a diameter of about 500 micrometers and extends to innervate muscle tissue in the most distant part of the mantle. The other third-order axons have intermediate destinations.

The rudiments of how this network functions can be inferred from its structure. When stimuli reaching the first-order neuron exceed some threshold that signifies alarm, the cell activates the second-order axon, which in turn produces a synchronous excitation in all the third-order axons. The speed with which a wave of depolarization propagates along a nerve fiber depends on the diameter of the fiber; since the longest of the third-order axons are also the thickest, the synchronization of the signals is preserved and all the muscles of the mantle contract simultaneously. The contraction forces the water in the mantle out through the funnel near the head of the animal. Thus the squid can escape from danger by jet propulsion.

The giant synapse we have studied is the one between the second-order axon and the largest third-order axon. When the synapse is examined in detail, it is found that the membrane of the presynaptic terminal (the second-order axon) is smooth. In contrast, the membrane of the postsynaptic terminal (the third-order axon) has multiple branchings that divide repeatedly and end as a net of thorn-shaped extensions. In the giant synapse as many as 5,000 such extensions face the presynaptic membrane. This structure is unusual. In a typical synapse the postsynaptic terminal is smooth, whereas the presynaptic terminal consists of protuberances at the end of an axon that are called synaptic boutons (from the French word for button).

Nevertheless, a microscopic examination of the places in the giant synapse where the presynaptic membrane abuts a postsynaptic thorn shows morphology quite typical of a synapse. Inside the presynaptic terminal at such places there are synaptic vesicles and two other intracellular structures that are notably abundant at synapses: mitochondria, the organelles that make energy available to the cell, and subcysternal systems, which consist of folds of intracellular membrane. (The function of the latter is not known.) In the same areas the postsynaptic thorn has a thickened membrane, which is recognized in the typical synapse to be the site where receptors are found. The receptors are molecules embedded in the postsynaptic membrane that react with arriving molecules of neurotransmitter.

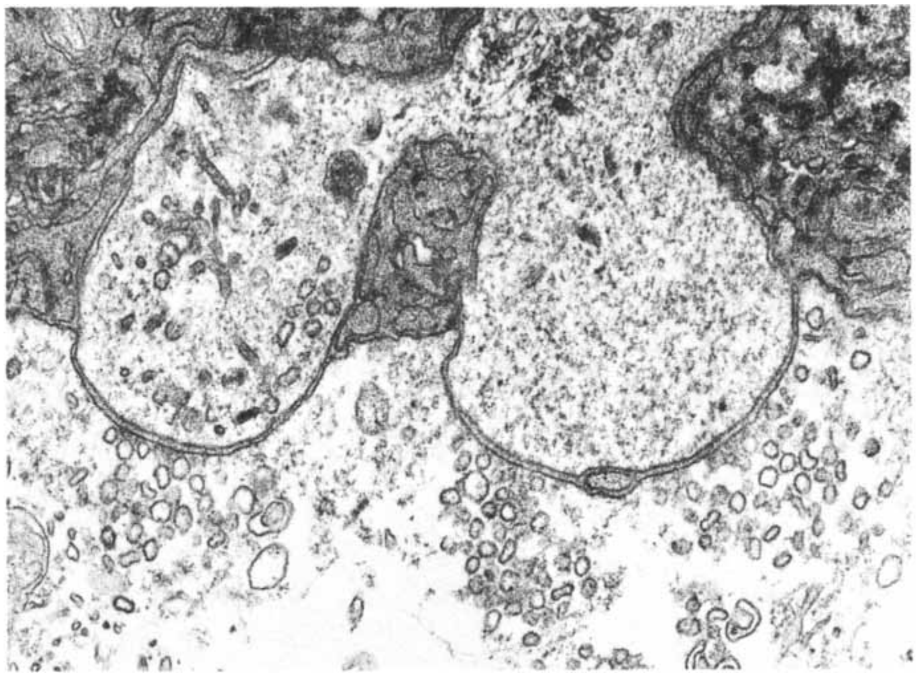
It was in the giant synapse of the squid that it was first demonstrated directly that the transmission of a neural signal across the synapse requires the release of a transmitter substance. This intermediate chemical step (first proposed in the course of research on the union between nerve and muscle) was established in the 1950's and the early 1960's through a number of studies by Theodore H. Bullock, Susumu Hagiwara, Ichiji Tasaki, Noriko and Akira Takeuchi, Ricardo Miledi and C. R. Slater. The studies capitalized on the size of the giant synapse: the workers were able to position the tip of a microelectrode in the presynaptic terminal and the tip of a second microelectrode in the postsynap-



GIANT SYNAPSE is made up of two apposed membranes. The postsynaptic membrane of the third-order axon has some 5,000 spiny extensions. They face the presynaptic membrane.

tic axon so that the voltage across the membrane of each terminal could be recorded simultaneously. In addition the workers could vary the concentration of ions or drugs in the solution in which they had isolated the synapse (along with a few centimeters of the second-order and third-order axons). Thus they could study the role of the ions and the action of the drugs.

It emerged from the various studies that when an action potential propagating along the second-order axon reaches the presynaptic terminal, it causes a depolarization of the postsynaptic axon after a delay of about a millisecond. On the other hand, an action potential at the postsynaptic terminal (induced by electrical stimulation of the third-order axon) has no effect on the presynaptic terminal. It further emerged that the degree of postsynaptic depolarization increases with the amplitude and the duration of the presynaptic action potential. Yet the presynaptic action potential becomes incapable of causing postsynaptic depolarization if all the calcium is removed from the solution in which the synapse is bathed. The combination of all these findings (showing, among other things, synaptic delay and unidirectionality) was strong evidence that signals are not transmitted electrically across the giant synapse. They are carried by a chemical messenger.



ELECTRON MICROGRAPH of the giant synapse suggests how it transmits signals. The presynaptic terminal of the synapse is at the bottom; it is abutted by two bulbous protrusions arising from a single spiny extension of the postsynaptic axon. Inside the presynaptic terminal at the places of abutment are accumulations of saclike organelles. They are synaptic vesicles, which are thought to contain a neurotransmitter: a substance released by the presynaptic terminal that acts on the membrane of the postsynaptic terminal. The signaling is thus a secretory process. The electron micrograph was made at an enlargement of 74,000 diameters by David W. Pumplin of the University of Maryland at Baltimore School of Medicine and Thomas S. Reese of the National Institute of Neurological and Communicative Disorders and Stroke.

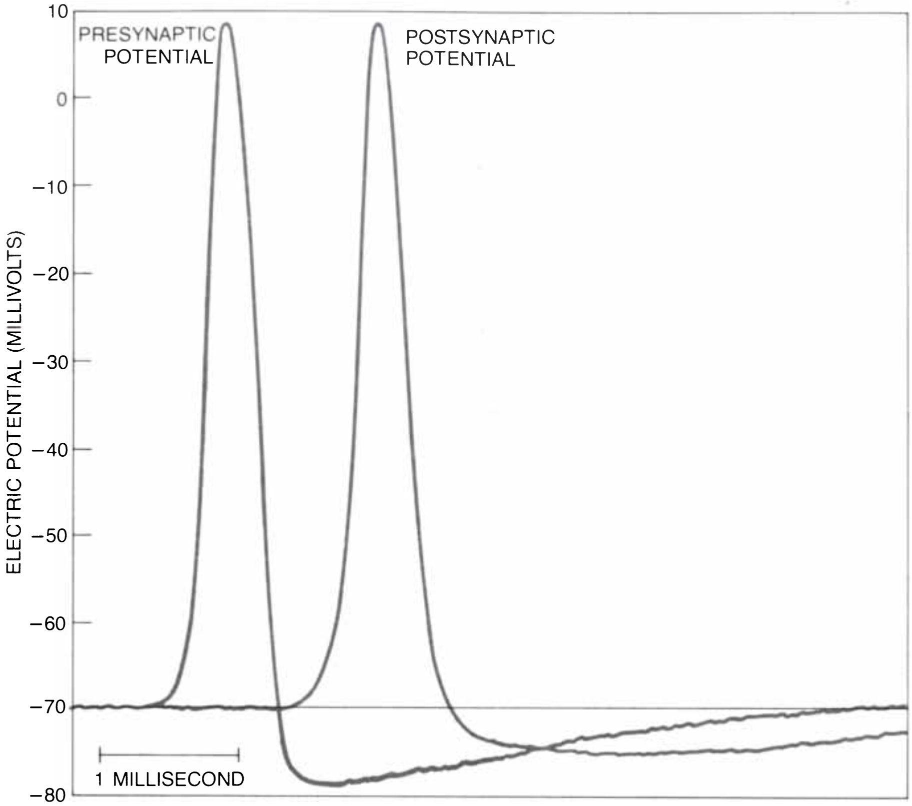
In 1966 Bernard Katz and Miledi, working at the Zoological Station in Naples, and two groups working at the Marine Biological Laboratory in Woods Hole, Mass., one group consisting of Kiyoshi Kusano, D. R. Livengood and Robert Werman and the other of James R. Bloedel, Peter W. Gage, David M. J. Quastel and me, made a further discovery. Signal transmission across the giant synapse does not require the arrival of a presynaptic action potential. It requires only that the presynaptic terminal be depolarized, which can be accomplished not only by an action potential but also by passing an electric current into the terminal through a microelectrode that impales it. The discovery was significant because it separated the mechanism of electrical excitability responsible for the action potential from the mechanism responsible for the release of transmitter from the presynaptic terminal.

The amount of transmitter released by the presynaptic terminal under artificial stimulation turned out to depend in a curious way on the degree of presynaptic depolarization. A slight depolarization of the presynaptic membrane causes a small postsynaptic depolarization; hence one infers that a relatively small amount of transmitter is released from the presynaptic terminal. A larger presynaptic depolarization leads to a larger response, but only up to a point: the response decreases if the presynaptic depolarization is greater than 60 millivolts with respect to its resting value of −70 millivolts, or in other words if the presynaptic potential attains a value more positive than −10 millivolts. If the presynaptic depolarization is great enough, so that the presynaptic potential becomes greater than +100 millivolts, the postsynaptic response can actually disappear. Even then, however, brief synaptic transmission is detected at the end of the experimental pulse as the presynaptic potential returns to its resting value.

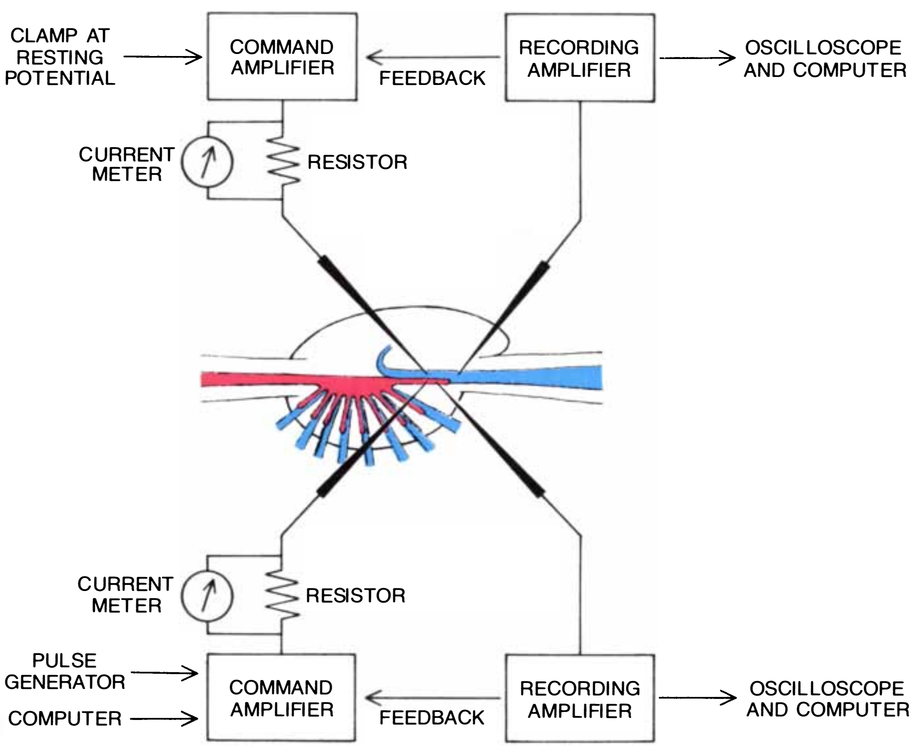
These observations were interpreted in terms of the effect of depolarization on the flow of calcium ions. When the neuron is in its resting state, the membrane is impermeable to calcium ions, which have a much higher concentration outside the cell than inside. Depolarization evidently opens channels in the membrane that allow calcium ions to pass; as a result the ions flow inward, propelled by the electromotive force arising from what remains of the electric potential across the membrane as well as the lesser concentration of calcium inside the terminal. If the depolarization is sufficiently strong, however, the influx of ions is opposed by electrical forces. The calcium ion carries two positive electric charges, and when the interior of the cell becomes electrically positive, the ions are repelled. Thus even though the membrane channels are opened by the depolarization, the calcium ions cannot enter the cell and neurotransmitter is not released. The degree of presynaptic depolarization at which synaptic transmission is abolished was named the suppression potential, and the postsynaptic response at the end of a pulse was named the off postsynaptic potential.

The early results of the study of the giant synapse lent support, then, to the hypothesis, first proposed by Katz, that calcium triggers the release of neurotransmitter. Still, the calcium hypothesis came of age only when the results of three further lines of investigation became known in the late 1960's and early 1970's.

First, Katz and Miledi demonstrated that calcium ions are capable of generating action potentials, but only at the presynaptic terminal, and that these action potentials are accompanied by the release of transmitter. In particular they demonstrated that if the ability of the presynaptic membrane to pass sodium and potassium ions is blocked by drugs, action potentials can nonetheless be detected along with a subsequent postsynaptic depolarization. What modulates the voltage across the membrane of a neuron as an action potential propagates is the redistribution of ions and hence of electric charge between the in-



NORMAL ELECTRICAL ACTIVITY at the giant synapse resembles the activity at synapses in all vertebrate and invertebrate animals. When the electric potential across the membrane of the presynaptic terminal is altered from its resting value of −70 millivolts by a spike of voltage called an action potential, which is transmitted along the length of the second-order axon, the presynaptic terminal releases neurotransmitter. The arrival of the transmitter at the postsynaptic terminal alters the permeability of the postsynaptic membrane to ions. The resulting ionic current through the membrane stimulates the postsynaptic terminal to develop a voltage spike.



VOLTAGE-CLAMP CIRCUITRY imposes a predetermined pattern of voltage on the presynaptic terminal or the postsynaptic terminal or both of them. One microelectrode monitors the voltage across the membrane of the presynaptic terminal. Its measurements serve as feedback for a second microelectrode, which injects current into the terminal. Instruments monitor the amount of current injected to give the terminal a "clamped" voltage, set by a pulse generator or a computer. A similar method clamps the postsynaptic terminal at its resting potential.

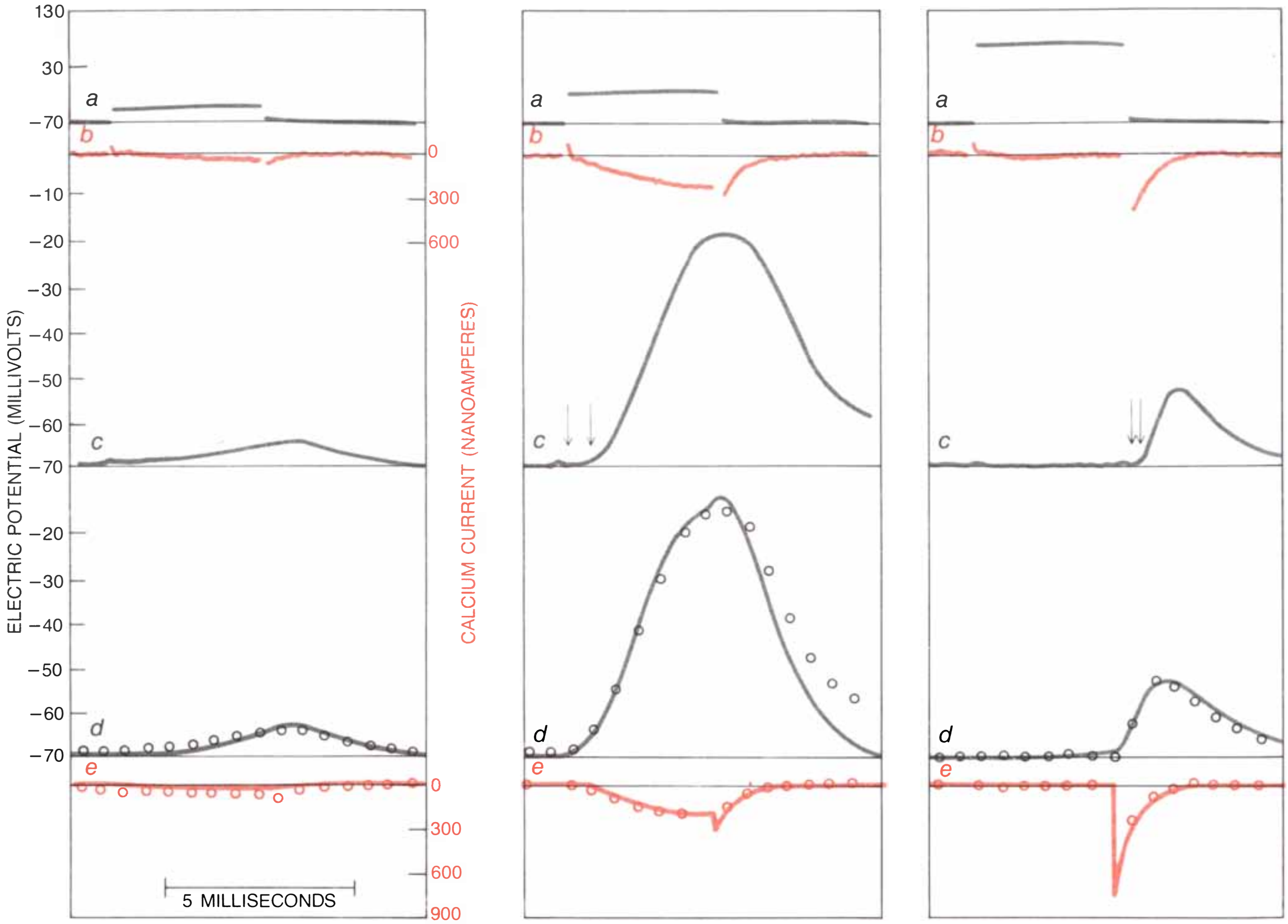
terior of the cell and the extracellular environment. Sodium ions flow in; then potassium ions flow out. The action potentials Katz and Miledi detected must have depended entirely on the passage of calcium ions. From their results they inferred that the presynaptic membrane has calcium channels. The action potentials were abolished when calcium was removed from the bathing solution.

Second, Miledi did a much simpler experiment. He showed that in the absence of extracellular calcium the mere injection of calcium ions into the presynaptic terminal causes a postsynaptic depolarization. Therefore the injected calcium must have caused transmitter to be released.

The third line of work was done by John R. Blinks, Charles Nicholson and me. We showed that if the presynaptic terminal is filled with aequorin, a protein that emits light when it is exposed to calcium, light can be detected in the presynaptic terminal during normal synaptic transmission. It follows that the concentration of calcium in the terminal increases at such times. A further experiment by Nicholson and me showed that the light signals synaptic events quite faithfully. For example, when we applied a suppression potential to the presynaptic terminal, we failed to detect light during the depolarizing pulse. We did detect light when we stopped the pulse and the postsynaptic terminal responded with an off potential. This demonstrated that the suppression of transmission is indeed due specifically to a suppression of the entry of calcium ions.

Beyond all doubt, then, calcium is the agent that triggers synaptic transmission. Two points remained to be clarified: the relation between depolarization and the entry of calcium and the subsequent relation between the entry of calcium and the release of transmitter. With this in mind my colleagues and I decided to make a change in our experimental technique. In essence we had been using a current clamp: we had injected into the presynaptic terminal a constant electric current and measured the resulting presynaptic depolarization (the changing voltage across the membrane of the terminal). The membrane has the electrical properties of resistance and capacitance, that is, it resists the flow of electric current and also stores electric charge, causing a delay in its response to a current pulse. In general, therefore, the procedure of current clamping resembles the charging of a capacitor in parallel with a resistor and the measurement of the voltage across the pair.

The difficulty with the current-clamping method is that the electrical properties of the membrane of a neuron are more complex than those of a capacitor in parallel with a resistor. Specifically,

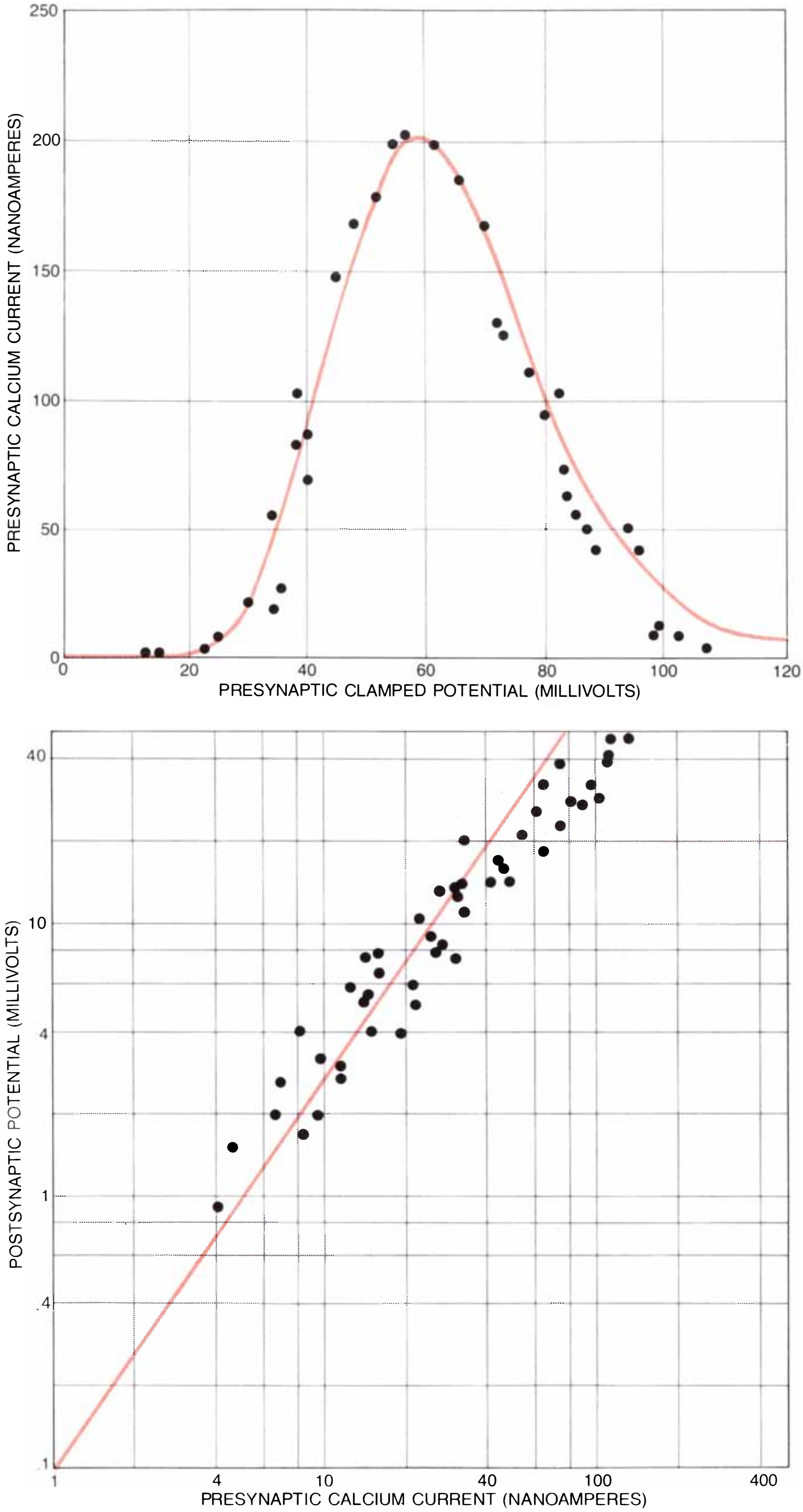


VOLTAGE-CLAMP EXPERIMENTS done by the author and his colleagues subjected the presynaptic terminal of the giant synapse to an artificial "step" in voltage (*a*). By means of drugs the terminal's membrane was made impermeable to all ions except calcium; thus the current across the membrane (*b*) represented the flow of calcium ions only. The current was measured by monitoring the current the clamp circuitry injected to counteract the change in voltage that the calcium current causes. In the graph at the left the voltage step is 30 millivolts. A small calcium current flows slowly into the presynaptic terminal. The calcium influx causes neurotransmitter to be released, a fact that can be inferred from the change in the potential of the postsynaptic terminal (*c*). In the middle graph the voltage step is 52 millivolts. The calcium current increases faster and builds to a higher level; the postsynaptic response is also more pronounced, although it does not begin until nearly a millisecond (*arrows*) after the onset of the voltage step. In the graph at the right the voltage step is 130 millivolts. The calcium current is suppressed and so is the postsynaptic response. At the end of the voltage step, however, a "tail current" of calcium flows. The postsynaptic terminal responds in only .2 millisecond (*arrows*). The data collected in voltage-clamp experiments yield a model of synaptic transmission: a mathematical description of how the synapse responds to voltage. At the bottom of each panel the postsynaptic potential (*d*) and the presynaptic calcium current (*e*) resulting from the model are compared with the experimental data (*dots*).

the changing voltage across the membrane can open channels that allow various kinds of ions (including calcium ions) to enter and leave the neuron. In short, the membrane's conductivity to ions is voltage-dependent. The movements of ions (which constitute an electric current) cause further changes in the voltage, which cause still further changes in the membrane, and so on. The most extreme result of this membrane self-modulation is an action potential. Learning the properties of the membrane when it is actively changing its voltage and no longer responding passively to experimental manipulation is problematic. Yet the active properties of the membrane are central to a nerve cell's functioning.

We determined, therefore, to end our work with current-clamping and take the opposite tack: we would employ the voltage-clamp technique that Kenneth S. Cole had devised at the Marine Biological Laboratory. A. L. Hodgkin and A. F. Huxley of the University of Cambridge had employed the technique to elucidate (in the third-order giant axon of the squid) the electrical events underlying the action potential. In the voltage-clamp technique a burst of current drives the membrane of a neuron very rapidly to a given level of depolarization. The level is then kept constant ("clamped") in spite of the movement of ions through channels that have opened in the membrane. The clamping is accomplished by an electronic feedback circuit that varies the continuing injection of current.

The voltage-clamp technique has two advantages. First, the amount of current injected when the clamp is in effect must exactly balance the current of ions; hence the metering of the applied current constitutes a measurement of the ionic current. It is a measurement that cannot otherwise be made, since the microelectrodes record only voltages. Second, the step in voltage that marks the onset of the clamping is much faster than the changes it induces in the structure of the membrane. This means the clamp circuitry charges the membrane's capacitance well before ionic currents have started to flow through membrane channels. It follows that the ionic currents can be distinguished from the injected current that merely gets stored in the terminal; therefore we could hope



RESULTS OF VOLTAGE-CLAMP EXPERIMENTS are summarized in these graphs. The relation between the clamped value of the presynaptic voltage and the plateau level of the calcium current that flows into the presynaptic terminal (*top graph*) is nonlinear: the current has its maximum for a clamped voltage of about 60 millivolts, and it is negligible for clamping above about 110 millivolts. On the other hand, the relation between the calcium current and the postsynaptic response (*bottom graph*) is quite linear. Both axes in the bottom graph are logarithmic.

in our further work with the giant synapse to study the full time course of changes in the presynaptic membrane's conductivity (or equivalently the opening and closing of channels and the passage of ions through them) for a given steady level of depolarization. Indeed, we could hope to derive from measurements of the current of calcium ions at various levels of clamped depolarization a mathematical model of the presynaptic calcium conductance, that is, a mathematical expression that would suggest how the membrane's calcium channels change in response to a changing pattern of voltage.

During the summers of 1975 through 1978 Kerry Walton and I did a series of voltage-clamp experiments at the Marine Biological Laboratory. We blocked the conductance of the presynaptic membrane to sodium and potassium, so that our measurements would represent only the flow of calcium ions. We found to our delight that the calcium current was readily measured. It turns out that the membrane's conductivity to calcium depends in a complex way on the level of clamped depolarization.

Consider the time course of the current. In response to a small step of depolarizing voltage the calcium current increases quite slowly: it reaches a plateau only after several milliseconds. In response to a larger voltage step it attains its plateau value in a fraction of a millisecond. Consider also the amplitude of the plateau. Almost no calcium current is detected unless the voltage step is at least 15 millivolts. For a larger step the plateau amplitude is markedly greater. For a step of 60 millivolts (which gives the terminal membrane a clamped potential of −10 millivolts) the amplitude is maximal. For still larger steps the amplitude decreases, a result that Katz and Miledi's discovery of the suppression potential would lead one to expect. For steps greater than 140 millivolts the current is again quite small.

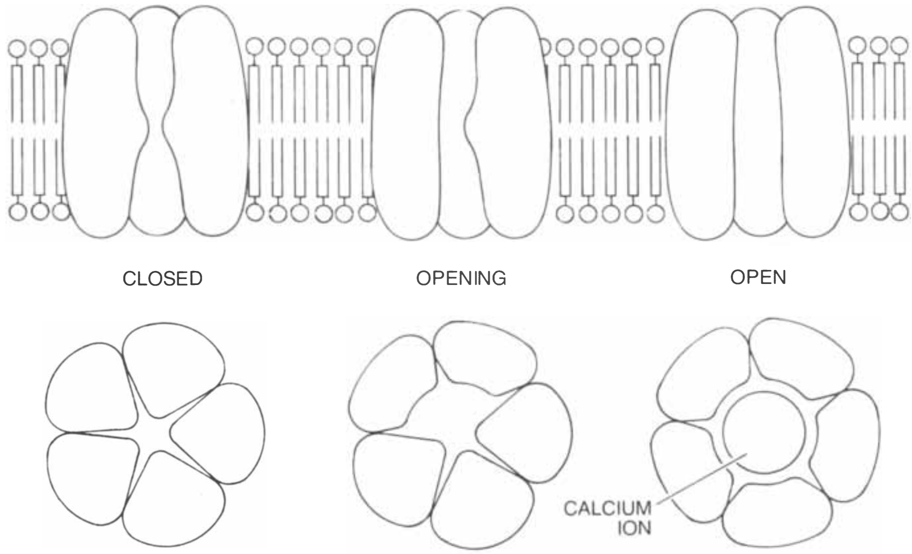
At the end of an episode of clamping the presynaptic potential falls from its clamped value back to the resting value of −70 millivolts. The return to resting potential (which can be accomplished quickly by the voltage-clamp circuitry) is accompanied by a second and distinctive flow of calcium ions into the presynaptic terminal. It is called the tail current, and it has been observed for ions other than calcium. It has no delay: it starts as soon as the voltage step ends.

The tail current can be explained as follows. Throughout the voltage step the calcium channels in the presynaptic membrane are open, and at the end of the step they close. The closing of the channels, however, is slower than the return of the potential to its resting value. The channels therefore remain open briefly after the potential has decayed, at a time when the electromotive force again is favorable for the entry of calcium ions.

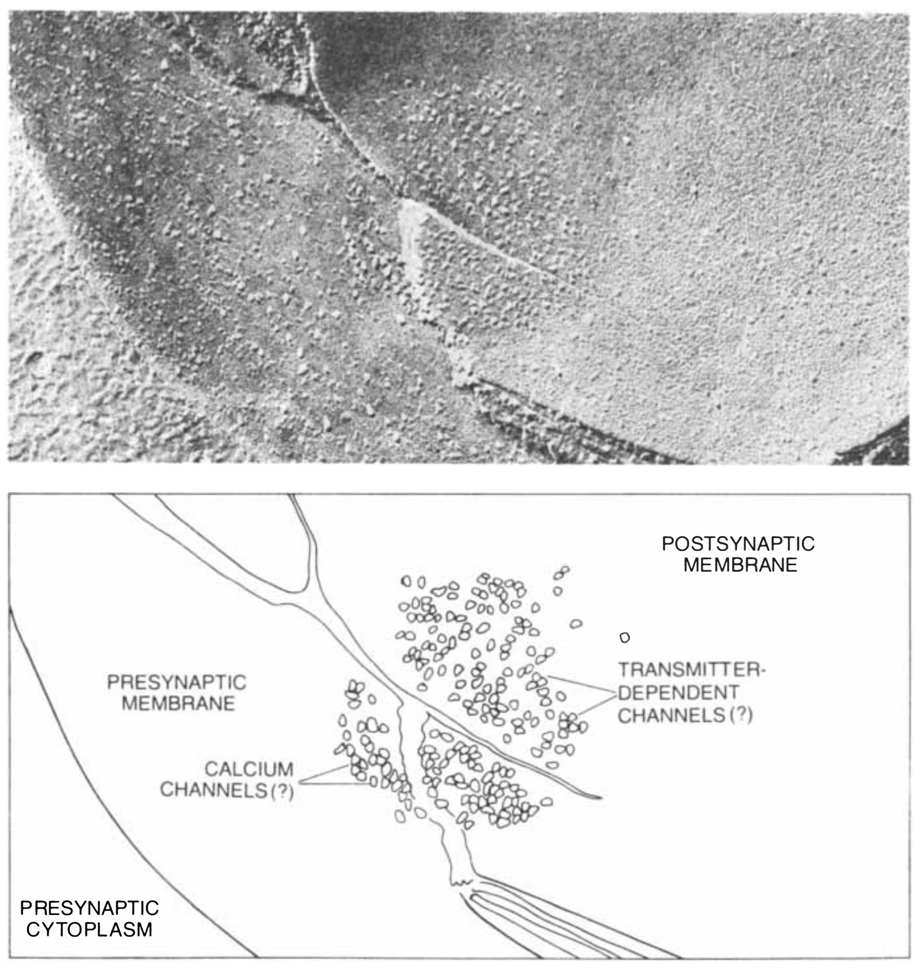
Doubtless the tail current is responsible for the release of transmitter that causes the off postsynaptic response. It is remarkable, then, that the off response begins a mere .2 millisecond after the tail current starts to flow. It is all the more remarkable because the normal postsynaptic response to the arrival of an action potential at the presynaptic terminal is delayed by as much as a millisecond. One must conclude that a large part of the synaptic delay is due to the time required for the calcium channels to open when the presynaptic membrane is first depolarized. The subsequent events in synaptic transmission must be fast indeed. In particular, once calcium ions have entered the presynaptic terminal they must act quite rapidly to get transmitter released.

The next goal of our work was to devise from our data a mathematical description relating the calcium current to any given pattern of voltage. To this end we began a collaboration with Izchak Z. Steinberg of the Weizmann Institute of Science in Israel. The heart of such a mathematical model is a description of how the calcium channels in the membrane respond to depolarization. The hypotheses with which Steinberg, Walton and I began were similar to those proposed by Hodgkin and Huxley for sodium and potassium channels. We assumed that a driving force arises from the difference between the concentration of calcium ions inside the presynaptic terminal and the concentration outside it. We further assumed that an increase in the presynaptic membrane's conductivity to calcium ions can be triggered by voltage-dependent changes in the structure of calcium channels that cause them to open. One of the simplest possibilities is that each channel consists of a certain number of subunits and that each subunit has two possible states, s and s'. Probably each state corresponds to a different shape. If all the subunits in a channel are in the state s' (the activated state), the channel is open; otherwise it is closed. The probability that a subunit will enter state s' is determined by the voltage across the membrane; the greater the extent of depolarization is, the more likely the subunit is to be activated. One further assumption is that the subunits do not interact with one another but that each subunit changes its shape independently.

From these assumptions (together with the laws of thermodynamics) it is possible to construct a family of equations expressing the time course of the calcium current in response to a clamped presynaptic voltage. The current is proportional to the rate at which



CALCIUM CHANNELS in the membrane of the presynaptic terminal must open in response to a change in voltage in order for the calcium current to flow. The model of synaptic transmission devised by the author and his colleagues suggests that five independent changes in the structure of each channel are needed to open it. Hence each channel may hypothetically be represented by a rosette consisting of five proteins each of which extends through the membrane. The voltage must change the shape of all five proteins before calcium ions can enter. The membrane of a neuron and its terminals consists itself of two layers of lipids (fatty molecules).



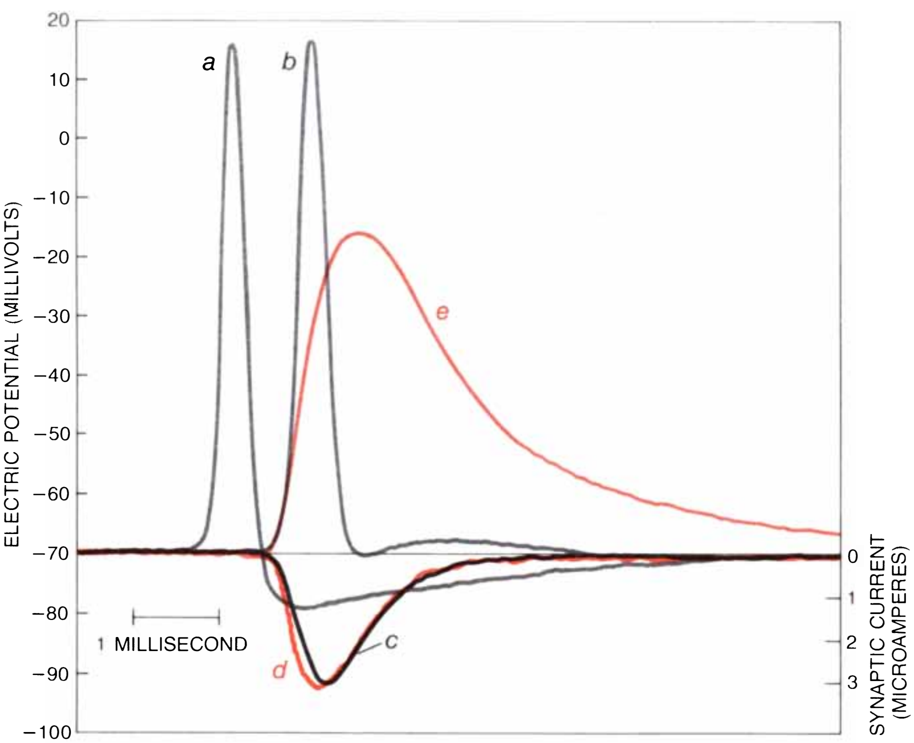
SMALL PARTICLES in the membrane of the presynaptic terminal may be the calcium channels. Particles in the postsynaptic membrane may be channels that open in response to the arrival of transmitter molecules. The tissue shown in this scanning electron micrograph was frozen, then cracked. The cracking characteristically splits a neuronal membrane down the middle (that is, between its lipid layers). Each particle is thus embedded in the membrane. The electron micrograph was made by Pumplin and Reese at an enlargement of 105,000 diameters.

ions flow through an open gate multiplied by the fraction of the gates that are open. Each equation incorporates a different assumption about the number of subunits per gate. In the equation that best matches the time course and voltage-dependence of the calcium current we had measured in our voltage-clamp experiments the number of subunits is five.

One hypothetical configuration for the channel (a configuration offered in the absence of any direct evidence) is a set of five identical proteins, each one spanning the thickness of the presynaptic membrane; the five protein molecules might be arranged in the form of a rosette. When all the proteins have the shape *s'*, they circumscribe a channel that allows the entry of calcium ions. Recently, however, the study of individual calcium channels by several investigators has suggested an alternative configuration in which each channel is circumscribed by a helix of proteins.

The voltage-clamp experiments show that the calcium current saturates, or attains a maximum value. Hence each channel must be able to pass only a certain number of calcium ions per unit of time regardless of the driving force. To explain the saturation we assume that each channel has an energy barrier. Perhaps a part of each protein molecule juts slightly into the channel and has a positive electric charge. Because the charge would tend to repel an entering calcium ion the driving force would have to "push" ions through the channel and their rate of flow would have an upper limit. The voltage-clamp experiments also show that the falloff in calcium current at the end of a voltage step is simply an exponential decrease and not the more gradual falloff that would indicate a complex sequence of events. This finding fits the hypothesis that a change in the shape of only one subunit in an open channel suffices to close it.

With a mathematical model for the response of the calcium channels to a given clamped level of voltage we could calculate the response of the channels to an arriving action potential. We approximated the action potential by a series of incremental changes in voltage and then we employed the model to predict the calcium current that would result from the succession of increments. We do not know the number of calcium channels in the presynaptic membrane, nor do we know the conductivity of an individual channel, and so the model's results could suggest not the actual magnitude of the calcium current but only its time course and its relative magnitude. Nevertheless, it appeared from the results that the calcium current starts to flow at the end of the action potential as the presynaptic voltage returns to its resting level.



ARTIFICIAL ACTION POTENTIAL produced by the voltage-clamp circuitry is shown to have effects on the giant synapse identical with those of a natural action potential. First a natural action potential (*a*) is recorded by a microelectrode positioned in the presynaptic terminal. Normally it produces a postsynaptic action potential (*b*). Now, however, the postsynaptic terminal is voltage-clamped at its resting potential. Under this circumstance the transmitter molecules arriving at the postsynaptic terminal open channels in the membrane and ions flow through them, but the clamp circuitry injects current in precisely the right amount to keep the net current zero and the voltage unaltered. The amount of current the clamp injects (*c*) is therefore equal in time course and magnitude to the "synaptic current" through the postsynaptic membrane. Next the membrane channels on both sides of the synapse that open in response to a change in voltage (as opposed to the arrival of transmitter) are incapacitated by drugs, so that they no longer pass sodium and potassium ions, the basis of a natural action potential. If the presynaptic voltage-clamp circuitry delivers the voltage pattern of the action potential recorded earlier, the resulting synaptic current (*d*) has the same latency, amplitude and time course as the one caused by natural events. If the postsynaptic terminal is then unclamped but the drugs remain, the postsynaptic terminal responds to the artificial action potential by showing a prolonged change in voltage (*e*). The change represents the passage of ions through transmitter-dependent channels. Ordinarily the postsynaptic change in voltage serves to open the voltage-dependent channels in the membrane, so that a postsynaptic action potential is generated.

Our third goal was to relate the presynaptic inflow of calcium to the release of neurotransmitter. Since the delay between the two events is only some .2 millisecond, we hypothesized that the calcium enters the presynaptic terminal quite near the site where transmitter is released. We further supposed the latter site would be marked in electron micrographs by the presence of synaptic vesicles.

Electron microscopy of the giant synapse does indeed show "active zones." They are characterized by a gathering of vesicles in the presynaptic terminal and a thickening in the membrane of the postsynaptic terminal. Scanning electron micrographs made by the freeze-fracture method reveal additional structure. In the freeze-fracture technique a section of tissue including the giant synapse is frozen and then cracked. The tissue tends to cleave down the middle of either the presynaptic membrane or the postsynaptic one. Under the electron microscope the freeze-fracture specimens show that the presynaptic membrane at an active zone has embedded in it as many as 1,500 small particles. We hypothesize (with other investigators) that they are the sites of calcium entry.

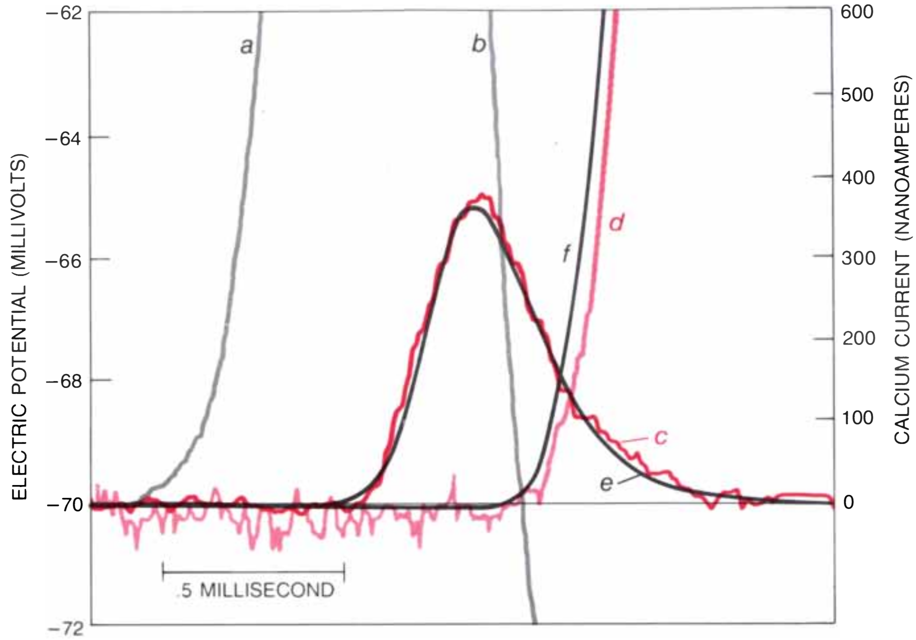
In order to incorporate into our model the relation of calcium entry to transmitter release we made some further assumptions. They are again the simplest possibilities. In the presynaptic terminal we assume that calcium binds to a molecule we call the fusion-promoting factor. In response to the binding a certain part of the factor molecule enters an active state by way of a first-order kinetic reaction (a reaction whose overall rate depends only on the concentration of a single reactant, in this case the fusion-promoting factor itself).

The activated fusion-promoting factor causes synaptic vesicles to fuse with the presynaptic membrane so that they release their content of neurotransmitter. The rate of the reaction depends only on the concentration of the activated factor. Then the activated factor returns to an inactive state. Meanwhile the neurotransmitter molecules are opening channels in the postsynaptic membrane so that ionic currents flow and the membrane becomes depolarized. The model resulting from these assumptions proved capable of reproducing the postsynaptic responses we had measured in the voltage-clamp experiments, including the postsynaptic off potential at the end of a presynaptic depolarization exceeding the suppression potential.

Thus far our voltage-clamp studies had yielded a set of experimental results amenable to a mathematical description of the presynaptic calcium current and its relation to the amplitude of the postsynaptic potential. Ultimately, however, the model must be tested against actual measurements of the calcium current during the course of an action potential. In 1979 and 1980 the test was attempted. In a series of experiments done by Mutsuyuki Sugimori, Sanford M. Simon and me at the Marine Biological Laboratory a presynaptic action potential and the resulting postsynaptic action potential were recorded simultaneously in the giant synapse. The recorded changes in membrane potential were stored in the memory of a digital computer. We then employed the recorded signals in experiments of a new kind.

As in the earlier voltage-clamp experiments the voltage-dependent conductance of the membranes for sodium and potassium was blocked chemically, leaving the presynaptic membrane with only its voltage-dependent conductance for calcium and leaving the postsynaptic membrane with only the ionic channels that are opened by the arrival of transmitter molecules. The new experiments differed in that we no longer clamped the presynaptic terminal at a constant level of depolarization. Instead the amplifier and the microelectrode that inject current into the presynaptic terminal were driven by the presynaptic voltage spike we had recorded earlier. In this way we artificially imposed on the terminal the voltage pattern of an action potential even though the normal basis of the action potential (the inward flow of sodium ions and the outward flow of potassium ions) was absent. The postsynaptic response to the artificial action potential was virtually identical with the response to the natural voltage spike; hence the artificial potential caused the release of transmitter from the presynaptic terminal with an identical amplitude and delay.

We felt justified, therefore, in thinking that further measurements made when ionic conduction was suppressed and the presynaptic terminal was stimulated by an artificial action potential would be a valid representation of natural events. Accordingly, we monitored the amount of current the voltage-clamp circuitry injected into the presynaptic terminal to make the artificial spike. Since sodium and potassium currents were being blocked, the injected current had two components. Part of it depolarized the membrane and thereby made up for the absence of the sodium and potassium currents; the rest exactly counterbalanced the calcium current that was flowing unblocked through the membrane. Next we chemically blocked the calcium current and repeated the experiment. The difference between the two results is the calcium current alone. It compares closely with the calcium current predicted by our model.

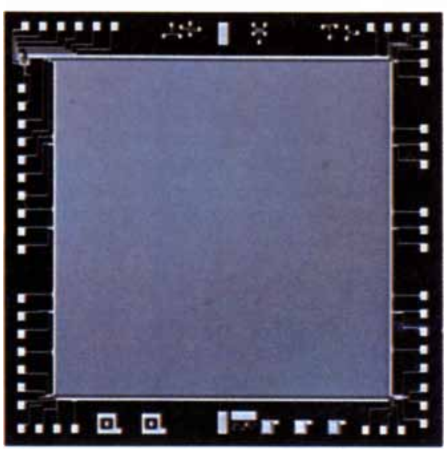


ROLE OF CALCIUM in synaptic transmission can be determined by means of the artificial action potential because the artificial voltage spike can be applied while the flow of various ions across the presynaptic membrane is being blocked pharmacologically. First the artificial spike is applied while sodium and potassium are being blocked. On the scale of the graph only the upshoot (*a*) and fall (*b*) of the spike are visible. The amount of current the clamp circuit must supply to produce the spike is monitored; it represents the sum of the amount that changes the voltage across the membrane in the absence of flows of sodium and potassium and the amount that precisely offsets the unblocked flow of calcium. The experiment is repeated with calcium also blocked. The difference between the two results corresponds to the calcium current alone (*c*). One concludes that an action potential causes the release of neurotransmitter from the presynaptic terminal by causing a current of calcium ions to enter the terminal during the falling phase of the voltage spike. The postsynaptic response to the transmitter (*d*) begins soon after. The graph also shows the calcium current (*e*) and the postsynaptic response (*f*) predicted by the model of synaptic transmission devised by the author and his colleagues.

Quite recently our results have been confirmed by the work of Steven J. Smith, Milton P. Charlton and Robert S. Zucker at the Marine Biological Laboratory and that of Miledi and Parker in Naples. The work employed a dye that changes color in the presence of calcium. The dye is injected into the presynaptic terminal of the giant synapse. Its change in color there suggests a time course for the calcium current much like the one we measured by voltage-clamping.

Many prospects are now before us. For some years I have been tantalized by the possibility that synaptic transmission is a modified form of neuronal growth. For one thing, the concentration of calcium in developing neural tissue seems to control the rate at which a growth cone, which is the growing tip of an axon, adds new cell membrane to its surface. Then too the membrane of a growth cone has a voltage-dependent conductance for calcium ions.

Perhaps the synaptic terminal can be regarded as a modified growth cone in which growth has been subdued, so that there is no longer any permanent increase in the area of the membrane. In an embryonic neuron vesicles in a growth cone might fuse with the growth-cone membrane to increase the extent of the membrane and give it newly minted proteins. Later in the life of the organism the vesicles in the presynaptic terminal that arises from the growth cone would fuse with the membrane of the terminal, evert their content of neurotransmitter and get taken back into the terminal. (Such recycling was proposed some years ago by John E. Heuser and Thomas S. Reese.) The mechanism that serves growth and plasticity in the developing nervous system would then come to serve neuronal signaling. The understanding of the molecular processes underlying this sequence of events holds the highest promise in the search for the ways of the brain.



NEW VIEW OF A SPIRAL GALAXY was obtained by means of the microelectronic light detector known as a charge-coupled device (CCD). A CCD sensor similar to the one used to record the image of the galaxy is shown enlarged about six diameters in the photograph at the left; the microcircuit, which is constructed on a square chip of semiconducting silicon 10 millimeters on a side, has 250,000 pixels, or individual picture elements, in its imaging section (*central square area*). The picture of the galaxy was made by first exposing three separate CCD frames at the telescope through blue, green and red filters and then reading them out through a set of identical filters onto color photographic film. A similar procedure can be followed with a color-television monitor by having the data from the three CCD frames drive the set's three electron guns. Color composites of this type are particularly helpful for identifying astronomical objects or regions that differ in their temperature and hence in the wavelength of the light they emit. Here, for example, the galaxy's spiral arms, which include many hot young stars, are bluer than the cooler dusty regions between the arms. One can also pick out a number of foreground stars of various colors and at least one faint background galaxy (*bluish oval at top center*). The CCD images were obtained with the aid of the 100-inch du Pont telescope at the Carnegie Institution of Washington's Las Campanas Observatory in Chile. This galaxy, designated NGC 1232, is in the constellation Eridanus at a distance of about 100 million light-years.

Charge-coupled Devices in Astronomy

Microelectronic technology has presented astronomers with a sensitive new radiation detector that is expected to improve the accuracy of many crucial observations

by Jerome Kristian and Morley Blouke

Astronomy, the oldest of sciences, concerned with nature on the largest scale, is being aided by the newest developments in microelectronics, the most advanced technology of the very small. The tiny new television sensors called charge-coupled devices (CCD's) are providing a new view of the heavens, and with unprecedented sensitivity. The history of astronomy is in large part the story of a continual search for more efficient and more accurate ways of measuring the meager light from stars. The kinds of detectors available to astronomers have always limited the data that can be gathered and the problems that can be attacked, and they have constrained, if not dictated, the direction of the advance of knowledge. Over the past few years the search for more light has combined with explosively fast developments in electronics, solid-state technology and computers to provide astronomy with the means to make measurements that were undreamed of 50 years ago and were only a vague hope just five or 10 years ago. These measurements promise dramatic new insights into the dynamic processes of the growth and decay of stars, galaxies and the universe.

Before we describe the new detectors let us briefly survey the history of astronomical light detection. Until the 17th century astronomy was done exclusively with the unaided eye. The eye is a very good light detector, perfectly tailored to its everyday uses, but it has its limitations for astronomy. Its efficiency can be as high as a few percent, which is respectable even compared with some modern light-detecting devices. The eye, however, responds only to a limited range of colors, from blue through red. Radiation in the neighboring ultraviolet and infrared regions of the spectrum is invisible. The eye can also discern subtle differences in light intensity but is a poor judge of absolute brightness. Although the eye works well over a remarkably wide range of brightnesses, it cannot store light for more than a few tenths of a second. No matter how long you look at the night sky you will not see stars fainter than some limiting magnitude.

The invention of the telescope in the early 1600's opened a new world. By collecting and concentrating the light from a larger area the telescope in effect increases the eye's sensitivity. As a result fainter objects can be seen, an improvement that was immediately evident to Galileo, who remarked that with his rudimentary instrument he could see "stars, which escape the unaided sight, so numerous as to be beyond belief." In order to observe fainter and still fainter objects astronomers have continued to build larger and more ingeniously designed telescopes.

Until late in the 19th century, however, the detector at the focus of the telescope was still the eye. Then came the new art of photography, which offered such marked advantages that it quickly became the chief detection method for astronomy. A photographic emulsion is not much more sensitive than the eye, but it has the great advantage that it can build up a picture of a faint object by accumulating light for a long time. A photograph is also a permanent record that can be taken to the laboratory, studied and kept for later reference. Modern photographic emulsions are sensitive to a wider range of wavelengths than the eye is: from the ultraviolet region to the near-infrared. Brightness can be measured photographically to an accuracy of better than 10 percent. There is still a limiting faintness, however, beyond which an object cannot be detected on a photograph; for long exposures the ever present background light from the night sky eventually saturates the entire emulsion.

After World War II photomultiplier tubes became widely available for accurate measurements of the brightness of astronomical objects. In a device of this type a photon, or quantum of radiation, strikes the surface of a photoelectric material, which responds by emitting an electron. The electron is accelerated by an electric field to a high energy, thereby constituting a small current, which is amplified within the tube to generate a measurable pulse of electricity. Photoelectric surfaces have efficiencies as high as 20 percent, and they can be made sensitive to a wide range of wavelengths: from the far-ultraviolet to the infrared.

The response of a photomultiplier is linear, that is, the output in units of electric current is directly proportional to the input in units of light intensity. Accordingly it is a straightforward matter to measure the intensity of the detected light with great accuracy. Individual photoelectrons can be distinguished, and indefinitely long exposures can be made for extremely faint objects. A major disadvantage of the photomultiplier is that it can observe only a small part of the sky; as a result stars must be measured one at a time, and extended objects such as galaxies must be sampled point by point, a task that is very costly in terms of telescope time.

More recently the technology of television and electronic image amplification has been adapted to astronomy, with the aim of combining the accuracy and unlimited exposure time of the photomultiplier with the extended field of view of the photographic plate. Such an approach should make it possible to accurately measure many stars or an entire galaxy with a single exposure. Various devices of this type have been proposed and tested for astronomy, and the range of possible schemes is still expanding. Most of the devices tried so far start with the light striking a photoelectric surface; a variety of methods are then employed for amplifying and measuring the resulting flow of electrons. Charge-coupled devices operate in a somewhat different way, which we shall now describe.

Recording a pattern of light with a CCD is rather like measuring the distribution of rainfall over a field by setting out an array of buckets before the rain and afterward moving the buckets on conveyor belts to a metering station where the amount of water in each bucket is recorded. In a CCD the "buckets" are electron-collecting zones of low electric potential created below an array of electrodes formed on the surface of a thin wafer of semiconducting silicon. The zones, called potential wells, are moved about within the device to an output amplifier by changing the voltage on the electrodes in a systematic manner.

When a photon strikes the silicon, it is very likely to give rise to a paired entity consisting of a displaced electron and the "hole" created by the temporary absence of the electron from the regular crystalline structure of the silicon. Hence a CCD resembles a photomultiplier tube in the sense that both devices convert light into electric charge. The method of capturing and measuring the charge, however, distinguishes the operation of a CCD from that of other light detectors. In the case of the CCD, when a photon creates an electron-hole pair,

OPERATING PRINCIPLE of the CCD imager is depicted in this sequence of schematic diagrams, each of which corresponds to a small segment near the top edge of the device. The strips of gray and white bars function as a system of electronic conveyor belts. The white bars represent zones of low electric potential, called potential wells, in which the photoelectrons (*colored dots*) are collected; the gray bars are zones of higher electric potential that act as barriers to keep the electrons in the potential wells. The three vertical strips in each diagram are electron-conducting channels "buried" in the body of the device's imaging section; the horizontal strip across the top is the serial output register. Three pixels are shown in each channel. Each pixel is in turn subdivided into three parts: one part low (the potential well) and two parts high (the potential barriers). The heights of the three parts can be changed by means of three sets of electrodes called gates (not shown here), which run across the surface of the chip at right angles to the channels and work in concert to move the electrons along the channels. The electrons are kept from moving sideways out of the channels by permanent barriers called channel stops (*thick black lines*). In *a* the CCD is being exposed. Photons, or light quanta, enter the chip from the rear. Each photon can liberate one electron from the regular crystalline structure of the silicon. The electrons are promptly stored in the nearest potential well. After the exposure is finished the image is read out by moving the potential wells with their trapped charge packets in a systematic fashion. First the level of the next barrier toward the output register is lowered to the same level as the well. The electrons then divide between the two wells. Finally, the level of the original well is raised so that it becomes a barrier (*b*). The effect of this operation is to move the electrons one-third of a pixel upward. After two more shifts (*c, d*) the entire pattern of charge has been moved one full pixel upward and the electrons that were in the top row of pixels have been deposited in the output register. The same technique is now applied to move this row of pixels along the output register toward the left (*e, f*). An amplifier at the end of the output register measures each charge packet in turn, thereby reading out an entire row. The process is then repeated, with each row read out until the entire chip has been emptied of information.

the electron is immediately collected in the nearest potential well, whereas the hole is forced away from the well and eventually escapes into the substrate.

A CCD imager can be thought of as an array of serial shift registers. The image-forming section is covered with closely spaced columns, called channels. The channels are separated from one another by narrow barriers called channel stops, which prevent the charge from moving sideways. Each channel is in turn subdivided along its length into pixels, or individual picture elements, by a series of parallel electrodes (also known as gates), which run across the device at right angles to the channels. Each row of pixels (that is, one pixel per column) is controlled by one set of gates. A picture is read out of the device by a succession of shifts through the imaging section, with all the rows simultaneously moving one space at a time through the body of the device [*see illustration on opposite page*].

At each shift the last row of pixels passes out of the imaging section through an isolating region called a transfer gate into an output shift register. Then, before the next row is transferred, the information is moved along the output shift register, again one pixel at a time, to an amplifier at the end, where the charge in each pixel is measured. This final step constitutes a measurement of the original light intensity registered in each pixel. The technique for moving the electric charge about in this way is called charge coupling, and that is how devices operating on this principle got their name [see "Charge-coupled Devices," by Gilbert F. Amelio; SCIENTIFIC AMERICAN, February, 1974].

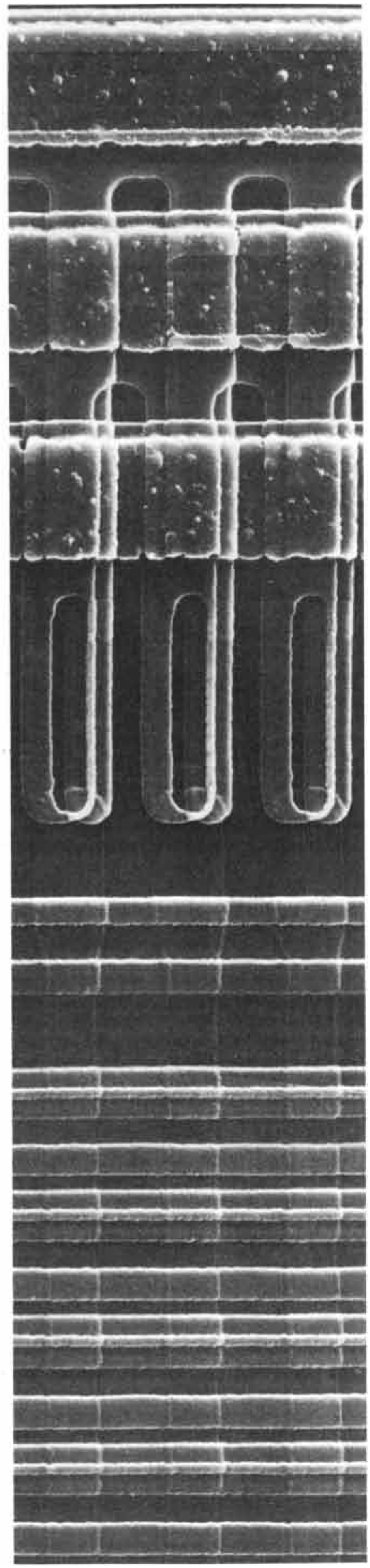
In effect the pattern of light falling on a CCD imager builds up an electron replica of itself, with more electrons created and collected where the light is brighter. The basic physics of the process is quite linear: doubling the number of photons at any pixel will double the number of electrons, until the potential well corresponding to that pixel is finally filled with electrons.

The first devices based on the charge-coupling principle were invented in 1970 by Willard S. Boyle and George E. Smith of Bell Laboratories. The first large image-forming CCD's, with more than 10,000 pixels each, were introduced in 1973. The advantages of such devices for astronomy were soon recognized by Gerald M. Smith, Frederick P. Landauer and James R. Janesick at the Jet Propulsion Laboratory of the California Institute of Technology.

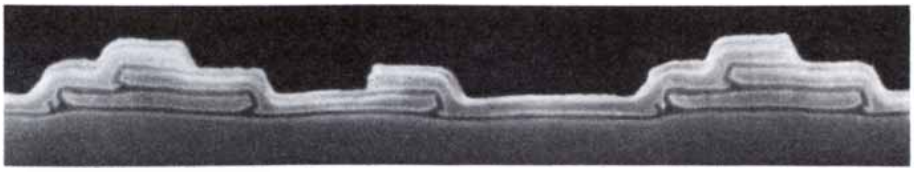
Later that year the Jet Propulsion Laboratory joined with the National Aeronautics and Space Administration and Texas Instruments to initiate a program for the development of large-area CCD imagers for space astronomy, in particular for the Galileo mission to Jupiter. The devices made by Texas Instruments for this program evolved from a thorough study of several approaches to the problem of designing and fabricating such microcircuits. CCD imagers have also been made by Fairchild Semiconductor, the RCA Corporation and the General Electric Company (GEC) of Britain, but here we shall discuss only the Texas Instruments detector.

In the Texas Instruments CCD all the circuitry for collecting, moving and counting the electrons is constructed on a single chip of silicon. The CCD is fabricated on a substrate of moderately high-resistivity *p*-type silicon: material in which the main charge carriers are positively charged electron holes. The overall fabrication process is similar to one originally developed by workers at Bell Laboratories. The first step is the creation of the five-micrometer-wide channel stops by diffusing boron ions through a mask into the exposed parts of the silicon substrate and then growing a thick (about 10,000 angstrom units) layer of silicon dioxide in those areas. Next the "buried" channels are created by implanting phosphorus ions in the areas not covered by the thick oxide. The phosphorus ions extend some 2,000 or 3,000 angstroms into the silicon. The dose of phosphorus converts the area below the surface into an *n*-type semiconductor: one in which the main charge carriers are negatively charged electrons. The *pn*-diode structure so formed localizes the potential wells at a position far from the interface between the silicon substrate and the superposed layer of insulating silicon dioxide.

The purpose of the buried channels is to enable the device to transfer charge



MICROCIRCUITRY ON SURFACE of a CCD imager made by Texas Instruments is enlarged some 1,500 diameters in this scanning electron micrograph made by John H. Tregilgas and one of the authors (Blouke). The area seen in the micrograph corresponds roughly to the one shown in the diagrams on the opposite page. A segment of the imaging section of the device occupies the bottom third of the micrograph. The raised horizontal bands are the electrodes for controlling the movement of the rows of charge packets upward through the channels. Each pixel is covered by three of these bands, which overlap to form a steplike structure. The less prominent vertical bands are the channel stops. The two raised horizontal bands just above the imaging section define the transfer gate through which the top row of pixels is transferred into the horizontal output register. The transfer gate isolates the imaging section of the chip from the output register. The output register is overlain by vertical structures that control the movement of the charge packets toward the amplifier (out of view to the left). The three broad horizontal bands near the top are aluminum contact strips that supply the voltage to the electrodes over the output register.



CROSS SECTION OF A PIXEL in the imaging section of the Texas Instruments CCD is enlarged about 7,500 diameters in this scanning electron micrograph. The cross section is along a vertical cut in the chip whose surface appears in the illustration on the preceding page, and it shows 1⅓ pixels. The characteristic steplike structure is formed by the three overlapping layers of polysilicon (polycrystalline silicon) that serve as the electrodes in this section of the device. Each first-level electrode is a polysilicon ribbon five microns wide (about a twenty-fifth of the diameter of a human hair) and 7.6 millimeters long, running in and out of the plane of the page. On a single chip there are altogether about 40 feet of such ribbon in the three sets of electrodes; the total volume of the polysilicon would fit into a cube half a millimeter on a side.

more efficiently by keeping the signal electrons away from the interface between the silicon and the silicon dioxide, where they can become trapped during the charge transfer. In general whenever a transfer is made from one well to the next, a small amount of charge is left behind. One effect of this residue is a blurring of the image; electrons that were generated at neighboring pixels are mixed together. In a square array 800 pixels on a side the charge packet farthest from the amplifier is transferred 4,800 times (2,400 times in each direction); in order to lose no more than 10 percent of the charge the loss at each transfer must be held to less than two electrons per 100,000. It is a measure of the power of the present technology that such a stringent standard is now routinely met.

The next step in making a CCD imager is to build the electrodes for collecting and moving the charge. After the formation of the buried channels a layer of silicon dioxide 1,200 angstroms thick is thermally grown on the surface to provide an insulating base for the electrodes. A layer of polysilicon (polycrystalline silicon) 5,000 angstroms thick is then grown on top of the oxide layer and is heavily "doped" with phosphorus to increase its conductivity. The first set of electrodes is made from the polysilicon layer by removing unwanted material by means of a standard photolithographic technique. The unprotected gate oxide between the electrodes is etched off and a new gate oxide of the same thickness as the original one is grown over the exposed channel. Simultaneously a somewhat thicker oxide layer is grown over the polysilicon to electrically isolate this first set of electrodes from the electrodes that are later formed over them.

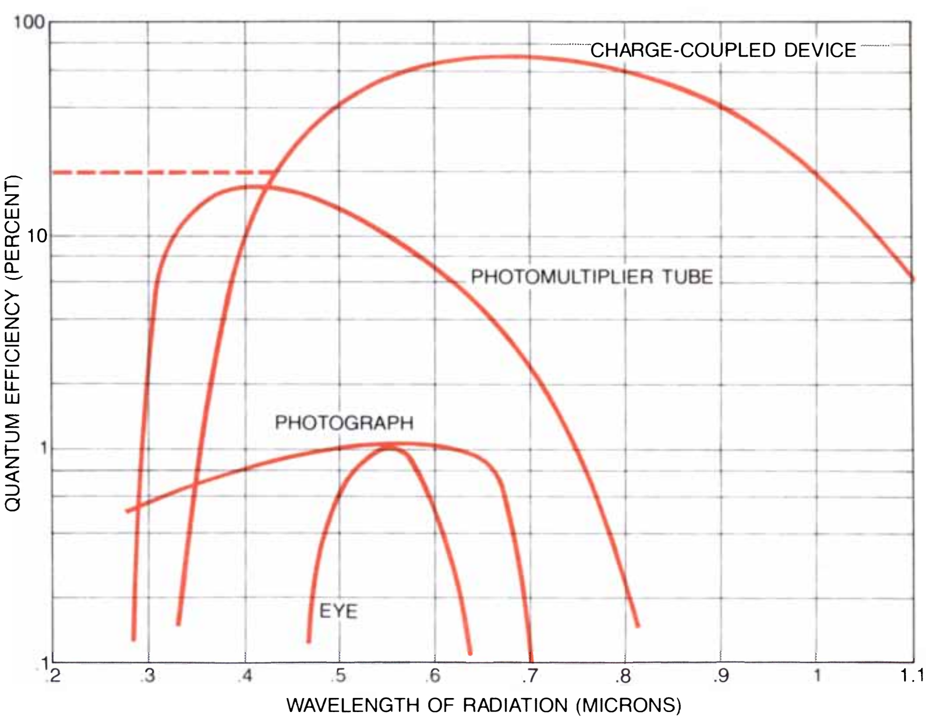
A second layer of polysilicon is then grown, doped and patterned to form the second set of electrodes. This step is followed by another etch-and-regrow cycle and by the deposition and doping of a third level of polysilicon, from which the third and final set of electrodes is made. The chip is finished by selectively etching holes in the oxide layer over the diodes and over specific contact points in the three polysilicon layers. Finally, a pattern of aluminum strips is formed by vapor-depositing aluminum over the entire surface and then defining the leads photolithographically. The aluminum makes electrical contact with the diodes and the polysilicon gates, and it leads to peripheral bonding pads where the chip can be connected with its external control circuitry.

Both silicon and polysilicon strongly absorb short-wavelength radiation, that is, radiation at the blue end of the visible spectrum. In order to improve the sensitivity of the CCD to blue light for astronomical observations, the imager is operated with the light striking the device from the rear; for this reason the imaging area is thinned to a thickness of about eight micrometers. A rim of unthinned silicon some 650 micrometers wide and 200 micrometers thick is left for support during mounting and for bonding the chip to the external leads.

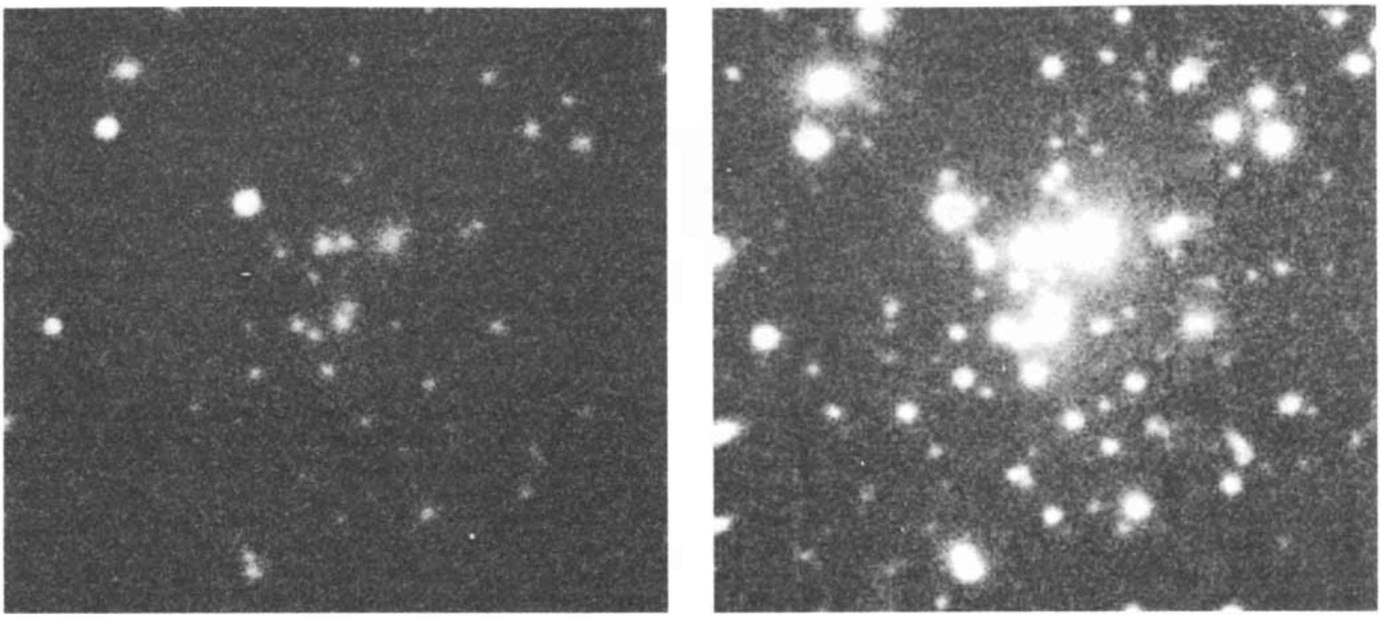
Two serial output registers are provided, one at the top of the array and another at the bottom, so that the image can be read out in either direction. Each register feeds into a simple two-transistor amplifier. One version of the Texas Instruments CCD has a square array of pixels with 800 pixels on a side, for a total of 640,000; each pixel is itself a square measuring 15 micrometers on a side. The total imaging section occupies a square area 12.2 millimeters on a side in the center of the chip, which has a total area of three square centimeters (17.8 by 17.8 millimeters).

How does the new CCD imager compare with other light detectors available for astronomy? There are six major criteria for making such a comparison: quantum efficiency, noise level, dynamic range, color response, photometric accuracy and geometric stability. We shall briefly discuss each of these characteristics in turn.

An ideal detector would have a quantum efficiency of 100 percent, that is, it would generate a measurable response for every photon that struck it, without introducing noise, or spurious signals. In addition it would be sensitive to light of all colors, and it would give an accurate value for the brightness of the light at every point in a scene. It could be used



LIGHT DETECTORS available for astronomy today are compared in this graph in terms of their quantum efficiency (a measure of sensitivity) and their spectral range. The CCD imager is superior at all wavelengths except those in the ultraviolet region of the electromagnetic spectrum. The broken line represents the effect of a technique devised by James A. Westphal of the California Institute of Technology to overcome this deficiency by coating the CCD with coronene, an organic phosphor that converts photons of ultraviolet radiation into photons of light.



PHOTOGRAPH AND CCD IMAGE of an extraordinarily rich cluster of galaxies are compared. The cluster, Abell 1689, is in the constellation Virgo at a distance of about a billion light-years from our galaxy; its velocity of recession is a sixth the speed of light. Almost every object in both of the pictures is a galaxy. The photograph (*left*) was made with the 100-inch telescope on Mount Wilson by William C. Miller; the exposure time was 90 minutes. The CCD image (*right*) was made with the 60-inch telescope on Palomar Mountain by Donald P. Schneider, John G. Hoessel and James E. Gunn; the exposure was 25 minutes. Even though the photograph was made with almost three times the light-collecting area and more than three times the exposure, the CCD image reveals objects that are much fainter.

for indefinitely long exposures, in order to allow very faint objects to be detected, and it would be able to measure accurately both faint and bright objects in the same picture. Finally, it would be able to record the positions of the incoming photons accurately, so that the exact position of a star or of the spectral lines from a star could be determined.

Quantum efficiency is a measure of the sensitivity of a detector. Photoelectric materials now in service typically emit an electron for every five to 10 incident photons; therefore they have a quantum efficiency of between 10 and 20 percent. The quantum efficiency of the eye or a photographic plate is not as easy to define, but the concept is still useful for rough comparisons. The superiority of the CCD in this respect is evident from a plot of the quantum efficiency and the color range of several detectors [*see bottom illustration on opposite page*]. In contrast to the eye and the kinds of photographic emulsions commonly used in astronomy, which have a maximum quantum efficiency of a few percent, the new CCD detectors have a quantum efficiency that goes as high as 70 percent.

Since a more efficient detector yields more data, an observation made with such a detector can be done in less time or with a smaller telescope. For many problems in astronomy the largest telescopes available are now being used to their limit; improving the detector in such a case is equivalent to building a larger telescope. Because of the gain in power the new detector can pay for itself in a few nights of observing.

The general term "noise" refers to any process that contributes to errors of measurement or distortions of information. In astronomy there is an ultimate source of noise that cannot be overcome. Light is quantized in the form of photons whose arrival at any point in time and space is represented statistically by a characteristic distribution function; hence the number of photons that strike even a uniformly illuminated detector differs from area to area in a given time interval or for different time intervals in a given area. The visible effect of this photon noise is to introduce a certain granularity into any image, an effect that can obscure faint details.

Most detectors introduce additional noise of their own. In a photographic emulsion the light-sensitive particles are not distributed uniformly; they tend to clump together, producing the effect known as graininess, which hides features that are small or have a low contrast. Electronic detectors of all kinds generate noise as a result of the constant thermal agitation of their constituent atoms or molecules. These motions can sometimes give rise to spurious events that are indistinguishable from the externally caused ones that are being measured. For example, in a photomultiplier an electron will occasionally gain enough thermal energy to escape from the photoelectric surface and mimic a photoelectron ejected by an incoming photon. This thermal noise can be reduced by cooling the detector. For CCD's in astronomy, where long exposures are required, it is necessary to cool the detectors to a temperature of −100 degrees Celsius or lower.

Noise in electronic detectors also arises from the electronic components themselves. Such sources of noise are characterized in terms of equivalent photon events. One important attribute of the Texas Instruments CCD is its low noise level: the equivalent of 10 photoelectrons or less. That means the extraneous random noise introduced by the read-out process is the same as the photon noise associated with the detection of only about 100 individual photons. Although this noise level is low, it is nonetheless significant in astronomical applications, where the interval between the arrival of individual photons from the same object may be measured in seconds or more.

The dynamic range of a light detector is the ratio of the maximum detectable light intensity to the minimum detectable light intensity. The minimum level is usually determined by noise, the maximum by the fact that most detectors "saturate" in some way at high exposure levels. An ample dynamic range is needed for measuring faint objects near bright ones. That is a common situation in astronomy: bright stars and faint stars are scattered together throughout the sky; in quasars and most galaxies the nucleus is much brighter than the outer regions; faint spectra are "contaminated" by bright emission lines caused by artificial lights and the natural airglow (the faint light emitted by certain molecules in the earth's atmosphere). By increasing the dynamic range of a detector more photons can be collected before saturation, the relative noise due to statistical photon processes can be reduced and fainter objects can be detected.

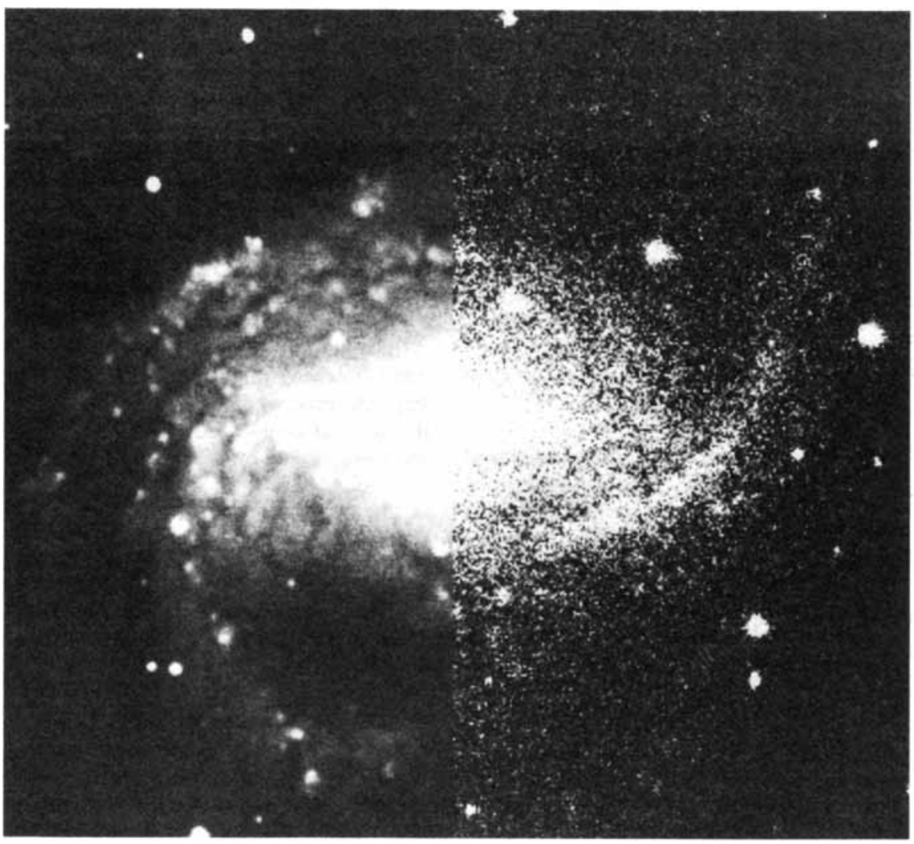
In a CCD imager the upper limit of the dynamic range is established by the filling of the potential wells with electrons. The Texas Instruments CCD has a dynamic range of about 5,000. In comparison, the dynamic range of a photographic plate or a television system is typically less than 100.

Photometric accuracy refers to the ability of a light detector to measure the exact brightness of an object. This factor becomes more important in astronomy as a particular study evolves from the stage of observing, reporting and classifying to the stage of attempting to understand the physical processes operating in the object being observed. Astrophysical theories, if they are to be most useful, must provide quantitative predictions, and the observations employed to test the theories must be accurate enough to decide between competing predictions.

An accurate measurement of brightness requires that the detector respond in a known and reproducible manner, so that a given light input always yields the same output. One can then measure the brightness of an unfamiliar object by comparing it with a source of known brightness. Furthermore, in practice it is helpful if the detector's response is linear, or directly proportional to the input, since such a relation greatly simplifies the task of analyzing the data.

The CCD imagers, like photomultipliers and many of the new television-type detectors, are linear devices. Photographs, on the other hand, are inherently nonlinear in several ways. Moreover, photographic plates can be used only once, and their characteristics are not reproducible, even for plates from the same batch of photographic emulsion, so that for the highest attainable photometric accuracy each plate must be individually calibrated, a laborious process and one that is itself not very accurate.

Geometric stability refers to the ability of a detector to record exactly where in the sky a given object is. Objects such as quasars and pulsars are often discovered solely by the radio waves they emit; on the basis of the radio emission their position in the sky can often be determined to an accuracy of a fraction of a second of arc. To identify such a radio source with a visible object in what might be a crowded field of stars calls for a light detector with a geometric accuracy at least as high. In general the geometric stability of photographic plates is quite good, whereas that of some of the new television-type detectors is rather poor. The CCD imagers have exceptionally good geometric properties, since the individual picture elements are defined by the physical structure of the chip.



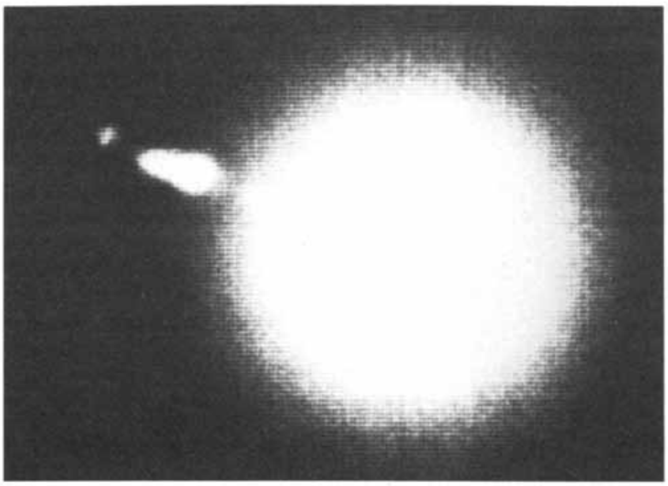
EFFECT OF NOISE on a CCD image is demonstrated in this mosaic picture of the bright spiral galaxy NGC 6951. The clarity of the original short-exposure image (*left*) is limited by "photon noise" (that is, random fluctuations in the number of photons striking the detector). In the right half of the picture the noise level of the original image has been artificially increased by computer processing but the image is otherwise unchanged. The most obvious effect of the added noise is an appearance of increased granularity. On closer inspection the loss of information in the altered image is evident in the reduced visibility of fine details and of low-contrast features such as H II (ionized hydrogen) regions and dust lanes in the galaxy's spiral arms.

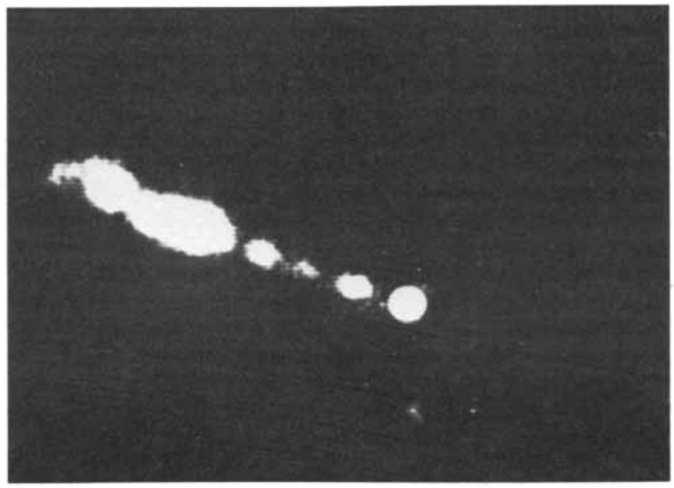
The first astronomical pictures recorded with the Texas Instruments CCD were planetary exposures made in 1976 with the 61-inch telescope on Mount Lemmon operated by the Steward Observatory of the University of Arizona. In 1977 one of the early Texas Instruments detectors was tested with the 200-inch Hale telescope on Palomar Mountain, with support from the National Science Foundation. The excellent results obtained from this early work led to the proposal to include a CCD imaging system in the Space Telescope, the 94-inch astronomical instrument scheduled to be put into orbit on a Space Shuttle flight in 1985 [see "The Space Telescope," by John N. Bahcall and Lyman Spitzer, Jr.; SCIENTIFIC AMERICAN, July].

The construction of the CCD camera for the Space Telescope is well under way at the Jet Propulsion Laboratory under the direction of a group of astronomers headed by James A. Westphal of Cal Tech. The development of this highly sophisticated detector has benefited greatly from the close interaction of detector designers, engineers and astronomers. Observations with the CCD imager have now been made at other observatories and have led to improvements in the performance of the device.

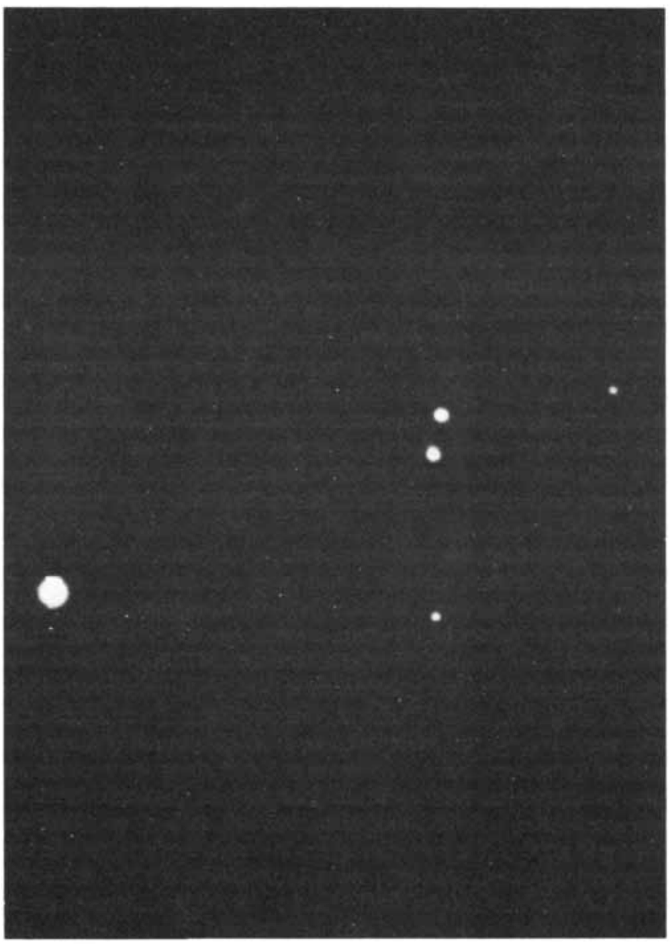
In the process of testing the new detector important data have been obtained on a wide range of astronomical topics, from planets to quasars. Some of the results of the early work accompany this article. For example, a CCD imager has been used to study the active galaxy M87 and has led to the suggestion that its nucleus harbors a massive black hole [*see top illustration on opposite page*]. Moreover, the galaxy that forms the first known gravitational lens was discovered with a CCD imager [*see bottom illustration on opposite page*].

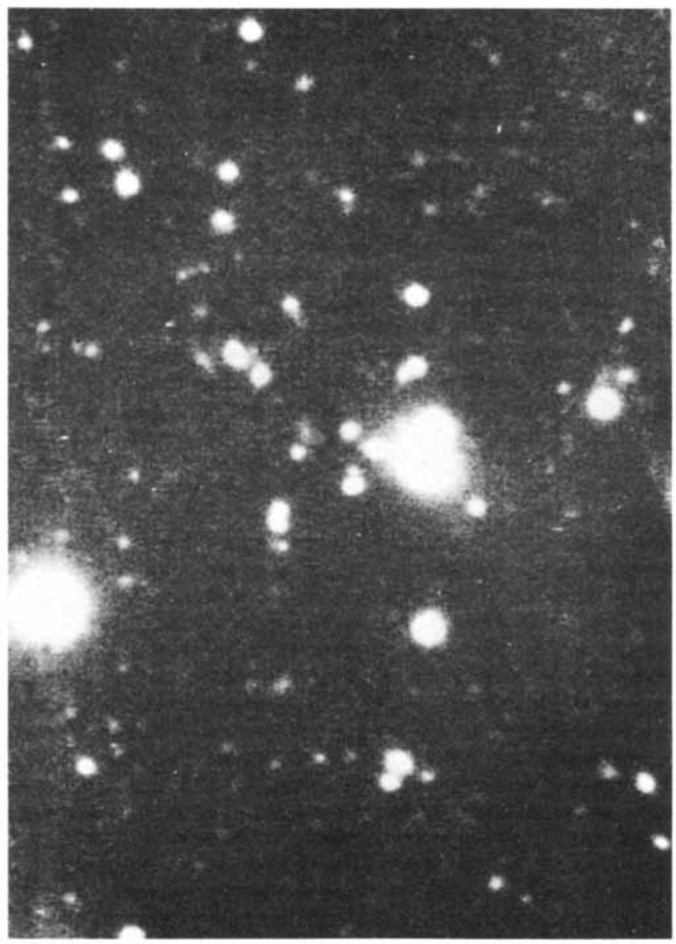
At present earth-based telescopes equipped with CCD's are being used to carry out a survey of selected small areas of the sky to detect objects fainter than the 25th magnitude: some 50 times fainter than the photographic limit of the 200-inch telescope. Red-shifted spectral lines attributable to the high recession velocities of very faint and distant galaxies and quasars are being measured. CCD imagers have been used to study the galaxies in which at least some quasars appear to be embedded, and to discover faint galaxies near some qua-



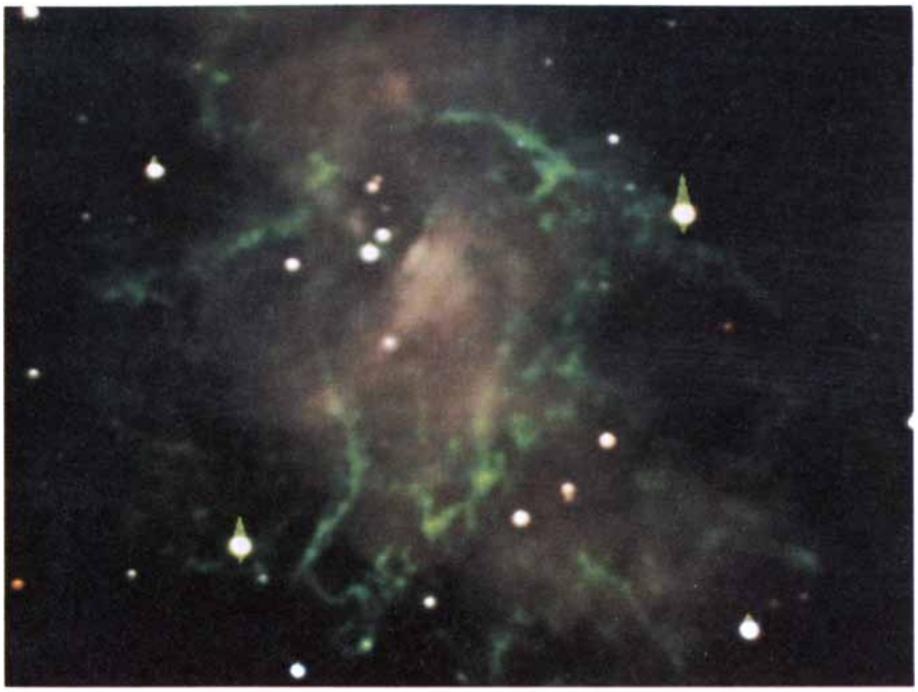


GIANT ELLIPTICAL GALAXY, M87 in the Messier catalogue of extended celestial objects, has been studied with the aid of the CCD imager in an effort to understand the nature of the violent event that is evidently taking place in its nucleus. M87 is a member of the nearest great cluster of galaxies, situated in the constellation Virgo at a distance of some 60 million light-years. In addition to being one of the brightest galaxies known M87 is a source of both radio waves and X rays. Its most peculiar feature is a jetlike series of bright spots reaching out 7,000 light-years from its center. In the CCD image at the left the outer part of the jet can be seen extending beyond the main body of the galaxy. In the CCD image at the right, made from the same data, the light from the galaxy has been removed by computer processing to show only the jet and the nucleus of the galaxy. The light from the jet, which is thought to be produced by the spiraling of energetic electrons in a magnetic field, is much bluer than the light from the normal stars that make up the rest of the galaxy. It has been suggested that the nucleus of M87 harbors a massive black hole that supplies the energy needed to account for the X-ray and radio emission from the galaxy and the radiation from the jet. CCD image was made with the 200-inch Hale telescope on Palomar Mountain.





FIRST KNOWN GRAVITATIONAL LENS was identified with the aid of this pair of photographs, both of which were reproduced from a single CCD frame recorded with the 200-inch Hale telescope on Palomar Mountain. The two photographic exposures show exactly the same region of the sky. The two bright objects at the center of the shorter exposure (*left*) originally appeared to be a pair of quasars designated Q0957 + 056; they are only six seconds of arc apart, and they have identical spectral properties. They are now known to be duplicate images of a single quasar, caused by the focusing of its light by an intervening galaxy. In the longer exposure (*right*) the lens galaxy can be seen as a fuzzy, elliptical image that completely blots out the lower image of the quasar and part of the upper image as well. This picture, which records objects fainter than 25th magnitude, is one of the deepest astronomical pictures that has ever been made.



CRAB NEBULA is seen in this CCD-based composite photograph as two components: (1) a network of gaseous filaments (which appear here as green) radiating at a set of discrete wavelengths characteristic of the emission from a hot gas and (2) a continuous emission characteristic of the radiation produced by the spiraling of energetic electrons in a magnetic field. Energy is being supplied to the nebula by a pulsar at its center. The pulsar, the first of its kind observed at wavelengths other than those in the radio region of the spectrum, is thought to be a superdense neutron star spinning at the rate of 30 revolutions per second. The pulsar is the remains of a star that exploded in A.D. 1054 as a supernova and the nebula is the debris from the explosion. The false-color composite was made at the Jet Propulsion Laboratory of Cal Tech by combining the data from three CCD frames, which were made originally at the 200-inch telescope through green, red and infrared filters. Since the eye is not sensitive to infrared radiation, the CCD frames were reproduced on the photographic film through blue, green and red filters, with the result that the picture has been "blue-shifted" by one color band. Hence the filaments, which appear green, are actually red, because the strongest of their emission lines lie in the red.

PLANETARY NEBULA, the well-known Ring Nebula in Lyra, is an expanding shell of gas ejected from the star visible at the center. The gas shines by fluorescence: it absorbs ultraviolet radiation from the hot central star and reradiates it at visible wavelengths. In this two-color, CCD-based composite photograph the nebula appears green on the inside and red on the outside because the higher-energy photons from the star are absorbed closer to the center and heat the gas there to a higher temperature, exciting shorter-wavelength (greener) emission.

sars. They have also played an important role in studies of the evolution of stars in galaxies other than our own and in analyses of the recently discovered rings of Jupiter. As these examples indicate, the new detectors are already serving in the study of a great range of objects and problems.

The new detectors generate enormous amounts of data, and their use would be quite limited if it were not for concurrent advances in computer technology. A single picture made by a CCD imager 500 pixels on a side holds the same amount of raw information as a 100,000-word book, and a single night's observing can yield as many as 100 such pictures—the equivalent of an encyclopedia! In order to be of any value to the astronomer these data must be stored, calibrated and analyzed, and the information of interest must be extracted and presented to him in a timely and comprehensible form. It would be next to impossible to do this without a modern digital computer. Computing is already a major activity at most astronomical observatories, and it will surely become much more important as the new detectors become widely available.

CCD detectors are not perfect, and they will not totally supplant other astronomical light detectors. The main disadvantage of the CCD imager at present is its small size, which is not likely soon to exceed a few million pixels in the space of a few square centimeters. In comparison, the 100-inch du Pont telescope at the Las Campanas Observatory in Chile makes photographs on plates that are 20 inches on a side: 4,500 times larger than the largest available CCD. Although the CCD chip is simple and rugged, it takes considerable technical skill and capital investment to assemble the electronic equipment needed to collect data from a telescope with a CCD, to say nothing of the computer power needed to wring scientifically interesting results from the data.

Furthermore, there are many problems in astronomy for which the attributes of a CCD imager are simply not needed, or for which its limitations make it unsuitable. For example, for surveys of large areas of the sky in which extreme accuracy and the capability of detecting very faint objects are not needed, it is doubtful that anything will soon replace photographic plates, with their unique combination of simplicity, low cost and ease of handling, storage and reproduction.

Nevertheless, the new detectors are a powerful addition to the astronomer's tools. They have already proved their value in solving difficult astronomical problems, and as they become more available they will help to solve some of the many fascinating problems with which modern astronomy is blessed.

Our high-speed video can take 2,000 separate pictures of this single second.

Photo courtesy United States Air Force.

Or, in the split-screen mode, up to 12,000 images per second can be captured. And that rate can be held for almost a full minute of continuous recording.

It may (but shouldn't) be a surprise to find Kodak at the leading edge of high-speed videography. Imaging is our business. So, as traditional motion-analysis methods neared the sheer physical limits of moving parts, film strength, and shutter function, we were already at work on non-silver alternatives and extensions. The result is the world's fastest electronic sensory ability available today.

Captured phenomena can be replayed instantly, in slow motion. Can be advanced or reversed, frame by frame. Or you can select any desired frame. And freeze it. The system's electronics allows you to analyze that frame. Instantly.

With the aid of a computer, you can input accumulated raw positional data and associated elapsed times. Compute displacement, speed, and acceleration. Generate stick figures and perform angle analysis. Fascinating and revealing impressions of rapid motion can be seen and measured. Instantly.

Thanks to the recent combination of three significant advances, the science of motion and analysis will likely never be the same.

1. A metal-oxide silica (MOS) photocapacitive, solid-state image sensor developed by our research laboratories.

2. An innovative magnetic tape, utilizing vertical recording on isotropic media, developed jointly by Kodak-Pathé (France) and Spin Physics, Inc., a Kodak company.

3. State-of-the-art microgap recording heads, developed by Spin Physics, that lay down 34 tracks on a ½-inch tape.

The sensor's array of charge-accumulation cells (192 x 240 pixels) can record a total of 46,080 bits of spatial information within the frame area. Unlike other video photosensors, the MOS device is unaffected by accidental overexposure to light. And since there is no time lag, there is none of the ghosting or blur found in conventional video systems.

That's not all. For additional information, write Spin Physics, Inc., a Kodak company, Dept. 280, 3099 Science Park Rd., San Diego, California 92121. Or call 714-481-8182, extension 2800.

Kodak

Kodak. Where technology anticipates need.

Nuclear Boom

The accident that destroyed the core of a nuclear-power reactor at Three Mile Island in 1979 seemed likely for a time to be equally damaging to the entire nuclear-power industry in the U.S. The long-term prospects are still uncertain, but for now the industry seems to be growing faster than it has at any other time in its 30-year history. Nuclear reactors accounted for 12 percent of the electricity generated in the U.S. last year; this year they should become the second-largest source of electric power, exceeded in energy output only by coal. One utility company in New England reported that in May 85.7 percent of its electric-power production came from nuclear sources.

Some 82 power reactors are now licensed for operation in the U.S.; another 68 are under construction and eight more are on order. Nuclear-power production is increasing at a rate of 7 percent per year. Because the total consumption of electric power is growing much slower the fraction of all electricity generated by nuclear reactors can be expected to increase quite sharply. In effect nuclear plants are replacing fossil-fuel plants rather than supplementing their output.

The period of rapid growth could end abruptly, however, when the reactors now being planned and built are completed. In the past year there have been no new orders for power reactors to be built in the U.S. and eight projects have been canceled. Public unease about the safety of reactor technology is only one reason for the downturn. The cost of building, operating and maintaining a nuclear reactor has been rising steeply and future demand is uncertain. As a result it has become more difficult for utility companies to persuade their customers and regulatory bodies that new plants represent a prudent investment.

Elsewhere in the world the long-term commitment to nuclear power seems to be appreciably firmer. Worldwide nuclear generating capacity increased 13 percent in 1981, and 10 new reactors were ordered. (They will be built in France, Japan, India and Romania.) The French program is particularly vigorous: France has become the first country whose primary source of electricity is nuclear power. Last year 38 percent of all electric power in France was generated by nuclear reactors.

Sequential Sequencing

Determining the exact sequence of paired nucleotide bases in a gene, and thereby reading the text of a message written in the genetic code, has lately become a matter of great importance in molecular biology. Many of the new methods of industrial microbiology require knowledge of a sequence of nucleotides so that a gene can be synthesized and inserted into a bacterium, where it is replicated and translated into a protein. Furthermore, it is generally easier to decipher the gene that codes for a protein than it is to analyze the protein itself. Even so, determining the sequence of nucleotides in a segment of DNA remains a formidable task. The existing methods are time-consuming and cumbersome and rely on random processes. Now Guo Fang Hong of the Chinese Academy of Sciences, who is currently working at the British Medical Research Council Laboratory of Molecular Biology in Cambridge, has devised a method for rapidly finding such sequences; he describes his work in *Journal of Molecular Biology.*

In the late 1970's two methods for finding the sequence of as many as 400 consecutive nucleotides were developed. One method was devised by Walter Gilbert and Allan M. Maxam of Harvard University and the other by Frederick Sanger of the Medical Research Council Laboratory of Molecular Biology. Most genes, however, are at least several thousand base pairs long. To determine the sequence of such a long segment of DNA it has been necessary to break up the strand into many fragments of 400 or fewer base pairs. One fragment after another is then sequenced. As sequences accumulate, overlapping regions are identified and the code for the entire segment is slowly filled in. As the determination proceeds, however, the probability increases that any fragment will turn out to include a large proportion of nucleotides whose sequence has already been determined. Such techniques, which are known as shotgun methods, therefore become very slow as they approach the end of the sequencing process.

The sequence of nucleotides in a long piece of DNA could be found quickly and directly with existing methods if certain conditions were met. To understand how it might be done, imagine 10 copies of a piece of DNA 2,000 base pairs long. One of the copies is left intact, but 200 base pairs are cut from one end of the next copy, 400 base pairs are cut from the same end of the next copy and so on. The result is a set of fragments that range in length from 200 base pairs to 2,000.

If the standard methods of determining nucleotide sequences are applied to the fragments, they yield information about 400 base pairs in each fragment, beginning at the cut end and proceeding toward the uncut end. Because the fragments differ in length by only 200 base pairs the sequences determined in this way overlap. If the overlapping regions of each piece are aligned, the sequence of base pairs in the original DNA can readily be determined. This simple principle is the basis of Hong's method, but several ingenious manipulations are needed to obtain DNA fragments of the right length and in the right order.

Hong's starting point was the DNA of the virus called phage lambda, which infects bacteria. The restriction endonuclease designated Hind III was employed to cut out a piece of the viral DNA 2,327 base pairs long. (A restriction endonuclease is an enzyme that cuts DNA at a site defined by a specific group of base pairs.) The same endonuclease, along with other enzymes, was then utilized to insert the segment of phage lambda DNA into the circular double-strand form of the DNA of the virus M13. The M13 DNA served as a vehicle for introducing the lambda DNA into a strain of the bacterium *Escherichia coli.* The vehicle was constructed so that near one end of the lambda DNA there was a cutting site for the restriction endonuclease Sma I. Adjacent to the cutting site another group of bases formed a priming site, which is the point on the DNA molecule where techniques of sequence determination begin to yield information about the order of the base pairs.

The loop was put into the bacterium, where it was replicated many times. The replicas were removed and subjected to the action of the enzyme DNAase I. Unlike the endonucleases, DNAase I cuts the DNA molecule randomly. The procedure therefore yielded many linear molecules formed by breaking the loop in different places.

The linear molecules of interest were those that had been cut within the inserted stretch of lambda DNA. Hong found a way to separate these from the loops that had been cut within the M13 DNA. Each of the selected molecules thus consisted of the intact M13 DNA in linear form with a short segment of lambda DNA at each end. The length of the lambda fragments varied from one molecule to another.

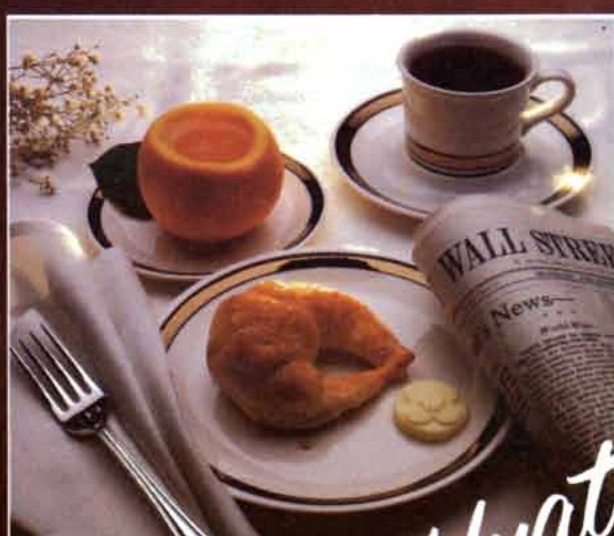
On each loop at the boundary between the M13 DNA and one of the two short pieces of lambda DNA was the Sma I site, with the priming site nearby. Sma I was employed to cut off the segment of lambda DNA extending beyond the cutting site. The remaining DNA was then rejoined into a circle. As a consequence of the position of the Sma I site and the priming site, the priming site on each linear molecule was joined to the

All things considered, Hyatt Hotels continue to be a touch above the rest.

Hyatt: It's what other hotels try to be. Over the years, and especially recently, other hotels have raised their rates within dollars of Hyatt without justifying that increase with the service and amenities that are, in total, uniquely Hyatt.

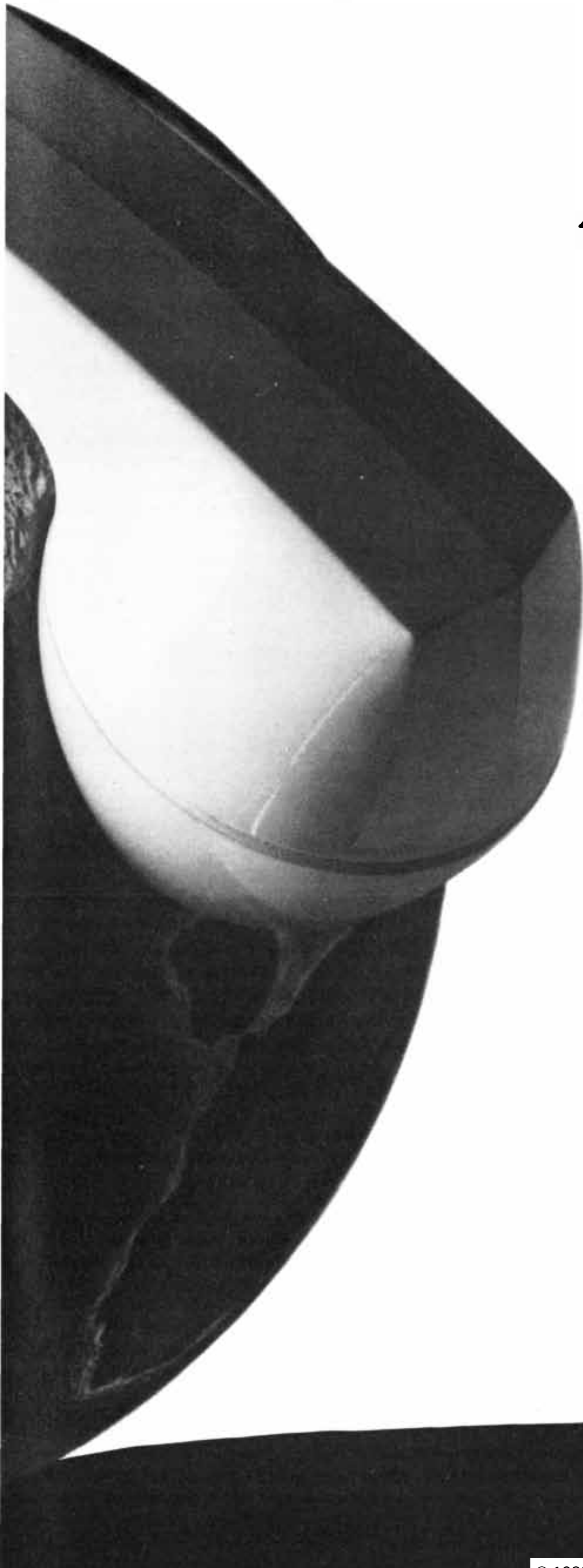
The world's most innovative hotel continues to lead the way. Hyatt is always pioneering new standards of contemporary hotel service at 109 prime locations across the U.S. and around the world. Personal attention and special touches are offered as a matter of course. Fresh flowers. Natural greenery. Just-squeezed orange juice. A complimentary newspaper in the morning. Express check-in and check-out.

Capture the Hyatt spirit at 109 hotels worldwide. Weekend packages are another special touch of Hyatt. For reservations, call your travel planner or 800 228 9000.



a touch of Hyatt®

HYATT HOTELS



"Improving telecommunications all over the world? GEE!"

(No, GTE!)

At an oil camp on a wind-whipped island off the tip of South America, a GTE digital switching system links the camp with the outside world.

Along a 6½-mile route between Brussels and the nearby city of Vilvoorde, a GTE optical communications system carries hundreds of telephone calls at once on high-speed pulses of light.

We brought solar telecommunications to the jungles of Colombia. Microwave telecommunications to Mexico and Spain. And satellite telecommunications to such countries as China and Italy.

We provide telecommunications for countries—like Algeria and Venezuela. Or just cities, if need be—like Vancouver, B.C., and Honolulu, Hawaii.

There's hardly a part of the globe that isn't affected, in one way or another, by GTE. Of course, the degree of involvement may vary. But, as far as we're concerned, the degree of importance is always the same.

GTE

cut end of the fragment of lambda DNA at the other end of the M13 DNA.

Each rejoined loop therefore included a piece of lambda DNA whose length depended on where the original loop had been cut open by DNAase I; since this enzyme cuts at random sites, there were a great many different lengths. The entire array was reinserted into *E. coli* and replicated. After replication the DNA loops were separated according to the length of the inserted lambda genetic material. The separation yielded some 160 samples. To determine the nucleotide sequence of the complete original lambda DNA it was necessary to have approximately 10 samples differing in length by about 200 nucleotides and having overlapping ends. Twenty-four samples that appeared to be of about the right length were selected.

To refine the selection the 24 samples were tested to see whether their ends overlapped. The method used to test for overlapping is part of the general method developed by Sanger for determining base-pair sequences. In the Sanger method four different reactions yield information about the position along the DNA molecule of the four nucleotide bases. Hong employed Sanger's method to find each unit of the base thymine. The sequence of thymine units was noted at the cut end of the piece of lambda DNA and about 200 bases away from the cut end. Because of the priming site the sequencing began at the cut end and proceeded toward the uncut end.

By comparing the thymine sequences it was possible to tell whether any two of the 24 fragments had overlapping ends; by this means Hong selected 11 overlapping fragments that covered the entire length of the intact lambda DNA. It was then a relatively straightforward task to apply the full Sanger method to the stretch of about 400 nucleotides beginning at the cut end of each segment and thereby determine the entire nucleotide sequence of the insert. Hong's method seems to hold considerable promise. He notes that "the length of the DNA along which the sequence can be read progressively appears to be limited only by the insertion capacity of the vector." The longest piece of DNA sequenced so far is about 6,000 nucleotides.

Wrinkles in Four Dimensions

Few branches of mathematics have seemed less likely to describe the physical world than topology. Physicists have occasionally described some features of a world that is topologically equivalent to a torus, but most topologists are more at home working in many-dimensional spaces than they are in the physically relevant spaces of three or four dimensions. Recently, however, two topologists have found that at least one four-dimensional space can be constructed whose overall structure is wildly different from the smoothly curved manifold of space-time to which physicists have become accustomed.

The possibility of a new structure for four-dimensional space-time is raised by the work of Michael Freedman of the University of California at San Diego, who has been seeking ways to classify all the topological manifolds in four dimensions. A topological manifold in ordinary three-dimensional space can be illustrated by a blob of clay with a fixed number of holes. The manifold is the two-dimensional surface of the blob, and it is said to be embedded in three-dimensional space. A manifold is classified topologically according to the number of holes that appear when it is embedded in a higher-dimensional space. For example, a doughnut and a coffee cup with one handle are topologically equivalent because each has one hole; indeed, if the coffee cup were pliable enough, it could be stretched and pulled into the shape of the doughnut.

Freedman's work on the classification of four-dimensional manifolds draws on the notion of a more abstract kind of hole: one that is literally a gap in space or a lacuna in the set of points included in a manifold. Such topological holes are the basis of a celebrated conjecture made in 1904 by Henri Poincaré. To understand the conjecture, consider a loop formed in a piece of string that is somehow embedded in the surface of a sphere but is still free to move on the surface. If the ends of the loop are pulled, the loop will collapse into a single point. If the same experiment is done on a torus, however, the loop may not be able to collapse. Indeed, there are two independent ways of embedding a loop in the surface of a torus so that the loop cannot be shrunk to a point. One way is to make the string pass around the hole of the torus; the other way is to make the string pass through the hole and around the solid "arm" of the torus. Paradoxically, therefore, a one-hole doughnut has two topological holes. Poincaré conjectured that if every loop of string embedded in a three-dimensional manifold can be shrunk to a point, the manifold is topologically equivalent to the three-dimensional "surface" of a four-dimensional sphere.

Poincaré's conjecture has still not been proved. In 1959, however, Stephen Smale of the University of California at Berkeley proved an analogue to the conjecture for manifolds of five dimensions or more. Freedman, in pursuing his classification, proved the conjecture for four-dimensional manifolds. In four dimensions the proof of the conjecture has two parts: it must be possible to shrink to a point not only any loop of string embedded in the manifold but also any two-dimensional sphere that can be embedded in it. For example, in ordinary Euclidean space a balloon can be shrunk indefinitely without catching on anything. If a marble inside the balloon represents a hole in space, however, the balloon cannot be shrunk without eventually catching on the hole.

In the course of his work Freedman was able to construct a remarkable four-dimensional manifold that has eight topological holes. In other words, there are eight independent ways a sphere embedded in the manifold can fail to collapse to a point. The possibility of such a manifold had been suspected for many years, but until Freedman's work no one had been able to show that it could actually exist. The manifold is called E8.

An important characteristic of a topological manifold is whether or not it can be smoothed. Smoothing is analogous to flattening a decal on a windowpane. If the decal is carefully applied, it will stick to the glass without any creases; if it is stuck first at its perimeter, however, creases can develop when the interior of the decal is flattened against the pane. Although the creases can be moved around and confined to a small area in the interior of the decal, they cannot be eliminated without unsticking some of the points at the perimeter. E8 is so convoluted that it always includes creases.

Although E8 cannot be smoothed, it remained an open question after Freedman's work whether it is possible to smooth the manifold that results from sticking two copies of E8 together, a structure topologists call E8 + E8. Recently Simon Donaldson, a graduate student at the University of Oxford, demonstrated that E8 + E8 cannot be smoothed either. Donaldson drew on the work of Clifford Taubes of Harvard University and of Karen Uhlenbeck of the University of Illinois at Chicago Circle, but his argument is a highly original one in its own right. When Freedman received word of it, he was able to perceive its consequences from his own previous work: Donaldson's result implies that there is more than one way to construct a topologically smooth four-dimensional space-time. Ordinary space-time, whether it is curved or not, is called R4; the new space-time has been referred to as fake R4.

What does fake R4 look like? No one can say for sure. According to Robion Kirby of Berkeley, it is "crinkled to a degree we've never seen before." Freedman likens it to the graph of the mathematical function sine $(1/x)$, which oscillates about the x-axis an infinite number of times for values of x between zero and 1. A physical space-time with the topological structure of fake R4 would look like ordinary space-time near an observer, but the surface of a distant four-dimensional sphere (defined as the set of points equidistant from the observer in space-time) would look almost unimaginably complex. Moreover, it is

not yet clear that there is only one candidate for fake R4; according to Freedman, it is just as likely that there is an infinite family of fake R4's. If there is, the widened perspective raises new questions, similar to those raised during the 19th century by the development of non-Euclidean geometries, about the privileged status of ordinary R4 as a description of the universe we inhabit.

Proprietary Plants

The Plant Variety Protection Act of 1970 is designed to encourage research and development in plant breeding by giving the creator of a new plant variety patentlike rights to its commercial exploitation. Minor amendments to the act were proposed in 1980, provoking congressional debate on the value and implications of the act itself. The debate led to a study of the impact of the act by the Department of Agriculture and the Food System Research Group of the University of Wisconsin's Department of Agricultural Economics.

Proponents of the act cite the economic incentives it creates for industrial activity in plant breeding. This activity, they argue, will increase the genetic diversity of plants as more varieties are developed, improve the yield of crops, release public plant-research funds for other purposes and make the U.S. more competitive in international agricultural markets. Opponents of plant patenting argue that on the contrary the act will decrease the genetic diversity of plants and slow the creation of new varieties by limiting the exchange of information and plant germplasm. It has also been suggested that the act has harmed small seed businesses by attracting larger multinational companies to the industry.

The act's effect on the genetic diversity of plants is one of the most crucial and confused issues in the debate. L. J. Butler of the University of Wisconsin at Madison, director of the recent joint study, agrees that the protection afforded by the act encourages the private development of new plant varieties. He adds, however, that industry may concentrate on developing varieties with the specific qualities considered desirable by farmers, particularly high yield and resistance to known pathogens. As a result agriculture could come to rely increasingly on genetically uniform populations of plants. If a new strain of pathogen were then to arise, an entire crop could be destroyed. Genetic diversity may offer better prospects for long-term disease resistance.

With respect to the potential for monopoly created by the interest of multinational companies in the seed industry as a result of the Plant Variety Protection Act, Butler thinks the takeovers of small companies cited by opponents of the act do not constitute market concentration. Although many small concerns have been sold or have closed, new ones have been formed, and the number of seed companies remains relatively stable. Thus there is considerable diversity among the producers of seed.

Fatal Encounter

On August 30, 1979, a comet collided with the sun. The collision apparently went unseen in the various programs that monitor the sun; Zdenek Sekanina of the Jet Propulsion Laboratory of the California Institute of Technology notes, for example, that solar photographs made every 20 seconds at the Big Bear Observatory "show no traces of any unusual phenomena." The collision was discovered only some time later, when images made by a solar coronagraph on board a U.S. Air Force satellite became available. The coronagraph includes a disk that blocks out the photosphere (the bright sphere of the sun) so that the fainter light of the surrounding solar corona can be recorded. Every 10 minutes the device records an image showing the corona at distances of between 2.5 and 10 solar radii from the center of the sun.

The coronagraph was put on board the satellite by Donald J. Michels, Neil R. Sheeley, Jr., Russell A. Howard and Martin J. Koomen of the Naval Research Laboratory, who also analyzed the images and who have discussed them in *Science*. The remarkable sequence of events, as the four investigators reconstruct it, is as follows. Late on August 30 a comet approaches the sun. The path of its head in the plane of the images is a straight line; the velocity of the head in that plane is 284 kilometers per second. The line on which the comet approaches does not intersect the center of the disk of the sun; from that fact and some others the investigators surmise that the Air Force satellite was "near but not in the [comet's] orbital plane." The head disappears, then for some two and a half hours the coronagraph makes no further images. The first image after that shows "tail material strewn along the path of approach." The tail material "persists for many hours," growing brighter than it was when the comet was approaching perihelion (the point in its orbit closest to the sun). Meanwhile the corona is becoming diffusely brighter around the half of the sun north of the tail. The comet's head is no longer seen. The brightness persists for about a day.

The investigators estimated the comet's orbit by finding the parabola whose projection onto the plane of the images would align best with the successive recorded positions of the head. It was an open question at first whether the orbit was prograde (in the same direction as that of the planets) or retrograde. As it turned out the orientation of the orbit matched well that of Ikeya-Seki, a comet considered representative of the group of comets known as the Kreutz sun-grazers, after the 19th-century astronomer H. Kreutz. The Kreutz sun-grazers, which have a retrograde orbit, approach as close to the center of the sun as one to two solar radii. If the new comet was indeed a Kreutz sun-grazer, so that it had a retrograde orbit, it would have been "very close (in angle) to the sun for many days prior to perihelion, lessening the likelihood of detection [from the earth]."

Because the tail of the comet was never visible on the side of the sun opposite the side of approach the workers place the perihelion at no more than 1.75 solar radii. Because the tail did not reappear on the same side of the sun they place it at no more than one solar radius. The question then is whether the comet collided with the sun or "simply lost all its volatile material through sublimation in a grazing encounter." The investigators note "that for particle diameters smaller than about .3 micrometer the outward radial force of the solar radiation exceeds the [sun's] gravitational attraction." It follows that small particles expelled from the comet were "quickly driven out of the comet's orbit." The particles would then have assumed "trajectories characterized by outward radial accelerations but still in the comet's orbital plane." The investigators suggest that the light scattered toward the earth from the broadened distribution of the particles accounts for the brightening of the tail after the head had disappeared.

The diffuse brightening of the corona demands a different explanation. For one thing it reached its maximum when the tail itself was fading. The investigators propose that the corona became filled with particles that "were products of the final disintegration of the cometary nucleus in the solar atmosphere." The particles were heated into rapid random motion and "blown outward by radiation pressure" with a distribution "somewhat resembling the ejection spray from a terrestrial explosion."

Sekanina, writing in *The Astronomical Journal,* offers a further analysis of how the comet was destroyed. His own study of the positional data embodied in the sequence of images from the Naval Research Laboratory leads him to calculate that if the comet was a Kreutz sun-grazer, the perihelion of its orbit was at about .35 solar radius: it is indeed "the first known case of a comet falling into the sun." (It is also the first case of a comet discovered by satellite.)

In the final image before the two-and-a-half-hour hiatus in the data the head of the comet (if it survived that long) was 2.3 solar radii from the center of the sun; it was therefore "hidden by the occulting disk of the coronagraph." The path Sekanina calculates for the com-

et would have had it approach the sun from behind (as viewed from the vantage of the coronagraph), "ready to plunge—if still surviving—into a region of the photosphere that was facing the earth." Perhaps the head disintegrated from a combination of heating and tidal stress even before it encountered the photosphere. If not, the comet would have "struck the sun" 17 minutes before it was due to reach perihelion.

Host and Guest

An animal is a self-contained ecosystem in which large populations of microorganisms occupy such external surfaces as the skin and the mucous membrane of the eye and such interior regions as the nose, throat, stomach, gut and genitourinary tract. Does the host find the symbiosis beneficial or harmful? It is certainly beneficial to ruminants: without the symbiotic rumen bacteria that ferment cellulose, the plant matter in a ruminant's diet would provide far less nourishment. What about other animals, including human beings?

The question has been a controversial one since the mid-19th century, when Louis Pasteur proposed that the "normal" microorganisms (he thus excluded pathogens) are essential to their hosts' welfare and probably to their survival. Pasteur's hypothesis was badly shaken when experiments with laboratory animals proved that successive generations could be raised quite free of the supposedly essential microorganisms. Indeed, Pasteur's colleague Ilya Ilich Mechnikov, who became director of the Pasteur Institute after its founder's death in 1895, supported exactly the opposite hypothesis. Ruminants notwithstanding, Mechnikov contended that the symbionts are competitors that seriously endanger their hosts.

Writing in *The New England Journal of Medicine,* Philip A. Mackowiak of the University of Texas Health Science Center at Dallas has summarized the evidence, mainly accumulated since the 1960's, bearing on both hypotheses. Among many findings that suggest a harmful role is the observation that certain strains of intestinal bacteria can convert nitrate, a common food preservative, into nitrite, a carcinogen. Other findings that were taken at first to indicate benefits have proved to be equivocal. For example, various gut bacteria synthesize vitamins K and B_{12} and riboflavin in quantities larger than they need. There is no evidence, however, that the surpluses contribute to the host organisms' welfare. Lactobacilli inhabiting the genitourinary tract have an inhibiting effect on the growth of the pathogenic bacterium responsible for gonorrhea. On the other hand, the majority of infections in leukemia patients can be traced to the microorganisms they play host to, and in patients with shigella the bloodstream is often invaded by the microorganisms ordinarily found only in the gut.

Much commoner illnesses may also be traced to resident bacteria. Mackowiak reports that analyses of astronauts' feces indicate a rise in the population of microorganisms that can induce diarrhea. The change seems to be attributable more to anxiety than to weightlessness, suggesting that the diarrhea associated with terrestrial travel abroad may often be caused less by infection than by the stress of foreign surroundings.

Mackowiak concludes that neither Pasteur's nor Mechnikov's hypothesis can be accepted absolutely. "Given the appropriate circumstances," he writes, "each member of the normal resident flora appears to have the capacity for helping or harming the host."

The Truly Needy

"A decent provision for the poor is the true test of civilization," said Samuel Johnson, who did his part by giving to beggars and taking in the destitute. More substantial provisions for the poor had long been embodied in British law and social institutions. Indeed, virtually all the measures for mitigating poverty that are now central to social policy were already well established early in the 17th century, several generations before Johnson's time. They included government-sponsored work programs and schools, the direct transfer of wealth by means of taxation and support payments, subsidized medical care and the distribution of food and fuel at reduced prices.

The scope and the organization of poor relief in 17th-century Britain have been reviewed over the past 10 years in a series of publications by Ronald W. Herlan of the State University of New York at Brockport. He has concentrated chiefly on London, but most recently, writing in *Proceedings of the American Philosophical Society,* he has considered the situation of the poor in Bristol from about 1600 to 1660. At the start of this period Bristol was a prosperous city and port, second only to London as a center of commerce, but the decades that followed were difficult ones. Trade declined, and then the civil war of the 1640's severely disrupted the economy and led to heavy taxation. Bristol was besieged twice, and there were periods of famine and epidemic disease. One estimate is that 20 percent of Bristol's population was living in poverty by 1660. The number of people who required assistance to meet their basic needs was occasionally large enough to strain the resources of the city.

Revenue for government programs of poor relief came primarily from the taxes called poor rates, which were "designed to purchase stocks of materials to set the able-bodied but unemployed poor to work, to teach poor children a gainful occupation by paying their apprenticeship costs and to support the incapacitated poor with weekly doles." A system of rates-in-aid was set up to divert tax funds from the wealthier parishes of the city to the poorer ones. In times of economic distress taxpayers were assessed additional amounts, and the Common Council of the city took emergency measures, such as buying quantities of grain, butter, coal and other commodities for distribution at less than market cost. In the worst years special assessments were imposed on the members of the council itself, and the city borrowed from its wealthier citizens.

Public assistance was supplemented by private charity, which most often took the form of bequests or endowments. "Through the mandatory taxes the state assured its disadvantaged members that a floor of subsistence existed to maintain them. At the same time voluntary contributions were encouraged to buttress the statutory system." It appears the amounts raised by the two methods were roughly equal.

The administration of both the public and the private funds was generally the responsibility of the parish churchwardens. Herlan examines in detail the records of Temple, a parish south of the River Avon with a sizable poor population. In addition to other sources of revenue Temple parish received a share of the profits of an annual trade fair. An accounting for the church year 1654–55 shows that "parish land rentals, bread receipts and other sources amounted to £131 9s. 8d. St. Paul's Fair produced a profit of £50 11s. 9d. while poor rates (£25 10s. 2d.) and rates-in-aid from four Bristol parishes (£32 6s.) contributed another £57 16s. 2d. Total outlays to assist the poor came to just over £113 or approximately 59 percent of the £190 10s. 5d. disbursed in that accounting year." A decade later the fraction of parish funds going to poor relief approached two-thirds.

Herlan points out that compassion was by no means the only sentiment motivating efforts to help the disadvantaged. The poor rates were instituted as part of a "national scheme adopted to prevent begging in the streets, to control social distress and to regulate the volatile labor market." The measures taken in Bristol throughout the 17th century, Herlan notes, demonstrate "the continuous commitment of that community's leaders to dampening possible social upheaval from below." Moreover, whether the assistance given was prompted by generosity or self-interest, the resources available were sufficient only to sustain the poor, not to eliminate their poverty. "Abolition of their wants was not even a remote possibility."

JAPANESE TECHNOLOGY TODAY.

The beauty and splendor of the traditional kimono adorns seven-year-old Maiko Takahashi as she presents one of the latest developments from Japan, the unique one-chip speech recognition LSI.

This special 1982 annual report from Japan explains the new technologies, the management policies and the traditions that have made Japan's product quality control a worldwide standard.

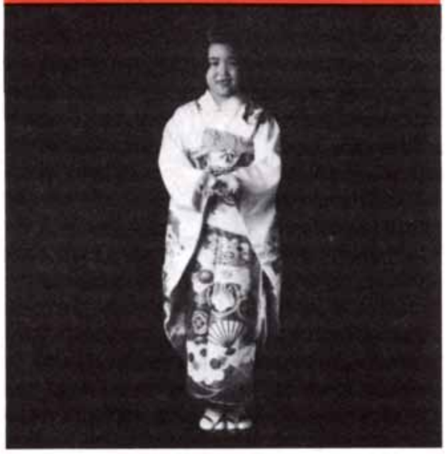
SCIENTIFIC AMERICAN

Advertisement

MATSUSHITA IN AUDIO TECHNOLOGY:


SEIKO
SEIKO QUARTZ
WED 2

JAPANESE TECHNOLOGY TODAY



THE AUTHORS

The 1982 edition of "Japanese Technology Today" is coauthored by Dr. James C. Abegglen and Mr. Akio Etori.

Dr. Abegglen is Vice-President and Director of the Boston Consulting Group, Inc., and concurrently Professor of International Management at Sophia University, Tokyo. He has authored numerous articles and books including *The Japanese Factory,* a pioneering study of Japanese corporate organization.

Mr. Akio Etori is an award-winning and respected writer on science and technology, having covered these areas for over 20 years. He is managing editor of *Saiensu,* the Japanese language edition of *Scientific American* and Japan's leading and most prestigious publication in this field.

THE COVER

Photographer Henry Wolf contrasts the tradition of kimono, which dates back thousands of years, with one of 1982's latest technological achievements from Japan, the world's first one-chip speech recognition LSI from Matsushita Electric, Japan's largest consumer electronics company. By integrating spectrum analysis and pattern-matching circuitry on one chip, the unique speech recognition LSI (MN–1263) can recognize up to 64 words of a registered voice. Think of the implications for consumer products, for the physically handicapped and for automated production lines including robotics.

Art direction and supervision of text and illustration were provided by Ted Bates Advertising/New York.

The striking increase in Japanese self-confidence regarding the nation's economic performance has come together with new technologies to cause still another of Japan's periodic national preoccupations. Japan's latest "boom" is being called the new industrial revolution. Signs of the new preoccupation are many. Alvin Toffler rode his *Third Wave* to national prominence, the best-seller list and a featured series on Japanese national television. A leading newspaper held an international symposium in May on the "Impact of the New Industrial Revolution." A world's fair of technology, Expo '85, is planned at Japan's new science university at Tsukuba to symbolize Japan's arrival at world technological leadership.

Implicit in all this is the view that Japan is leading this revolution. The economic journal *Toyo Keizai* published a special issue called "The Impact of the New Industrial Revolution." The leading article, by one of the revolution's prophets, Professor T. Ishii of Tokyo University, claims that "Now Japan's century has begun," citing especially Japan's position in microelectronics and machine tools and its potential in biotechnology.

A sharp increase in Japanese self-confidence should come as no surprise. Exceptionally vulnerable to raw material and energy supply problems and price increases, Japan has come through the difficult past decade in a better position regarding inflation, unemployment and real economic growth than any other major economy. The Japanese trade balance is again firmly positive despite import price increases and increasing trade barriers. Japan's products have won a reputation for high quality. The achievements are many, and a degree of self-satisfaction is appropriate.

This is not to say that the evidence indicates that Japan is the technological leader of a new industrial revolution, if indeed such a revolution is in fact under way. It is clear, however, that in several areas Japanese technological progress is proceeding at an exponential rate, and that in some of these areas Japan has a strong and even leading role. An assessment of Japan's overall position requires a closer look at developments and at the part Japan is playing in them.

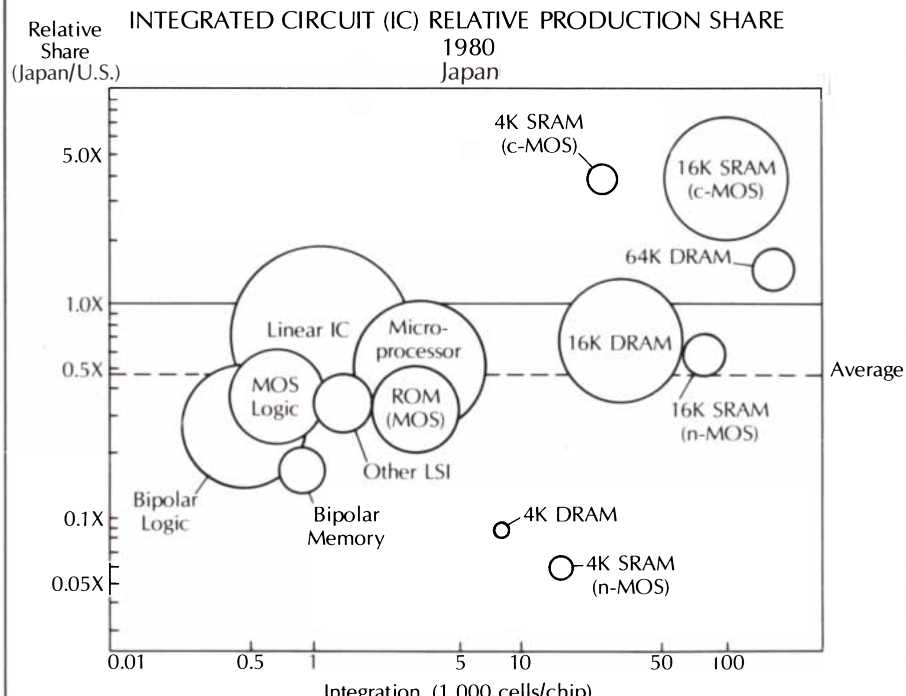
Discussion of the "new industrial revolution" in Japan begins with microelectronics, and in particular, semiconductors. From this basic technology, the revolution is seen especially as comprising applications in manufacturing technology. Here Japan's position in numerically controlled machine tools and in robots is emphasized, with robots arousing a more intense interest in Japan than elsewhere.

Computers are seen as part of the revolution, of course, not least in terms of computer applications in CAD/CAM, but also in wholly automated factories. Office automation has a second position to factory automation in Japan, but facsimile machines, copiers, word processors, telecommunications and optical fiber highways to link these and other kinds of office equipment are a focus of discussion and of development effort.

This seems to be a reversal of the U.S. pattern. More attention has been paid in the United States, until recently at least, to developments in the office. No doubt, this reflects in part the continuing shift of the U.S. economy from manufacturing to service. Japan, as will be further noted below,

TABLE 1

INTEGRATED CIRCUIT (IC) RELATIVE PRODUCTION SHARE 1980 Japan



Sources: Electronics, Dataquest, Newspaper Articles has concentrated on factory technology and methods, to its considerable competitive trade advantage. The United States seems now to be trying to refocus its efforts, but remains most active in office technologies.

Japan has long been highly receptive to automation. Take vending machines as a simple example. There are as many vending machines in Japan as in the United States, although Japan has half the population of the United States. In a Tokyo neighborhood, vending machines dispense rice, milk and batteries for household appliances, as well as the usual range of juices, soft drinks, beer and the like. Japanese vending machines supply hot noodle meals as well as reading material, and their distribution is aided by the fact that vandalism is not a problem.

Machine tools similarly represent the initial level of automation. Numerical controls increase the level of automation, while robots represent a still higher level, and intelligent robots will extend automation further yet. This appears to be more an evolutionary process than a revolution. It is perhaps the increasing pace of evolutionary change in these technologies, as in semiconductors, that causes the current, rather romantic Japanese view of a "new revolution."

The driving forces behind these changes in the factory and workplace are, of course, the semiconductor and the products that semiconductors—or, more broadly, microelectronics—have made possible. The Japanese position in semiconductors is well known, publicized in part by the protests of U.S. producers against Japanese competitors and their impact in the marketplace. There are at least three aspects of the Japanese situation with regard to semiconductors that represent a recurring Japanese pattern in movement to new technologies.

First, the basic research and invention in semiconductors was carried out in the United States, with Bell Laboratories playing a major role. The Japanese entered the field rather late, and based their work on U.S. technology. This is a recurring pattern and may be seen in striking fashion with respect to robots.

The second recurring aspect of the semiconductor story is that having entered the field, the Japanese began to close the technological gap. At least in terms of degree of integration, each successive generation of product found the Japanese in a stronger position than in the previous generation. The initial gap closes, rather than broadens. Thus, the Japanese came in a poor second after U.S. firms in 4K random access memories; a close second in 16K RAM; and in a dead heat, at least, in 64K RAM (*see Table 1*). Hitachi was the first firm, however, to announce the availability of 256K RAM to customers for sampling, in 1982, with scale production announced for 1983. The Japanese apparently are now ahead in this sequence of advances.

As a result, Japanese firms are now among the leaders in production of semiconductors and components. Nomura Securities estimated that in 1980 the sales of Nippon Electric in this product area were about equal to those of Motorola, with Texas Instruments still the leader. Hitachi, Toshiba and National Semiconductor are closely grouped together behind the three leaders in sales. These statistics, however, underestimate the U.S. position, since they take no account of IBM and ATT, which produce only for internal consumption, but must be among the leaders in the field. (Note, too, that no European firms are seriously in this competition. Microelectronics seems to represent the future, but the drama is being played out between Japan and America, with Europe out of it.)

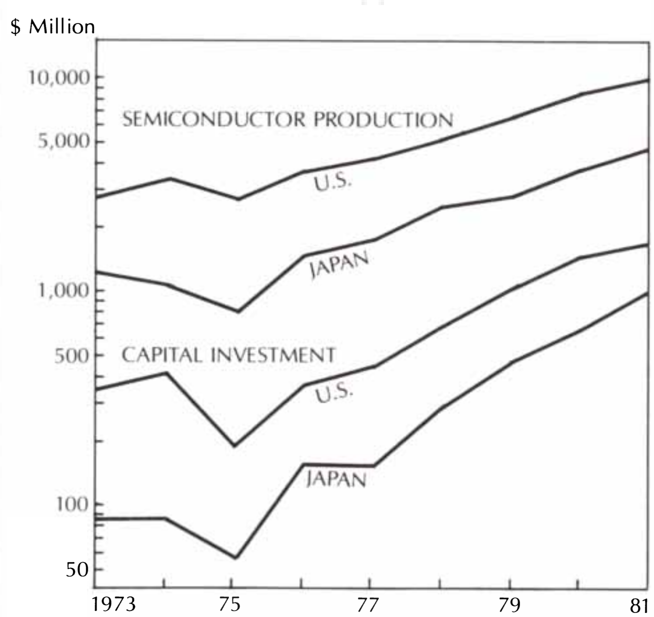
However, there is a third part to this pattern. As the chart indicates, while Japan's share is increasingly strong in a particular aspect of the technology, integration, Japan's overall position remains distinctly less powerful than that of the United States. Also, Japan is relatively weak in logic chips and has a position of no distinction in microprocessors (similarly an American invention.) This suggests that, as in factory automation, Japanese firms do less well in systems work and in inventing new technologies than they do in precision manufacturing and in engineering.

This is not to underestimate the position of the Japanese, but rather to suggest that there is a typical Japanese approach to technology and that Japan does not yet have a clear and important position over the whole range of technological development. In this context, another Japanese pattern should be noted (*see Table 2*). Very high rates of capital investment are required to stay in so fast moving a game as semiconductors. This is a game that the Japanese are quite prepared to play. If sheer amount of capital investment determines the outcome, the U.S. position is in some danger, as growth rates of Japanese investment substantially exceed those of the United States, even though U.S. rates are high.

The problem for the United States in regard to investment must be worsening now as a result of U.S. fiscal and monetary policies as compared with those of Japan. Current interest rates in Japan are approximately eight percent; U.S. interest rates are more than twice as high. Investment attractive to the Japanese company at its cost of money may well be impossible for a U.S. firm to undertake, because of its higher cost of money. This may well explain some of the divergence in investment rates indicated in Table 2 and may well force a further divergence, which will ultimately be reflected in the production figures. While the U.S. government tightens monetary

TABLE 2

CAPITAL INVESTMENT AND PRODUCTION IN SEMICONDUCTOR INDUSTRY



Average Growth 1973-1977	1977-1981
9.6%	24.0%
8.5	27.8
6.0	37.8
15.5	58.5

Sources: MITI, *Dataquest*

TWO YEARS AGO IT WAS JUST AN IDEA.

Foresight 7000 Pioneer Artists' "The Music of Melissa Manchester" Simulated TV picture.

It's just about two years since Pioneer Video came into being. In the beginning, it was a name and a commitment to the video industry. Back then, we vowed to do in the video industry what we had done so notably in audio.

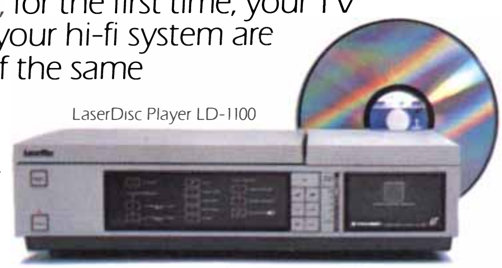
Large Screen Projection TV LS-501

Our first product was LaserDisc.™ It changed the industry. Already, it has become the highest standard for sight-and-sound reproduction.

We quickly followed the introduction of LaserDisc with the Pioneer® Projection Television. A television that offers nothing less than one of the sharpest, brightest pictures ever seen on a projection TV. It has turned the concept of big-screen TV from a good idea into an astounding reality.

This year we've done something perhaps even more innovative. We've changed the very nature of television itself. We introduced Foresight.™ One of the world's first true video-component systems. It's a system that combines a monitor that delivers a picture substantially better than anything ever seen on a standard television (astoundingly, more than 40% better video resolution); combined with a TV tuner capable of handling 127 cable-ready channels, all your video games, VTR and LaserDisc. Clearly, Foresight is the biggest step forward in the quality of television sight and sound since black-and-white changed to color.

But Foresight does more than offer a great picture. It offers great sound. You see, it comes with a stereo amplifier and 3-way speakers. The kind you'd buy if you were looking for audio alone. (Or you can use your existing amp and speakers.) So, for the first time, your TV system and your hi-fi system are both parts of the same system.



LaserDisc Player LD-1100

Two years ago we said, "This is just the beginning. Video is in its infancy. But you can be sure as video grows, Pioneer will be there leading in its innovation and development." In video, Pioneer is the perfect name for us.

PIONEER®

Does this look like an impossible photograph? It wasn't—for **National Geographic** pro and nationally-published **Black Star** photojournalist Jim Sugar. And the incredible Olympus OM-2.

Far above the city lights, Sugar flew—and fired his motorized, remote-controlled OM-2 attached to the plane's wing. Automatically, he took 36 exposures. **All were perfect.**

What makes OM-2 perfect for professionals makes it unbeatable for you. Its revolutionary OTF™ system sets exposure automatically, **off-the-film.** So if the light changes during exposure, OM-2 changes exposure. Instantly, automatically. As no ordinary SLR can do.

OM-2 lets you explore new realms in automatic flash photography, with its unique Olympus electronic flash units. The world's first with OTF™-automatic bounce, close-up and multiple flash control. OM-2 even confirms exposure accuracy. Most cameras can't.

And when you're shooting for the top, OM-2 is ready. With dozens of laser-sharp Olympus lenses, superbly-balanced compact motor drive, 300 professional components in all. From the OM System—largest compact system in sight.

See the OM-2 now at your Olympus dealer. And, for information and a free poster of Jim Sugar's incredible shot, write Dept. 222, Olympus, Woodbury, NY 11797. In Canada: W. Carsen Co. Ltd., Toronto.

OLYMPUS® OM-2

The camera with the press card

JAPANESE TECHNOLOGY TODAY

ing, Japan had in place several times the number of robots as did the United States and was very far in advance of the rest of the world. Professor M. Hasegawa of Waseda University reports on a study by the American Robot Association on robots installed in the world (*see Table 6*). This report uses a far more generous definition of robots for countries other than the United States (all statistics on numbers of robots are suspect because of definition problems). Whatever the definition, Japan's unusually fast rate of adoption of robots is clear. At the end of 1981, Japanese were boasting that they had 100,700 robots at work in their factories.

Mindful of problems of definition, it is clear that robot production and use in Japan is in a period of exponential growth. By some estimates there are now more than 200 companies in Japan producing robots, many for their internal consumption rather than—at this point—for the external market. In May 1982, Nikkei Sangyo Shimbun published the results of a survey of the industry. 63 companies took part in the survey, including most of the major robot producers. Robot production is referred to in the industry as a "doubling game" according to this survey, because the value of production increased by 108 percent in 1980 over 1979, and by 107 percent in 1981 over 1980.

Japanese producers, it should be noted, have some experience with doubling games. The exports of numerically controlled lathes doubled each year from 1970 to 1979. Production and export of video tape recorders have gone through a series of doublings. Hand-held calculator production more than doubled each year from 1967 to 1973.

This Nikkei Sangyo survey gives some notion of the intensity of effort key companies are expending on robot development. Hitachi is reported to have deployed 500 researchers in the field, and four other companies—Ishikawajima Harima, Mitsubishi Electric, Matsushita Electric and Toshiba Electric—each have more than 100 researchers on the devices. The thrust of development is toward higher speeds, sight and touch sensors and miniaturization. 24 of the companies surveyed are now exporting robots; 10 more plan exports in the near future.

The value of robot production for these companies is reported to have been about $200 million in 1980 and nearing $500 million in 1981. The Japan Industrial Robot Association projects values of well over $1 billion by 1985, and well over $2 billion by 1990. For many companies, robot production is now rather limited, but as the market size projections suggest, plans are ambitious. Kawasaki Heavy Industries, Japan's pioneer in robot production, plans to increase from its current production level of 70 units per month to more than six times that level within five years, to 430 units per month. Fanuc and Kobe Steel are aiming at 500 units per month in five years, while Mitsubishi Electric has the greatest reported ambitions. While Mitsubishi Electric now produces only 20 robots monthly, it states a target of 1,000 per month production within five years.

Japan produced some 24,000 robots in 1981 (*see Table 7*). Another doubling puts that at 50,000 in 1982. The typical producer is apparently planning to triple production over the next four to five years, suggesting output by then of 150,000 to 200,000 units per year. If these levels of output and installation are achieved, it might be appropriate to speak of a new industrial revolution.

As would be expected, this very rapid growth is resulting in a broadening of the products, both in types and applications of robots, and in types of users. It is fairly often charged that the Japanese definition of robots is broad, and takes in manual manipulators and fixed-sequence robots, while a more demanding and appropriate definition is that of a programmable, multipurpose manipulator. In the rather pedantic definition of the Robot Institute of America, a robot is a "reprogrammable, multifunctional manipulator designed to move material, parts, tools or specialized devices through variable programmed motions for the performance of a variety of tasks."

Japanese robot development is moving fast to bring the level of sophistication of output up to the most exacting definition. The year 1980 has been referred to in Japan as the first year of the robot era, because of the breadth of acceptance and the level of technology embodied in the machines by that time. Playback robot production nearly tripled in 1981 over 1980, while NC robot production increased nearly 10 times in that single year. These are no longer simple "pick and place" machines.

Just as the level of sophistication of the robots made in Japan is rapidly increasing, so is the range of users broadening. As in the United States, robots in Japan were initially sold largely to the automobile makers, and applications were dominated by arc and spot welding robots. However, in the first year of the robot era, as Japan sees it, the electric appliance industry moved ahead of the auto industry as the principle user of robotics. Thirty-six percent of shipments in 1980 were to the electrical products industry, and no doubt that trend has strengthened since. Matsushita Electric Industries has about 5,000 insertion robots in use now, but its president has announced that the company expects to have between 50,000 and 60,000 assembly robots in place by the end of this decade.

After this brief review of robotics in Japan, it is rather more understandable to hear the Japanese speak of the new indus-

TABLE 6

NUMBER OF INDUSTRIAL ROBOTS INSTALLED BY COUNTRY, 1981

Country	Number of Units
Japan	67,435
United States	4,100*
West Germany	11,400
USSR	3,000**
Switzerland	8,050
Czechoslovakia	531
United Kingdom	371

*Simple units are not included in the U.S. numbers.
**Estimate

Note: No data was available on several countries, including France and Sweden.

Source: Nihon no Saishin Gijitsu Series, *(Japan's Newest Technology)*, Robot Gijitsu Hyakka, p.16

TABLE 5

WORLD SHARE IN LEADING PRODUCTS
(Percent)

	Japan	United States	West Germany
Integrated Circuits (1979)	24	72	4
Computers (1979)	20	70	10
Video Tape Recorders (1979)	95	—	5
Numerical Controls (1979)	62	33	5
NC Machine Tools (1979)	58	29	13
Aircraft and Aircraft Engines (1978)	3	88	8
Antibiotics (1979)	62	30	8

Source: Moritani, ibid., p. 143

Even the sound of the door closing is by design.

Our doors speak well of the rest of our car. But their tight, solid sound is no accident.

Before we begin building the car, we engineer the best way to assemble it. Piece by piece.

Doors, trunk and hood fit the body with close tolerances all around. We align hinges and latches to work freely.

Window glass is held snug. And runs in the window channels smoothly. Triple rubber door seals help reduce outside sounds and wind noise. We've improved the insulation throughout the car.

The Honda Accord 4-Door Sedan is not only quieter, it's also roomier and more comfortable. You'll find it rides and handles better.

And while its many luxurious features are impressive, they are standard equipment. Like cruise control and variable-assist power steering, to name but two.

Honda's simple philosophy of building cars that work the way they were designed to work has made Honda number one in owner loyalty among all leading import cars*.

It has helped Honda cars retain a high resale value, among the best in the automotive industry.

You can fully appreciate how a Honda is made the next time you're in a Honda showroom.

After you kick the tires, listen to the doors.

HONDA
We make it simple.

*Source: R. L. Polk and Co. 1981 model year *New Car Buyer Analysis*. ©1982 American Honda Motor Co., Inc.

TABLE 8

CAPITAL INVESTMENT AND LABOR PRODUCTIVITY
(Percent)

	Rate of Equipment Investment to GNP		Capital Stock per Employee, Change per annum		Labor Productivity Change per annum	
	Japan	U.S.	Japan	U.S.	Japan	U.S.
1960-1973	15.9	10.0	10.4	2.6	8.5	2.2
1973-1979	16.2	10.4	6.1	0.9	3.0	0.5

Source: Economic Survey of Japan 1980/1981. Economic Planning Agency, Japanese Government. The Japan Times, Ltd., Tokyo, 1981.

trial revolution. It is the introduction of microelectronics into the full range of worker functions, with emphasis on what the Japanese are calling mechatronics, that is, the combination of more traditional mechanical technology with electronic technology. The NC machine tool and the robot are the epitome of this combination. These technologies are moving toward sensate machines and machines that produce machines. If these technical advances continue, and they almost certainly will, and if the social system can tolerate the impact of these rapid changes, the effect on production economics and on the nature of the manufacturing workplace will indeed be revolutionary. Japan is currently the leader in these changes. The Japanese economy and Japanese society may well lead the way into this revolution, to the unmanned factory, and to computer control of entire production systems.

The history of the Japanese position in robotics is in many ways reminiscent of Japan's movement into other technologies. A good deal of work on numerical control and on robots had taken place in the United States in the 1950's and 1960's. As is often the case in the United States, much of this work was directly or indirectly subsidized by the U.S. government. The U.S. military, especially the Air Force, sponsored NC work. The Atomic Energy Commission was concerned with robots, as was NASA. The article by R.U. Ayres *et al.* documents much of this history and notes that "the first industrial robot was marketed by Planet Corporation in 1959. It was (and is) a 'pick and place' device." (Ayres *et al., op. cit.,* page 87.) Unimation was next on the scene and remains the leading U.S. producer of industrial robots.

TABLE 7

ROBOT PRODUCTION IN JAPAN

Year	Units	Cumulative Units
1968	200	200
1970	1,700	2,300
1972	1,700	5,300
1974	4,200	12,000
1976	7,200	23,600
1978	10,100	42,300
1980	19,900	76,700
1981	24,000	100,700

Source: Japan Industrial Robot Association

A licensing agreement between Unimation and Kawasaki Heavy Industries in 1968 led to the first Japanese manufacture of robots. The Japanese also entered into licenses to import robot technology from Norway, Germany, Sweden and the United Kingdom. Again, this has not been an instance of Japanese invention (although no doubt invention will arise from current Japanese work). The initial discoveries and developments took place in the West. However, Japan has proved a more fertile ground for the nurturing of these initial discoveries, and is the undisputed leader in the field now.

The Japanese are now suppliers of technology. Ayres and associates note, "... In October, 1977, Kawasaki introduced its Unimate 6000 series multi-arm robots. While this was a Unimation robot, the continuous path technology the multi-arm was based on was developed by Kawasaki. Since Unimation did not make such a robot it distributed the Kawasaki Unimates through its sales organization in Europe and the United States." (Ayres *et al., op. cit.,* page 96.)

General Motors and Fanuc have agreed on a joint venture in the United States to manufacture a machine-tool-loading robot and an assembly robot designed by Fanuc, as well as a spray-painting robot now produced by GM. Hitachi has licensed its microcomputer-controlled robot technology to General Electric. Komatsu and Westinghouse are reported to be negotiating a technological exchange: Komatsu would supply its flexible manufacturing system knowhow to Westinghouse, and, of special interest, Westinghouse would in exchange supply CAD/CAM technology to Komatsu (another instance of the stronger U.S. position in software and systems technology.) In a little over a decade, Japan has taken in a new technology, applied and further developed it, and seized leadership; Japan is now supplying advanced technology to the leading manufacturers in the West.

It is this achievement in robots, and in other mechatronic and electronic fields that has done so much to increase Japanese self-confidence and to cause them to begin to see themselves as leading the world into a new industrial system. To note the limits of the Japanese achievement, in that the Japanese were not the initiators of the technology, is not to belittle the achievement. For many decades, U.S. confidence and pride in its technical achievements was not limited by the fact that much of the basic research and invention had been imported from Western Europe.

Perhaps it is worthwhile here to reiterate in the context of Japan a point that has been made before in other contexts. In a world in which scientific discovery has been increasing at an exponential rate, and in which scientific knowledge flows reasonably freely, it is not necessary for a nation to be the source of discovery and invention in order to achieve a high level of technological sophistication. Rather, if the economy is a dynamic one, growing rapidly and investing heavily, it will attract inventions and will be the place where those inventions most quickly reach commercial application and widespread use. The United States was in that position for a long time; Japan is in that position now.

To understand the ability of the Japanese to move so rapidly in microelectronics and factory automation, it is necessary to understand the context in which this progress is taking place. The driving force is investment in productive facilities and aggressive competition for market leadership. The facilitating mechanism is in part the government, which by its tax and other initiatives can encourage—or work to suppress—a direction of development. The setting is the workplace, and the quality of the labor force, its attitudes and motivations, are determinants of whether the technology can and will be applied.

The issues of high savings and investment have moved to the center stage of U.S. discussion in recent years through the "supply side economics" now associated with the Reagan administration. Whatever the merits of the approaches being followed in the United States to stimulate savings and investment, Japan serves as a prototype of supply side economics and a model for high rates of investment. Gross capital formation as a percent of gross national product has been 30 percent or

more in Japan for a long time, compared with 16 to 18 percent in the United States. A nation, like a company, can in the long run grow no faster than it invests. Japan has continued to invest heavily and to grow more rapidly than other economies.

In terms of investment in plant and equipment, Japan's rates of increases in investment have been very high (*see Table 8*). The growth rates of capital stock per employee have been several times higher in Japan than in the United States for the past twenty years. Not surprisingly, labor productivity has been increasing several times more rapidly in Japan than in the United States over the same period. While the level of rates of increase in Japan has fallen in the post–oil crisis period, the advantage in rate of increase over the United States remains unchanged. Only a sudden and quite improbable reversal of this trend could prevent Japanese levels of output (and subsequent income) from exceeding those of the United States in the near future.

These are averages, of course, and some care needs to be used in considering them. Japan has sectors that are doing very badly. Notable among these are the energy intensive, raw material and pollution intensive industries such as nonferrous metals, bulk petrochemicals, chemical fertilizer, paper and pulp and oil refining. These industries, among Japan's star performers in the late 1950's and after, are now in deep trouble, unable to compete with world-scale competitors, and in the process of phasing out much of their capacity as they yield to offshore sources.

Japan's strength, the focus of investment, and the source of the overall high rates of increase in productivity are now the mass production industries in the middle range of the technology spectrum. It is in the production of machinery—autos, cameras, copiers, electrical appliances, machine tools, and the like—that Japan now has competitive advantage. As seen, the shift within this broad area is to higher levels of technology and value-added, especially through the application of microelectronic technology.

In these production and engineering based industries, leading Japanese producers have achieved startling levels of production efficiency. In a recent study by an American automobile company, the conclusion was reached that Japanese productivity exceeds that of the U.S. company in its American and European plants by about three times overall, when comparable production units are compared (*see Table 9*). The Japanese company has higher direct labor productivity, much lower overhead rates, and the work force operates with much less supervision.

In this comparison of automobile companies, the U.S. observers reported no unusual technology in the Japanese plants; the machines used were familiar. What they found instead was that in Japanese plants meticulous care was taken in each step of the total production process, with less space required; machine cycle times were slower, but the machines were utilized to a far greater extent; setup time on the machines was very fast (five minutes, compared with four or five hours in U.S. plants); and the levels of work in process inventories were very low. Much of the increased efficiency is credited to the *Kanban* system of "just-in-time" production and inventory delivery.

Kanban deserves detailed analysis as a system, and no doubt the term will become increasingly familiar abroad as one of those rare Japanese words, like *tsunami* and *sushi,* that enter Western vocabularies. The system has already been misunderstood and abused in the West by some companies as a means of squeezing parts suppliers and subcontractors. It is not such a system. It is a highly sophisticated approach to production and inventory scheduling that, in time, is extended to suppliers as well to maximize the advantages of the entire system. It focuses on eliminating all waste in the system—waste of time and of materials.

Kanban was developed through the 1960's and 1970's, with Toyota as the pioneer. (Remember the Toyota Group as first Japanese users of a robot.) The system requires very high levels of investment in the most modern machine tools and tooling, and encourages the rapid introduction of automation in the factory. These automation devices provide the further great advantage of improving quality. Thus a loop is closed.

Japanese companies began early to develop quality control circles, efforts to motivate employees to work together to improve output levels and quality. *Kanban* as a system took advantage of the quality control circle by focusing on fully utilizing the workers' abilities. This in turn has led to a willingness to introduce more sophisticated production methods, and again, quality is improved.

What is noteworthy in this development by Japanese producers of *Kanban* is, first, the preoccupation of the Japanese company with production engineering and the focus of management attention on the shop floor. This alone is in marked contrast with many Western companies in the 1960–1980 period. The second aspect is the willingness to invest in machinery and tools to improve output and quality. The result is the Japanese advantage in overall productivity increases that have been noted, despite the existence of sectors in which competitiveness can no longer be maintained.

Frequently, too much is made in Western analysis of Japanese use of bank borrowings to explain this willingness to invest. Toyota has no bank debt, for example, nor does Matsushita Electric. These two giant companies have used bank borrowings in the past, during their periods of early growth. They no longer need to do so, and have paid down their borrowings. Certainly there is a willingness of Japanese banks to extend levels of credit that would be impossible in the West. This does not explain the investment level, however. Rather, it simply describes one source of investment funds. It might be noted too that the ability of Japanese banks to take very large positions in companies is due in no small part to their conviction that the central banking authorities will avoid a credit crunch and will maintain sufficient liquidity in the system to make these levels of loans tolerable. Not all Western central banks have such credibility.

Investment in plant and equipment is the one way for the Japanese company to secure its future. Businesses in Japan cannot be disinvested, through sale for example. A company cannot grow by purchase of an-

TABLE 9

PRODUCTIVITY COMPARISON OF
JAPANESE AND AMERICAN COMPETITORS
(Indexed to the Japanese Competitor)

	Japanese Competitor	American Competitor Plant 1	American Competitor Plant 2
Units/Day	100	105	86
Labor/Unit/Day			
Direct	79	146	121
Indirect	17	221	96
Salaried and Other	4	67	25
Total	100	434	252

other successful company. Yet with a fixed labor force, and the close identity between company and employee that prevails in Japan, the company's future must be secured, and this can only be done by improving cost position, quality, product design and general international competitiveness. This compulsion is no less for medium-sized and smaller companies. Indeed, it is this small business sector that in Japan has become a principal market for NC machine tools and promises to be a major market for robots.

One of the notable aspects of automated factory systems as they are now emerging is their suitability for small runs of parts and for rapid changes in production mix. To a large degree, this type of automation removes some of the advantage of scale production. Put another way, it makes possible the production of a wide variety of parts and products without incurring the cost penalty that might otherwise be encountered. NC machines and robots can be rapidly switched from one type of run to another without the machine downtime that results on traditional equipment. When *Kanban* is combined with this machine capability, to help ensure no cost handicap from parts and product inventory, the result is very high levels of productivity and cost efficiency. This is what is happening now in Japan's mass production industries.

Just as high levels of equipment investment and a management focus on production efficiency are the driving force behind Japan's move to a new industrial revolution, so the Japanese government plays an important role as facilitator of the process.

government was especially farsighted in identifying robots as a key sector for encouragement. The early interest in importing the technology and in entering into robot production was entirely on the part of the private sector.

However, as private sector interest grew, government attention was drawn to the field. Industrial robotics was one of the industries designated for promotion under the 1971 law for Provisional Measures to Promote Specific Electronic and Machinery Industries. This law included tax and financial provisions to promote the improvement of Japanese production technology. In recent years, government support for industrial robot development has become even more specific and concrete.

From 1978, under the Special Measures Law Concerning the Promotion of Specific Machinery and Information Industries, any company installing a robot can write off 12.5 percent of the cost in the first year in addition to normal depreciation. This law also allowed small- and medium-sized companies to obtain low-interest loans through the Small Business Finance Corporation in order to purchase robots.

As interest has intensified, other measures have been taken. According to the Japan Economic Institute's report No. 46, "24 robot makers and ten insurance companies formed the Japan Robot Leasing Company, Ltd. (JAROL) this past April (1980) to enable small and medium sized manufacturers that cannot afford to buy robots outright to lease them. Using funds supplied by the government-owned Japan Development Bank as well as their own monies, JAROL buys robots from builders and then leases them to qualified buyers on terms more advantageous than those offered by Japan's 40 or so leasing companies. (So far, the leasing companies have shown little interest in robots, as indicated by the fact that they handled only about 10 percent of all robots sales in Japan last year.) Equally important, JAROL which is similar to the highly successful Japan Electronic Computer Company, Ltd. provides

ices to its customers."

Indeed, now that the world has entered "the era of robots," the field of robotics has received the official cachet of becoming the subject of a project of the Ministry of International Trade and Industry (MITI). MITI had earlier sponsored work on unmanned factories and on flexible manufacturing systems. The field is not a new one for the Ministry, which is highly aware of and responsive to emerging interests of its constituency, Japan's manufacturers.

MITI's proposed project on robots is ambitious, as described by The Japan Economic Journal on May 18, 1982. "The Ministry of International Trade and Industry has decided to launch from fiscal 1983 a semi-long range joint government-industry project to develop highly intelligent robots for scientific, economic and social purposes.

"This is because it feels that the many different types of robots so far developed in Japan are still far from satisfactory or in the 'stage of infancy,' although Japan's robot industry may be internationally rated as the best developed.

"The seven-year project, starting from April, next year, will involve spending 17 billion yen, though this still is subject to the government's budgetary approval. It calls for spending 2 billion yen in the first year for fundamental research and undertaking of the entire research and development job by a joint research association to be formed by the Electro-technical Laboratory of MITI's Agency of Industrial Science and Technology, and about 10 major robot, computer and machine makers.

"The specific study themes will be to: 1) create a high-sensitivity sensor capable of 'seeing' and 'touching' for object recognition; 2) develop an ultra-midget high-efficiency processor to control its 'sense of vision'; 3) produce a new high-versatility robot arm placing everything at a high precision and moving everything including heavyweight objects; 4) develop a new high-efficiency motor specializing in working the wanted robots; and 5) reduce the weights of the proposed robots far below the existing ones and find out the best materials to make the wanted robots. Universally applicable types of intelligent robots are the most desirable."

This will be recognized immediately as a MITI initiative parallel to the program in very-large-scale integrated circuits of the 1970's. The era of the robot has indeed arrived if the field is seen to warrant an effort of this magnitude. If the amounts of government money involved seem small (and they are), it needs to be remembered that this is seed money and will be multiplied by the investments of participating companies in the program.

In other words, the field of industrial robots is now getting the sort of attention that the Japanese government has paid to com-

TABLE 10

EDUCATION LEVEL OF NEW ENTRANTS TO THE LABOR FORCE

Year	Number of Entrants (K)	Middle School (%)	Higher School (%)	College and University (%)	Total (%)
1950	832	86	14	Less than 1	100
1955	1,125	62	30	8	100
1960	1,376	49	42	9	100
1965	1,500	42	47	11	100
1970	1,374	20	59	21	100
1975	1,042	9	57	34	100
1980	1,103	6	55	39	100

Source: Appendix 132, Waga Kuni no. Kyoiku Suijun *(The Educational Level of Our Country)*, 1980 Edition, Ministry of Education, Tokyo, 1981

puters and semiconductors. As with those areas, the government did not take the initiative in directing the attention of industry to the field. On the contrary, the initiatives were from Japan's thrusting, probing private sector. Now, however, the government has put in place the tax, financing and development support to encourage rapid further growth in the field.

We come now to the heart of the matter in considering the probability of this new industrial revolution taking place in Japan. The technology is in place, or within reach. Management is willing and able to invest. The government is encouraging the development of factories of the future. Ultimately, however, the feasibility and pace of these potentially revolutionary changes in the mode of production rests with the ability and willingness of the work force to cooperate with and support the changes.

The real test of all this is yet to come. The population of robots remains small, although it is growing explosively. The full magnitude of the potential change is perhaps not yet clear. What are the prospects for the Japanese labor force accommodating itself to these changes in the 1980's?

Two great currents of change in the nature of the Japanese labor force must be considered when evaluating the probable reaction to full automation. The first is the explosive increase in the level of education over the past generation. The second is the profound change in the age structure of the working population that will occur in Japan over the next two decades.

A generation ago, in the mid-1950's, most of the new entrants to the labor force in Japan were graduated from middle school (*see Table 10*). That is, most children completed their compulsory education and then entered the labor force. It was not until the mid-1960's that as many as 10 percent attended colleges and universities before entering the labor force.

At the present time, virtually all new entrants to the Japanese labor force have completed higher school, and nearly none end their education at the middle school level. The proportion entering the labor force from institutions of higher learning is now very much higher in Japan than in Western Europe and is approximately at U.S. levels.

Clearly, the kind of work that these current new entrants to the labor force are capable of handling is very much different than those of 20 years ago or so. Furthermore, one can assume that the kind of work these higher school and college graduates are willing to do is very different than that of the middle school graduates of a generation ago. Routine assembly line work is unlikely to satisfy these new employees. Further, the price they can command on the labor market makes it uneconomical to employ them in routine jobs for very long.

Note in Table 10 as well that the number of new entrants to the labor force peaked in the mid-1960's in Japan and has been declining rapidly since. It is a simple and inescapable fact that the Japanese economy and Japanese companies are faced with the prospect of a work force diminishing in size and increasing in age.

The consequences of all this are wide ranging and profound. Consider first the issue of an increasingly demanding and capable work force, at an increasingly high price level. This is not a problem peculiar to Japan, although the change has taken place there in an especially short period of time. West Germany, for example, has many of the demographic characteristics of Japan, in the sense of a declining younger population and relatively high priced labor.

The problem of who will do the simple and dirty tasks has in Western Europe and the United States largely been dealt with by immigration. The *gastarbeiter* of West Germany, the Hispanic immigrant to the United States, and the North Africans in France have provided a new source of inexpensive and relatively undemanding laborers. This has had the effect of keeping in place longer than was appropriate those labor-intensive businesses and industries that properly, from a long-term industry structure point of view, should have been exported to less developed countries. That is, rather than exporting the jobs that are no longer appropriate for the economy's stage of development, labor was imported. Indeed, a striking characteristic of the United States in the 1970's was the large number of new jobs added in the economy—additions to the total employment. The United States at least seems to have moved toward being more labor intensive than capital intensive.

The consequences of the use of immigrant labor are only now becoming clear. First, its use has greatly delayed the restructuring of these Western economies, which are now trying to deal with high levels of unemployment and with the consequences of remaining overlong in inappropriate industries. Government attention is focused on preserving employment rather than on redirecting and increasing the flow of capital investment to higher-value sectors. Further, a major social problem and social cost has been created by the difficulties of integrating this immigrant labor into the society and providing social welfare facilities to deal with the economic and social problems now created.

In the case of Japan, there has been no immigration, despite the very large numbers of underemployed and low wage rate Asians in nearby countries. Of all the developed countries, Japan would seem in the best position to use immigrant labor to perform the low-level jobs, and from the rapid increase in the size of the economy, rapid run-up in wage rates, and decreasing supply of younger labor, Japan would seem most in need of immigrant labor. However, politically and socially, substantial immigration into Japan is simply not possible or acceptable.

As a result, there has been a massive transfer of textile, light assembly and other low-value-added operations from Japan to such neighboring, low-labor-rate countries as South Korea and Taiwan. This relieves some of the pressure on the proper utilization of a highly educated and priced work force. It also has very much assisted Japan in dealing with the industry restructuring being forced on the developed economies by changes in the world economic environment.

There remains, however, the question of who is to do the simple, repetitive jobs in the industries that the Japanese economy is now focused on. It is at this point that factory automation becomes feasible and attractive. In many jobs, automation can to a good degree substitute for less-skilled labor. The robot is becoming an economical-

TABLE 11

INTERNATIONAL COMPARISON OF SOCIAL SECURITY LEVELS

	Contributions as Percent of National Income	Percent of Population Over 65 Years of Age: At Present	Percent of Population Over 65 Years of Age: year 2000
Japan	12.7%	9.3%	15.6%
United States	16.5	11.2	12.2
United Kingdom	19.9	14.5	14.9
West Germany	28.7	15.3	15.5
France	28.9	14.1	14.2
Sweden	36.2	15.9	15.9

Source: Asahi Shimbun Reporting Ministry of Health and Welfare Findings


SHIFT INTO AWESOME.
NEW 5-SPEED DATSUN TURBO-ZX.
Quick. So quick, this fuel-injected whirlwind can reach the legal limit almost before you can finish reading this sentence.
So effortless, mere flicks of the new 5-speed shifter smoothly boost you to cruise velocity.
So surefooted, you surge confidently through curves on a combination of 15-inch Potenza radials and precisely tuned suspension.
So luxurious, the cockpit borders on opulence: T-bar roof, 4-speaker digital stereo/cassette and integrated air conditioning are but a few of its standards–all built to Nissan standards.
In 2-seater and 2 + 2, the Turbo ZX is awesome…and then some.
DATSUN
WE ARE DRIVEN
NISSAN
Product of NISSAN
© 1982 NMC-USA

ly attractive labor substitute. By one estimate of the robot industry, the relative price of a robot to a worker dropped in the 1970's from about four times to 1.4 times, owing to rapid wage increases on one hand and increasingly competitive pricing of robots through cost improvements in production on the other hand. This trend will certainly continue, so that the robot becomes a cost effective labor substitute, even without taking account of the fact that the robot requires no fringe benefits or holidays, can work extra shifts, needs no tea breaks, does not engage in labor disputes and provides a consistent level of quality output that human labor cannot supply.

The second great and inexorable change taking place with respect to the Japanese labor force is the shift in age structure. During a period of massive effort to bring the national budget under control, a 12 percent increase in social welfare payments in 1981 over 1980 became a major news story. Contributions to social security as a percent of national income are not yet high in Japan compared with other developed countries (*see Table 11*). However, this is not because the payments are small in terms of level of social security support. Rather, this is largely due to the smaller proportion of aged persons in the population of Japan now, relative to Western Europe especially.

However, within two decades Japan will have as large a proportion of its population more than 65 years of age as any other country in the world. This is due both to a steadily falling birth rate and to a quite remarkable extension of life expectancy in the postwar period with greatly improved health services delivery systems. The consequences of this reduction in the number of new entrants to the labor force and great increase in the aged population would seem to be a quite rapid decline in the potential for economic growth in Japan. A principal resource factor will be in slowly growing supply and quality (*see Table 12*).

The development of robots and their potential for further development may mean that Japan can continue to substitute capital for labor and in the future do so quite literally in that the capital will supply labor-equivalent robots. Japan's highly educated labor force can then be increasingly focused on those jobs in which only human capabilities can be employed. It is interesting to speculate as to whether this provides the potential to raise per capita incomes in Japan to unprecedented heights.

This review of the two sweeping changes in the labor force, education level and age structure, makes clear enough the need and opportunity for Japan to move toward very high levels of automation. The issue remains, however, of the willingness of the labor force to acquiesce in this move.

To what seems from a Western view an unusual degree, there has so far been little controversy in Japan about the shift to automation and the enthusiasm for the introduction of robots. This is understandable in that the robot population remains small and has been dedicated largely to repetitious, hazardous and dirty jobs that the work force can give up with pleasure and relief.

There is a further factor that limits Japanese resistance to automation. That is the general pattern of employment practices in Japan. In large firms at least, employment for the full career of the full-time male employee is assured. He will not be laid-off or discharged except under the most extreme of conditions. Further, his compensation remains largely determined by the length of his service with the firm. Thus, the introduction of technology represents no threat either to job security or even to compensation. Rather, to the extent that the new technology improves corporate profitability and security, it similarly improves the employees' compensation and security.

In the West, especially in the United States, new and labor-saving technology is properly seen as a threat, and the Luddites of the beginning of the industrial revolution have their counterparts today. Especially in Western Europe, the preoccupation with the potential threat to employment of the new technologies seems great. This is curious, because the history of the industrial revolution has been that each major advance in technological level brings with it correspondingly larger employment opportunities. It may well be that the current industrial structure rigidities of Western Europe cause this heightened anxiety about the employment impact of new technologies, especially microelectronics.

In any event, the pattern of employment relations in Japanese firms seems especially favorable to the rapid introduction of high levels of automation in the workplace.

This applies as well to the structure of industrial relations and to the issue of trade unions. As would be expected, large manufacturing companies tend to be the most heavily unionized. Overall union membership in Japan is about 30 percent of the labor force, a lower proportion than Western Europe but a higher proportion than in the United States. The striking feature of Japan's industrial unions is that they are organized along company lines rather than by skill or industry. Thus a rather close identification is possible between the interests of the union and the interests of the company.

This is not to say that Japanese labor relations are simply benign. The union is meant as a check on management on behalf of the labor force and operates as such a check. It is to say, however, that the strength and continuation of the union depends on the strength and continuation of the company, and there is a clear limit on the extent to which the union is prepared to damage or constrain the interests of the company.

The union of Matsushita Electric Industries has been one of the first to speak directly to the question of the introduction of robots. As was noted above, Matsushita Electric has announced that it plans the introduction of some 50,000 to 60,000 assembled robots over the next decade.

The union has recently taken a public position on the issue. It first made clear that it does not oppose the introduction of robots. It has, however, stated that it insists that three principles be observed regarding their introduction.

The first principle of the Matsushita union is that there be established a system for prior consultation with the union regarding plans for robot introduction. Second, the union expects that job security will be recognized and provided and that retraining and reeducation programs will be introduced to assist worker adjustment to new tasks. The third principle is that improvements in productivity will be distributed to

TABLE 12

JAPAN'S AGING POPULATION

	Total Population (K)	Working Population (K)	Age Structure of Working Population (%)		
			15-44	45-59	60-
1960	93,419	45,110	68.7	22.1	9.3
1970	103,720	51,530	69.5	21.7	8.8
1980	116,916	56,500	62.5	28.2	9.3
1990	124,261	63,500	57.5	31.2	11.5
2000	129,234	66,630	52.5	33.7	13.7

Note: Japan's population peaks in numbers about 2010 at about 132 million and declines thereafter to about 109 million in 2050 by these projections.

Source: Japan Economic Research Center

the labor force in the form of shortened work hours.

The Federation of Metal Workers has recently taken a similar position. It cited its affirmative stance toward robot introduction, but under four conditions; 1) discussion with the union; 2) employment and working conditions maintained; 3) distribution of productivity improvements; and 4) proper attention to safety and health aspects of robot introduction. This union group also took note of the need for special consideration for middle-aged and older workers, whose problems of retraining and reassignment are likely to be greater than those of younger workers.

This concern with the impact of the robot on older employees seems at first glance curious. It might be expected that it would be the younger employees who are more often on repetitious jobs. Reference to the earlier data on education helps to explain the point. In the Japanese labor force, in addition to the normal increased difficulties of educating older persons, the older part of the labor force is much less well-educated. Their potential for reeducation and retraining is less, and therefore they are a greater source of anxiety in terms of being displaced by automation. They are too, by the way, higher paid and it is in the company's interest to displace older employees first. Thus the union concern over automation in Japan focuses on older employees.

A quite detailed study of the probable impact of microelectronics on employment was published in May 1982 by the prestigious Association for Scientific Technology and Economics.[3] In this study, industry was divided into three groups, according to the expected impact of microelectronics on the industry. The greatest impact is expected on A Group—electric and electronic industries, information industry, telecommunications, precision instruments and aerospace. Transport equipment, general machinery and industrial machinery make up B Group, in which moderate impact is expected. Steel, nonferrous metals, chemicals, pharmaceuticals, food, petroleum, glass, paper and pulp and construction make up C Group, in which little impact on employment is expected.

Looking ahead to 1985, overall manufacturing employment is expected to increase slightly (*see Table 13*). In the A Industry Group, where the greatest impact from microelectronics is expected, a rather substantial increase in employment is expected. The C Industry Group will be negatively impacted in terms of employment, suffering a decline in number of jobs without the benefits of increased growth. The numbers overall simply confirm again the fact that overall employment benefits from advances in technology.

This analysis goes on to attempt a qualitative assessment of impact, and like the unions, identifies older employees as being in a difficult position. The conclusions are as follows:

There will be a reduction in employment positions for workers employed in simple tasks. Because of the necessity of shifting to specialized work, middle-aged and older persons who have difficulty in readjusting will present a problem.

Skill requirements will change and there will be a need for employees with multiple skills.

Employment will increase in research and development, planning, production line supervision, security and the like, and there will be a decrease in employment in production and material handling.

The broadening of the scope of the workplace will make for increased worker motivation. However, job opportunities for older workers and for female workers will be reduced.

Because of the changes in the workplace and in job situations, reeducation and retraining will be important.

TABLE 14

HOURS WORKED PER YEAR IN MANUFACTURING (Excluding Part-time Workers)

	Total Hours
Japan	2,164
United States	1,987
West Germany	1,766
France	1,813
United Kingdom	2,031

Source: Rodo Mondai Kenkyu Iinkai (Labor Problems Study Committee) (Nikkeiren)

The Japanese unions' request for shorter work hours in consequence of greater automation seems an entirely reasonable one. There has been virtually no increase in 25 years in the number of Japanese companies that work a simple five-day week. Further, the number of hours worked per year by Japanese workers in manufacturing continues to be larger than in other countries (*see Table 14*). In Japan there is not yet the practice of extended holiday closing of plants, although the number of holidays and vacation days per year is increasing.

The Japanese Ministry of Labor recently ruled that overtime hours worked in industry (with a number of noted exceptions) should no longer exceed fifty hours per month. While that seems a great deal of overtime, Japanese unions expressed the view that a new law will be needed to keep overtime within this limit.

The room for reduction in work hours is considerable and will be another factor in easing the adjustment to automation. From the labor cost side, the improvements in productivity and resultant reduction in labor costs per unit will make possible reductions in working hours and real increases in wages without driving unit labor costs up.

It may be concluded that in the factories of Japan the stage is being set for still another phase of increases in efficiency. There seem to be no insurmountable obstacles to moving to yet higher levels of automation and productivity, and the technology for the new moves and the management initiatives are in place or in development. This may not constitute a new industrial revolution, but certainly can result in massive changes in the nature of the factory and the

TABLE 13

INFLUENCE OF MICROELECTRONICS ON EMPLOYMENT IN 1985 (Thousand Employees)

	1980 Employment	1985 Employment	Change	Amount of Change from Microelectronics
A Industry Group	1,630	1,780	150	390
B Industry Group	1,980	2,040	60	20
C Industry Group	7,320	7,200	−120	−250
Total	10,930	11,020	90	160

[3] Microelectronics no Shakaiteki Kokusaiteki Eikyo (*The Social and International Impact of Microelectronics*). Kagaku Gijutsu to Keizai no Kai. Tokyo, 1982.

role of labor in the factory. The prospect of being able to free the labor force from dirty, hazardous, monotonous and dreary tasks to work at a higher level of competence and dignity is an exciting one.

As this discussion implies, in Japan the focus on automation has been in the factory. It is only now that the office is being included. The reason is rather clear. Owing to the complexities of the written language, the first Japanese language word processor did not reach the market until 1979. In that year, Toshiba introduced the first machine, priced at nearly $30,000. Despite the price level, production grew to 2,500 units in 1980 and to 10,000 units in 1981. Prices by early 1982 had dropped to under 1,000,000 yen, or to less than $4,000 per unit—nearly one-tenth of the initial price. (The performance and features of the machines have been increasing along with the price drop.) 1982 production is estimated to reach 60,000 units. Given this explosive growth, no company has a dominant position yet and 34 companies are reported to have entered into production. Robots were termed a "doubling game." Office automation equipment, at this very early stage, is a "tripling game," at least. Just as the output of word processors tripled from 1980 to 1981, so did the output in Japan of personal computers. This product had an earlier start, and 320,000 units were produced in 1981—290 percent more than in 1980. As in the United States, pricing and distribution of personal computers is moving the product rapidly in the direction of becoming a new household appliance, with youngsters playing with the machines in department stores, in discount shops and in specialized outlets. Nippon Electric Company dominates this market with a more than 40 percent share of production.

The delay in moving to office automation was in large part due to the problems of dealing with written Japanese. There are two phonetic scripts for writing in Japanese in addition to the Chinese-derived characters. All three writing systems are usable, and are often used, in a single sentence. Each phonetic script has about 50 symbols, while the ordinary written language requires no less than 2,000 and more generally 3,500 to 5,000 of the written characters. The problems of input, retrieval and output of this mass of symbols have been enormous. Only the development of more powerful memory systems has made Japanese language word processors possible, and thus their introduction was long delayed compared with Roman-letter writing systems.

It is generally agreed that the office system in Japan has substantial inefficiencies, though they have seldom been measured. The most striking problem, at least to Westerners, has had to do with the typing of letters in Japanese. The Japanese language typewriter is about the size of an office desk, with a cumbersome and extraordinarily slow method of retrieving the typeface for a character, stamping it and replacing it. The process is quite like a very old printing machine. Generally only formal, and for obvious reasons, very brief letters have been typed.

The results of these difficulties with typing have had their positive aspects. The facsimile market has become very large, since facsimile machines can transmit handwritten documents with ease. Facsimile machines worth a total of $450 million were produced in Japan last year, including small models advertised on television by the national telephone company. Japan has taken the lead in world facsimile technology and production, having been forced to focus on this product by the awkwardness of its written language.

The situation with copying machines is similar. Japan produced 900,000 copiers last year, the market dominated by Ricoh with an over 40 percent share. These tend to be smaller than Western machines on the whole, because they also are intended to make copying of handwritten documents reasonably rapid and inexpensive.

Penetration of the corporate market by these new devices—Japanese language word processors and personal computers—appears to be rapid. Despite their recent introduction, a survey by Nihon Keizai Shimbun to which 78 companies responded showed that 58 of the 78 companies already had a total of 792 word processors, and 55 of the 78 companies had a total of 3,210 personal computers.

Thus the devices for increased automation of office activities are beginning to be put in place. Production of the older products—facsimile machines, office computers and copiers—is increasing currently by 25 to 35 percent annually, with output of the newer devices at least tripling. The next step is to begin to integrate these devices into systems. Nippon Electric has already announced the installation of optical fiber highway systems that will allow integration of telephone exchanges, building maintenance systems and the various office devices in terms of an integrated carrying system for digital and voice messages and storage. Clearly, the Japanese office is on the verge of sweeping changes in productivity levels and job content. There is no evident reason to expect that the process will be slower in the office than in the factory as the technology becomes available. Indeed, if the suggestions of low office efficiency are accurate, the changes in the office may be more rapid and drastic than in the factory. Again, there is no reason to believe that resistance to these changes will be a substantial obstacle to there taking place.

The anxieties commonly expressed in the West about automation, including industrial robots, center on the impact on employment. As noted, this is a concern in Japan as well, but the concern is tempered by two facts. First, job security in Japan, in the large firms at least, is a reality. Second, Japanese unemployment levels are relatively low—slightly over two percent of the work force by Japanese definition, and perhaps twice that level by the broader U.S. definition.

Further, it was only a decade ago that the principal bottleneck to continued rapid economic growth in Japan was seen as a labor shortage. While memories of the anxiety have faded, it is clear that the drop in new entrants to the labor force and the aging of the labor force would limit Japan's economic growth potential. Automation of the factory and office may well be the solution to this problem. Machines substitute for humans, and that limit on growth is thus removed. Finally, a labor force as well-educated as Japan's must be deployed in more valuable and complex tasks than repetitious office and factory work. Automation can fill the void that is left as the labor force is upgraded in skills and sophistication.

When Japanese executives are asked to assess Japan's current technological position and prospects, there is a rather broad range of views. Whatever their judgment on the technological race, they tend to share a view of the Japanese as being different in important ways from other peoples. Almost without exception they see themselves and their companies in a competition against foreign firms.

Mr. Nobutoshi Kihara, Executive Managing Director of Sony, is unusually confident of Japan's position and of his company's development capabilities. "I really do not think the technological development of other countries is so high. I have seen many foreign social systems, and people who have the natural quality, the sense for innovation, are really very few. Polaroid's Dr. Land is one.

"You cannot produce a new product if you do not have that sense. Even if you hit on a real possibility, you will not come up with the right idea if you do not have that sense. MAVICA, Sony's announced new electronic camera, is just such a case. I myself, looking back on my work on MAVICA, realize that I have that sense. You can get a lot of people together, and get a lot of noise, but if you want a complete device, you finally come down to one or two people. I do not know about other companies,

CELICA SUPRA

IT TAKES THE RIGHT STUFF TO BE IMPORT CAR OF THE YEAR.

The experts at Motor Trend Magazine know the right stuff when they see it. That's why they named Celica Supra the Import Car of the Year — over competition like Datsun 280 ZX Turbo, BMW, and Mercedes Benz! Motor Trend described Supra as "the stuff car nuts dream about...one of those cars that only comes along every ten years or so."

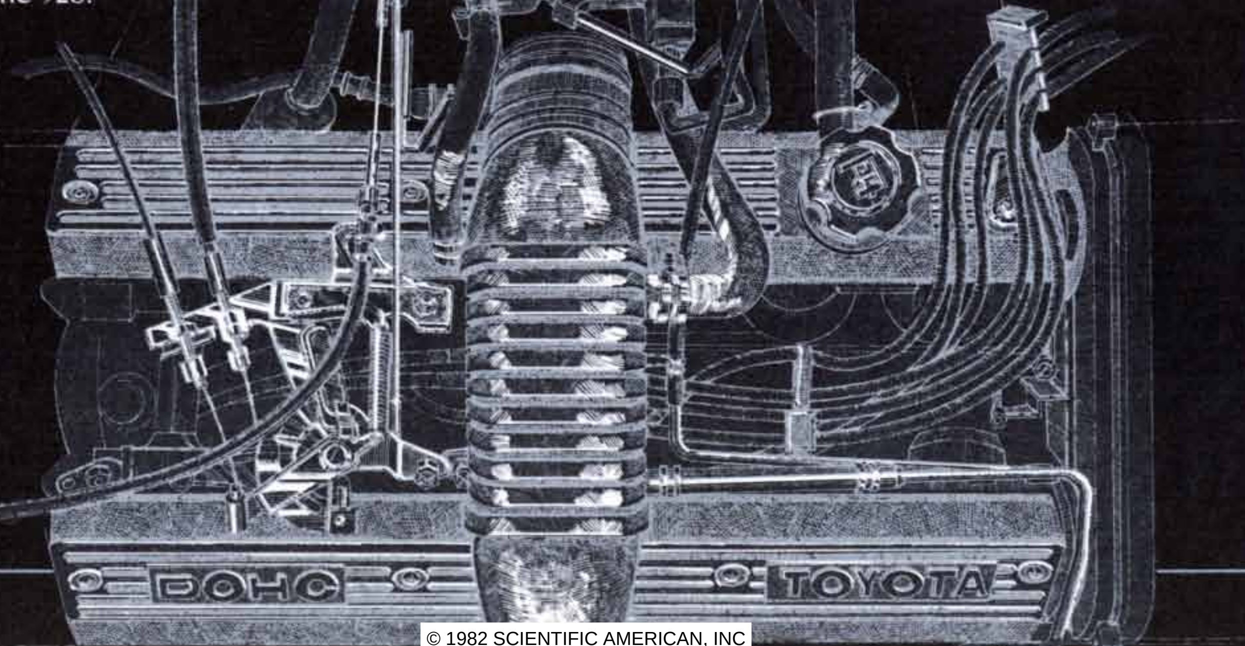
The right shape. An all-new aerodynamic wedge design so clean it actually has a lower coefficient of drag than a Porsche 928!

The right engine. Open the hood and those twin aluminum valve covers tell the world this is no ordinary power plant — this is a high-revving Twin-Cam Six! And it's electronically fuel injected for optimum response.

The right interior. A driver's sport seat with 8 different adjustments — even pneumatically-regulated lumbar support! Plus full instrumentation, and an electronic AM/FM/MPX stereo receiver with 5 speakers. All standard.

The right handling. 4-wheel independent suspension, with power-assisted 4-wheel ventilated disc brakes. Variable assist power rack-and-pinion steering. 5-speed overdrive transmission. And more.

The 1982 Celica Supra. As Motor Trend summed up, "By our standards, calling this car the right stuff...is an understatement!"



but there are a lot of these kinds of people in Sony."

Dr. Takeo Yamamoto, Managing Director of Pioneer Electronics, is similarly optimistic about the prospects for Japan. Speaking to the subject of industrial robots, he states, "The export of robots is often called the export of unemployment. But this depends on how robots are used. It would be best if they are used where human error is frequent or where human accuracy is insufficient. In any event, I think Japan will keep its lead in robots, because we are backed up with superior electronics."

Many Japanese executives see the general field of combined electronics and mechanics, "mechatronics," as especially suited to Japan's needs and strengths. Mr. Sakae Shimizu, Director and General Manager of Technical Planning and Coordination for Toshiba, believes that "Japan can hold its lead in production technology and mechatronics for the next ten years. In terms of product output, the energy problem is our Achilles' heel. We have only people as a resource. Since we are a country that sells human intelligence to buy food and energy, the development of mechatronic technology suits our needs. Further, I think that the Japanese people are given to the pursuit of precision. We get orders both from foreign and Japanese customers, but the Japanese customers are much more demanding. Thus, the reason for Japanese leadership in this field is both in terms of needs and in terms of special characteristics of the Japanese people.

"Further, the introduction of robots does away with dirty work, and in the case of Japan there is no resistance to this. Abroad, dirty work is done by immigrants in many cases, but there then is the problem of how to deal with the immigrants. Also in the Japanese case, in terms of increasing productivity, the unions are enterprise unions and there is no commitment to not changing jobs. In foreign countries, skill unions are frequent and there is a good deal of resistance to the introduction of robots. In Japan, there is, in the final reckoning, a high awareness of the advantages and disadvantages to the company, and therefore there is acceptance of the need for increased productivity."

Dr. Motokazu Uchida, Director, Research and Development Center, Hitachi, also believes that new technologies can more readily be introduced into Japan than elsewhere. "In the case of Japan, the Japanese have a special feature of liking new things. In order to survive in a competitive society, old things are discarded and new technology is moved to rapidly. The reason for Japanese superiority is perhaps due to this. Japanese consumers also have this special characteristic.

"Machines will become like people. People take information from their senses and give instructions from the brain to move their hands and feet. Machines also will pick up conditions from sensors and with information control from electronics will move according to instructions. Currently the most advanced aspect is the brain of microelectronics that manages information. The developments needed now are the sensors and activators.

"Currently a subject of much discussion in Europe is the impact of microelectronics on employment. This is not much discussed in Japan. The Japanese are flexible and have a high level of information. They feel that if some useless work is taken away, other work will be there. That is a special characteristic of the Japanese people, I feel. They don't worry that if a machine does certain tasks, all of the work will be taken away. The objective for introducing robots is to remove monotonous work that people dislike and hazardous work. There are simple tasks that machines can do and there is work that is not so simple that can only be done by people."

The general judgment that the competition for progress in technological development in electronics and related areas is between Japan and the United States is firmly held by Mr. Yoshihiro Yasui, Managing Director of Brother Industries. "Europe lags very much in a sense of computers. Their electronics industry itself is rather small. In the case of West Germany, importance is still placed on skill and the meister system still remains. Therefore they do certainly have an accumulation of technology, but further progress now is slow. In terms of advances in electronics, it is Japan and the United States. Getting a personal computer and playing with it wholeheartedly while still in grammar school, and then learning BASIC while playing with the keyboard—that is a reflection of this to be seen both in the United States and Japan."

The exceptionally high productivity of the Japanese steel industry is well recognized, and it is clear that this has come from considerable automation. Dr. Toshio Ikeshima, former Executive Vice-President of Sumitomo Metal Industries, analyzes the comparative positions of Japan and the United States regarding automation in this industry. "The biggest difference between Japan and the United States in our industry is in the operation of equipment by computer. We probably need not worry about being overtaken by the United States, whatever is done there. We at Sumitomo Metal Industries have many requests to provide computer control software. However, software applications to different manufacturing processes take a great deal of time. Further, because there are many computer makers, differences in equipment make a difference as well. Software workers need to be company employees, I believe. In the United States, their numbers are few and their compensation is high. It is unreasonable to expect the U.S. industry to hire large numbers soon.

"If one is investing in a new factory, it is rather easy and efficient to automate. But if the equipment is old and the factory is to be renovated, it is very time consuming and not very efficient to introduce computer controls. This is the usual problem of old equipment, not limited to the United States. Further, in America office functions are rather automated, but because they do not look on factory work as high class, software specialists in America are not much inclined to get into factory-related work."

Some executives see problems in Japan's efforts to build a leadership position in technology. Mr. Tsunesaburo Watanabe, Senior Managing Director, Olympus Optical Company, is proud of his company's new laser disk, announced for market introduction shortly. He feels his company has in hand an advanced memory device for computer use that is unique. However, he states, "Japan may hold its lead in technology but we must fill out our basic research more. Leaving out the biggest companies, corporate research is devoted only to product development. More national budget must be spent on national research centers. These should not be subject to governmental instructions, but, I think, should be focused on free research. Again, perhaps by joint research between universities and companies, there might be more than simply the expression of narrow interests."

Mr. Minoru Suzuki, Managing Director of Asahi Optical Company, also has concerns about Japan's future position and feels attention needs to be directed to the educational system. "I think it will be difficult for Japan to hold its technological position. Our present situation is nothing more than the result of working desperately hard and shows the difference in the quality of workers. From a technological point of view, while we certainly have some products that are superior, the United States is well ahead overall in leading technologies. Therefore, if they move to full automation with robots, we will face a difficult situation.

"In order to survive, it will be necessary both to move to automation and unmanned plants and to raise and extend the level of education of the personnel that are freed up. Such measures as going to a six-year from a four-year university training are needed. Specialist courses at the university are generally only two years, and that is too short. It is difficult to train scientists

in a company, since they need an atmosphere of freedom, and there is no room for that where everyone is working extremely hard."

Dr. Shunkichi Kisaka, Executive Vice-President, Matsushita Electric, has broader concerns about Japan's prospects. "In thinking about Japan's situation, a first point is that Japan is a country without religion. A second point is the national characteristic that, if a given direction is defined, everyone rushes in that direction together. For example, when it was understood that the thing to aim at was miniaturization, this became the most desirable goal for the Japanese. Given a clear direction, the sequence is investment–rationalization–cost reduction–quality improvement. But in life sciences, aerospace and oceanic development, there was no such group agreement. We can give up in those areas. We are ten to fifteen years behind and cannot catch up.

"On the other hand, when we start running in a direction, there is no brake. Perhaps we do not have the brake of religion. If you are religious, you believe that if while living you have lived to excess then when you die you must apologize as you go to the place of the gods. We do not have that; we are not that kind of country.

"Speaking of machines and electronics, the machines are working in place of people. We must rethink the role of people because of the problem of possible alienation. In the coming era, how will people use their time? Will they spend it in reading? Will they spend it in more imaginative work? Will they spend it in artistic activities? I do not know, but whether we like it or not, these are problems that will be arising."

While there is general agreement that robots will be increasingly used, Mr. Sho Masujima, Managing Director, TDK Electronics, introduces a note of caution. "I have the feeling that automation may be taken to extremes. When building a robot, do you make the robot fully electronic? Or do you include a mechanical controller in the robot? I feel that electronics are now being used to an unreasonable extent. It is necessary to keep a balance between the strong points of mechanics and the strong points of electronics.

"Japanese design is rather prejudiced toward electronics. In our case, while we at one point went entirely to robots for the automatic assembly of cassettes, we have now gone back a bit. We tried for complete automation through electronics, but found we had a good deal of loss, which caused us to return to some more mechanical systems."

The pioneering firm in robotics in Japan has been Kawasaki Heavy Industries. Mr. Gensuke Okada, Managing Director of that company, also believes that robots need modification. "Rather than a single function machine, as work requirements change there is need for a 'teach and repeat' capability. If taught, the machine will remember and repeat the function. This is of course a special feature of human workers. I believe the potential advantage of the robot is that it can do things in the same way as humans, but moreover does not tire and is not subject to failure. However, because we began with a single function machine, there has been too much holding to that concept, and we have not gotten rid of the notion. We need to move to attaching various sensors to industrial robots, although in truth, the kinds of functions necessary are extremely difficult to define."

Not all companies have moved to robots rapidly. The changes in technology require new ways of thinking and new skills, and changes are not always economical, in the view of Mr. Reijiro Ando, Director of Daini Seikosha. "At first, our product was not an electronic watch, but rather an electric watch. There have been problems in making the transition arising from the differences in point of view of the mechanical people and the electronic people. The first phase of effort is to achieve mutual understanding. The second phase is bringing the two groups together. However, the third stage is the most difficult. Electronic backgrounds are fine, and mechanical backgrounds are fine, but there are very few people who represent a combination of the two. It is no good to have gaps at the boundaries between the two approaches. From now on, we must have technicians combining these abilities. That is the third phase.

"In our own case, we are not yet using welding or painting robots. Even where we do not have complex operations, fixed machines are still in wide use. We do now have unmanned parts manufacture, but we have not yet gotten to unmanned assembly. Technologically we can use robots, but it would still be uneconomical."

However, despite some concerns over Japan's technological future and problems with moving rapidly into the application of electronics to production, the general view of senior Japanese executives about Japan's position and prospects is a positive one.

Mr. Hideo Sugiura, Executive Vice-President of Honda Motor Company, states, "There is a difference in the way in which Japanese and Westerners display creativity. That is, the base on which the countries depend is different. The Japanese focus on practical products, and thus tend not to show creativity in basic research. However, I do not think this is from lack of creativity. Of course the number of Japanese Nobelists in science is few, but that is because the numbers pursuing basic science are few. It might be said that the soil is thin.

"Japan is in the situation where in order to eat, we must sell things to other countries. Since 1945, we have been raised to think in that fashion. Previously we were a closed country. While we possess a unique culture, there was no exchange of information with the world. The West has a history of scientific research since the Renaissance, and it is this difference in historical climate—in addition to the thought that there is a thin layer of basic science—that gives an impression of lack of creativity.

"However, in terms of consumers and their needs, we are rather superior to other countries. America has a big American market, and it is enough to stay within it. Europe is able to limit itself to Europe. That is the difference in situation compared to us. The conception that you must adapt to suit others is foreign to Westerners. For us, if we are trying to get something sold, however minor, we are thinking of others and their life styles and tastes."

Similarly, Mr. Susumu Aizawa, Senior Managing Director, Epson Shimshu Seiki Co., Ltd., believes circumstances have limited Japanese scientific effort until recently. "It is said that to be successful one must be the first to grasp customers needs, but I believe one must imagine needs even more in advance. It is said that Japan is an imitative society that has taken the technology of advanced nations and prospered industrially without using imagination or creativity. I have a different view. Until now, Japanese engineers could only proceed in that way, introducing foreign technology and working full out to make it a reality. However, that does not mean that they are unable to develop new technology. Under Japan's conditions, efforts at a new technological revolution would have been absurd. But from now to the future, we have no other road but to move from leading edge to leading edge in technology. Seen in this way, the future strength of Japanese engineers will not be in technology improvement but will be directed to introducing entirely new technologies."

General optimism about the ability to introduce new technologies is represented by the remarks of Mr. Sachio Hata, Manager of the Electrical and Electronics Division of Toyo Kogyo. "We are finally moving toward unmanned production. First in parts manufacture, and then in assembly. Where volumes are high, we have already gotten near this target, and we are making the changes by using robots.

"Basically I think it can be done. There are still a number of problems that need to be solved. However, the issue of a conflict between automation and human happiness does not seem to trouble the Japanese. The Japanese view seems to be that the automation of production will free human beings and will greatly increase human well-being. Perhaps after all the Japanese are rather innocent. They seem to feel that it is a good thing to have the machine doing the work while being able to make the same money. Perhaps it is due to the high level of education, but they have responded cleverly to the changes made so far and have adaptability regarding changes made in the workplace."

Office automation is still an emerging field in Japan. Dr. Fusao Mori, Managing Director, Mitsubishi Electric, comments that "Office management is a matter of society and culture. In the case of the United States, office areas are divided very carefully. 'You are from here to there.' In Japan the area is vaguely defined and one cannot really say how far one's space extends. In strict logic, the U.S. way is more advanced. Similarly, English is easy to handle and fits a word processor very well. However, Japanese is difficult, and the introduction of a word processor was expected to be long delayed. In the event, it developed unexpectedly fast.

"A reason for the rather fast development of the Japanese word processor is that it is not yet at the level of sophistication hoped for ideally. The machine cannot make a judgment from context as to which character is to be printed. The problem is like that of using a dictionary while typing Japanese. Despite this, the machines are selling well. Electronic engineers are high-handed, aren't they? Word processors have only part of the functions they are likely to have someday. Still, there is a market that says that the overall value is sufficient. A number of electronic products are like that. For example, the computer is still in its first stages. If people could speak to the computer, in ordinary language, its power would be greatly increased.

"We can expect rapid application of these technologies in Japan. There are differences in social structure that affect this. The Europeans and Americans feel that if welding robots are used, there will be unemployment in the welding industry. However, in Japan, there is little anxiety because it is known that the worker will be moved to something else. The Westerners have individualism; the Japanese have companyism."

Sanyo Electric has a considerable commitment to office automation and to the information industry. Dr. Masaru Yamano, Executive Managing Director of the company, takes a world view of possible developments. "Industry receives the initial benefits from an information society. However, information is not simply a national matter. It is an international concern. Even now, there are systems by which computer calculations can be done from anywhere in the world using phone lines. Therefore, the development of the information society is not just a matter of Japan. We must develop a language for speaking throughout the world. This technology must not be a monopoly but must be open to all. I think it will be necessary to build systems that take account of the society and customs of others. Of course, it must be supplied economically, but I think the Japanese can do that.

"In the case of writing, the script will be Roman letters, but in the next generation resistance to English will be overcome and we can have a language that is a mixture of English and Japanese. The information is not our only major weapon, but I think it can work well for the prosperity of the next generation. Japan does not have raw materials. Therefore, industries that are based on raw material are of no use to us. In that sense, the information industry is the direction in which Japan must look."

Specific applications of electronics are being examined by major Japanese companies. Mr. Shoichi Ninomiya, Managing Director of Fujitsu, sees increased computer controls widely applied. "Our developments in automation have come out of our own needs in the production of semiconductors and electronic devices. This is true not only of Fanuc in numeric controls and robots but also in our new company called Fujitsu Automation. Automation is an absolute necessity to bring costs down and allow variation in product.

"There are of course a great many future applications of mechatronics. Take the example of vehicles. It is a nuisance to drive such vehicles as forklifts, bulldozers, and power shovels. These are machines that require various kinds of manipulations. If electronics were introduced, it would become possible to operate them very simply. However, it could not be done with a simple program. There would be a need to change responses according to specific conditions. Thus it will be important to have the judgment of an on-time sensor.

"It would be necessary to think both of an information-providing sensor and of computerized controls. The computer would need to be small and cheap, rather like a personal computer."

In optics applications of electronics, Mr. Shoichiro Yoshida, General Manager, Nippon Kogaku, refers to production technology. "The special feature of our company is optics. Until now optics and mechanics have been combined in optical equipment—cameras, telescopes, binoculars and the like. Optics and mechanics have fit well together. Now, however, with autofocus and other developments, electronics has entered the field, forming a trinity. Our present products might be called optical electronic machinery.

"In terms of future products, equipment for the production of VLSIs will require extremely fine etching of lines on the silicon base, and there will be a need for control technology for this process. This is a combination of electronics with the corresponding mechanics, and since light will be used, of course optics technology will be added."

Automobile producers, the first to apply new technologies to production, are also speculating about electronics applications in their products. Mr. Masatoshi Morita, Senior Managing Director, Toyota Motor Corporation, says, "The adaptability of electronics to the automobile will widen. I would guess that electronics parts and equipment will account for 10 to 15 percent of the cost. These applications are on a rising slope now, and where they will peak is not known. Looking at total costs, many speculate that it will become 15 percent of the total automobile cost. At present, depending on the model, it is at most eight percent and usually five to six percent.

"Perhaps now is the time for thinking of robots and their application to autos. We are seeing an aging society and under such conditions it is important to make driving safe and easy. However, I do not expect really major changes in the next 10 years between humans and the automobile. I think rather that it will be a matter of considering how much of the human element to include in the automobile."

Mitsubishi Motors is also planning extensive use of electronics in their products. Mr. Shinji Seki, Managing Director, notes that "There is no limit to the possible use of electronics in the auto. There is a link in automatic transmission controls. There is an advantage in introducing electronics to prevent skids and to improve fuel consumption.

"Again, there is the possible use of microprocessors to control emissions. The automobile telephone, as it becomes more widely used, will provide a simple means to provide information on road conditions after indicating destination. Voice recognition equipment when it is developed will allow hands-free driving. It should be possible to go that far."

Still, for all of their enthusiasm for the new technologies and product applications, Japan's business executives remain focused on the consumer and the consumer's needs and wishes. The Victor Company of Japan is a leader in moving Japan to a virtual world monopoly in the production of videotape recorders. Victor has been a pioneer in what may well be the next generation product, video disks. However, Mr. Toshiya Inoue, Senior Managing Director, notes his care in timing of market entry. "From a technical point of view there are no particular problems now with the video disk. However, the market introduction has been postponed. If we sent them to shops under present conditions, they would draw a large crowd of cus-

JAPANESE TECHNOLOGY TODAY

tomers. But whether people would buy or not is the question. For that reason, therefore, we are waiting a while. Even with humans, one wants a child born in a good environment, and might want to wait a while before the child is born. The product's level of development and reliability is high; it is easy to handle; the price is suitable. There is now the issue of timing."

Japan has moved very rapidly to technological parity with other nations in a wide range of product areas. From a strong base in microelectronics, applications of this technology to manufacturing in particular has now progressed to the point where Japanese businessmen can seriously discuss a new industrial revolution, a revolution to be led by Japan and powered by Japanese technology. Japan's relative weaknesses in basic research are clear, not least to the Japanese, and are likely to be overcome. The prospect is for Japan to become a major source of a whole range of discoveries and inventions for the world over the coming decades.

The West has been surprised by this development and seeks often to belittle or demean the achievement. The late Lord Snow, in his posthumously published book on modern physics, as an aside to his main arguments, spoke most effectively to these Western attitudes.

"...Japan is worth particular notice. The Japanese scientists, technologists, technocrats, have shown skills and originality in all this electronic apotheosis which quite out-class the West's. That ought to surprise no one who has given the most perfunctory attention to Japanese visual art or literature or pure science. For hundreds of years, the culture has been wildly original, something oddly different from any other among the sons of men. It was an instance of Western blindness not to discover that simple fact."[4]

[4]C. P. Snow. *The Physicists*. Macmillan London Limited. London, 1981, pages 159–160.

The publishers, authors and Ted Bates Advertising/New York acknowledge the important contributions by the members of the international business community whose messages appear on these pages:

American Honda Motor Co., Inc.

Canon USA, Inc.

Japan Air Lines Co., Ltd.

Matsushita Electric Industrial Co., Ltd.

Mazda Motors of America, Inc.

Maxell Corporation

Nissan Motors Corporation in U.S.A.

Olympus Camera Corporation

Pioneer Electronic Corp.

Seiko Corporation of America

Sony Corporation of America

Sumitomo Metal Industries, Ltd.

Toshiba America, Inc.

Toyota Motor Sales U.S.A., Inc.

US JVC Corp.

In addition, we acknowledge the significant contributions to this section by the following distinguished business leaders who were interviewed:

Mr. S. Aizawa
Epson Corporation

Mr. R. Ando
Daini Seikosha

Mr. S. Hata
Toyo Kogyo

Dr. T. Ikeshima
Sumitomo Metal Industries

Mr. T. Inoue
Victor Co. of Japan

Mr. N. Kihara
Sony Corporation

Dr. S. Kisaka
Matsushita Electric

Mr. S. Masujima
TDK Electronics

Dr. F. Mori
Mitsubishi Electric

Mr. M. Morita
Toyota Motor Corporation

Mr. S. Ninomiya
Fujitsu

Mr. G. Okada
Kawasaki Heavy Industries

Mr. S. Seki
Mitsubishi Motors

Mr. S. Shimizu
Toshiba Corporation

Mr. H. Sugiura
Honda Motor Co.

Mr. M. Suzuki
Asahi Optical Co.

Dr. M. Uchida
Hitachi

Mr. T. Watanabe
Olympus Optical Co.

Dr. T. Yamamoto
Pioneer Electronic Corporation

Dr. M. Yamano
Sanyo Electric

Mr. Y. Yasui
Brother Industries

Mr. S. Yoshida
Nippon Kogaku

Special acknowledgment to the officers and staff of Japan Society for their contributions to "Japanese Technology Today." Founded in 1907, the Japan Society is the primary organization in the United States dedicated to enriching and strengthening the ties between Japan and the U.S. by providing a binational forum for the exchange of views and dialogue. For more information about Japan Society write to 333 East 47 Street, New York, N.Y. 10017 or call (212) 832-1155.

1982 Mazda 626. Just one look and you'll see its exceptional value versus BMW 320i.

There are many similarities between the Mazda 626 Sport Coupe and the BMW 320i. But as you begin to study these two exceptional road cars, one major difference stands out.

The 626 stands for exceptional value.

It's also an exceptional automobile. The 626 handles like a sports car. Rides like a full-size car. Looks like an expensive European road car. And gets gas mileage like an economy car.

37 EST. hwy. mpg **27*** EST. mpg

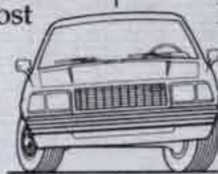
The 626 is an engaging mixture of sophistication and sport. A car that is quick and precise through turns. Yet roomy and comfortable through long trips.

Perhaps best of all, the 626 offers you all this at an astonishingly low price.

Proof that among the world's fine road cars, the Mazda 626 Sport Coupe not only stands out. But stands *alone* as a remarkable value.

A ROAD CAR THAT KNOWS HOW TO HANDLE THE ROAD.

The 626 owes its superb handling characteristics, in part, to its almost perfect balance. 51% of its weight rests on the front wheels, 49% on the rear. In addition to this nearly ideal weight distribution, the 626's center of gravity is just 20.3 inches off the ground.



The cornering ability of the 626 is particularly impressive. For example, skidpad tests conducted by *Road & Track* magazine showed that the Mazda 626 scored 0.760g for lateral acceleration. In two years of testing, only one road car in its class, the Audi Coupe, scored higher.

ADVANCED AERODYNAMICS AID PERFORMANCE.

Aerodynamic efficiency contributes to a car's performance. So the Mazda 626 was designed to reduce air turbulence, from the slant of its grille to the shape of its rear window. No wonder the 626 achieves a low drag coefficient of 0.39. (The Mazda RX-7 sports car attains 0.34.)

EXCEPTIONAL RESALE VALUE.

Like other world-class road cars, the Mazda 626 has exceptionally high resale value. According to the Kelley Blue Book, Nov.-Dec. 1981, the 1979 Mazda 626 Sport Coupe has retained a full 99.0% of its original suggested retail price.

STANDARD FEATURES ON THE MAZDA 626.

- 2.0 litre overhead cam 4-cylinder engine
- 5-speed overdrive transmission (3-speed automatic optional)
- Steel-belted radial tires
- Power-assisted front disc brakes
- Full-width velour upholstery
- AM/FM stereo radio
- 60/40 split fold-down rear seat backs
- Remote trunk release
- Front and rear stabilizer bars

1982 Mazda 626 Sport Coupe $7545**

*1982 EPA estimates for comparison purposes for 626 Sport Coupe. The mileage you get may vary with trip length, speed and weather. Actual highway mileage will probably be less.

**Manufacturer's suggested retail price for 626 Sport Coupe. Actual prices set by dealers. Taxes, license, freight, options (wide tires/al. wheels $350-$400) and other dealer charges extra. Prices may change without notice.

Performance data and comparisons with vehicles based on '81 tests and information. For '82 they may vary, as may availability at dealers of vehicles with specific features.

mazda The more you look, the more you like.

Spektrum
DER WISSENSCHAFT
POUR LA
SCIENCE
édition française de
SCIENTIFIC
INVESTIGACION Y
CIENCIA
SCIENTIFIC
AMERICAN
サイエンス
SCIENTIFIC
AMERICAN
1
1981
日本版
LE SCIENZE
edizione italiana di
SCIENTIFIC
AMERICAN
科学
SCIENTIFIC
AMERICAN
SCIENTIFIC
AMERICAN
$2.00
March 1981
INTERNATIONAL

Artificial Intelligence

Computer programs not only play games but also process visual information, learn from experience and understand some natural language. The most challenging task is simulating common sense

by David L. Waltz

The idea that the digital computer will someday match or exceed the intellectual abilities of human beings has been put forward repeatedly ever since the computer was invented. In the early days of computing the assertion was without an empirical basis: the programs of the computer could no more be expected to reason, plan, learn, sense, formulate concepts, use language or think creatively than could their mechanical predecessors in the art of simulation: the puppet, the windup toy soldier and the music box. Moreover, the idea that intelligence itself might require a theoretical framework for its explanation was scarcely acknowledged. Intelligence was considered a transparent concept, and if it were to be recognized in a machine, the intellectual powers of the machine would have to be almost indistinguishable from those of a human being. Thus early computers were called electronic brains, and John von Neumann, one of the architects of the kind of computer in commonest use today, formulated explicit analogies between the computer and the brain. Many people still associate data with human knowledge, the operation of a program with decision making, the trace or ongoing record of the program's operation with the stream of consciousness and the acquisition of data with learning.

In the past 20 or 25 years the new discipline of artificial intelligence has put to rest some of the more naive analogies between the computer and the brain and has begun the task of placing the concept of intelligence on a theoretical footing. For the practitioners of the discipline the computer is a laboratory in which to develop new ways of thinking about thinking. Computer programs written by investigators in artificial intelligence have demonstrated conclusively that in certain activities (including activities most people would say require intelligence, such as playing games) the computer can outperform a human being. Recent programs have demonstrated that the computer can even develop elaborate theories about a limited domain such as arithmetic from a few simple axioms. At the same time the understanding of various features of human intelligence has been considerably enriched by the attempt to describe analogues of those features in the detail necessary for writing a program. As a result the analogy relating the performance of the computer to that of human intelligence has broadened and matured.

Consider learning: In people learning seems to be intimately related to growth and changes in the physical structure of the brain, but the hardware of a computer does not undergo any analogous changes. The programs of a computer, however, can change in a variety of ways: they can accumulate data, organize it and even modify themselves. Hence workers in artificial intelligence have suggested that complex programs, and particularly those that can alter their own operations, may be good models of human learning. Such models, like other models in science, do not duplicate the phenomenon they are meant to represent in all its detail. For example, the operations of a computer that embodies a "learning" program bear no explanatory relation to human learning as it is realized in the firing patterns of brain cells and nerve cells. Nevertheless, the computer model of intelligence is flexible enough to be formulated at whatever level of detail is considered appropriate for highlighting the essential functions of human thinking.

Here I shall attempt to convey an idea of the kinds of results that can be achieved by programs in artificial intelligence. Most of the programs I shall discuss are well-known approaches to major problem areas in the field, and my descriptions are intended primarily to make clear the main contributions of each program. I have therefore favored programs whose operation can be explained succinctly. What follows should be read more as an introduction to the field than as a summary of it.

One of the constant themes in artificial intelligence has been how best to explore a range of possible actions in the pursuit of well-defined goals. In general for every action taken new possible actions become available, and so the planning of a sequence of actions must consider a branching structure of possible states called a search tree. The tree is generally shown inverted, with the root at the top. The root is the current situation, the branches are the possible actions and the ends of the branches are the possible outcomes. One or more of the outcomes may correspond to the goal of the search.

Heuristic Search

Exploring the full range of possible actions in all possible sequences would be equivalent to tracing every branch of the tree and would guarantee that the optimum sequence would be identified. In many cases, however, the search tree is so large that a complete search is not feasible. Therefore in most artificial-intelligence programs heuristic principles, or informal rules of thumb, are incorporated so that the most promising actions are selected for further examination early in the search and the less promising ones are eliminated from full-scale consideration. Heuristic principles are effective and practical in most circumstances and they shorten the search, but their application cannot guarantee that the optimum result will be found. Heuristic-search methods are commonly employed in programs that enable a robot to plan the manipulation of objects or to plan its own movement in space.

Programs that play games such as chess, checkers and backgammon are informative examples of the application of the heuristically guided search. Many of the programs now play at the level of the expert or master human player; a program called Mighty Bee, written by Hans Berliner of Carnegie-Mellon University, defeated the world backgammon champion in 1979. Probably the most intense interest has been focused on programs that play chess, and virtually all chess programs devised since the 1950's are based on a heuristic-search model developed by Claude E. Shannon at Bell Laboratories.

When a chess program is choosing a move, the program must somehow evaluate the position that results from each candidate move. The evaluation can be done by means of a scoring system such as one based on the value of each piece and on the gravity of the threat to any piece that can be captured on the next move [*see illustration on page 121*]. Many programs also incorporate such factors as the strength of the pawns' positions, the degree of control over the center of the board and the number of pieces mobilized. The program plays to maximize the score.

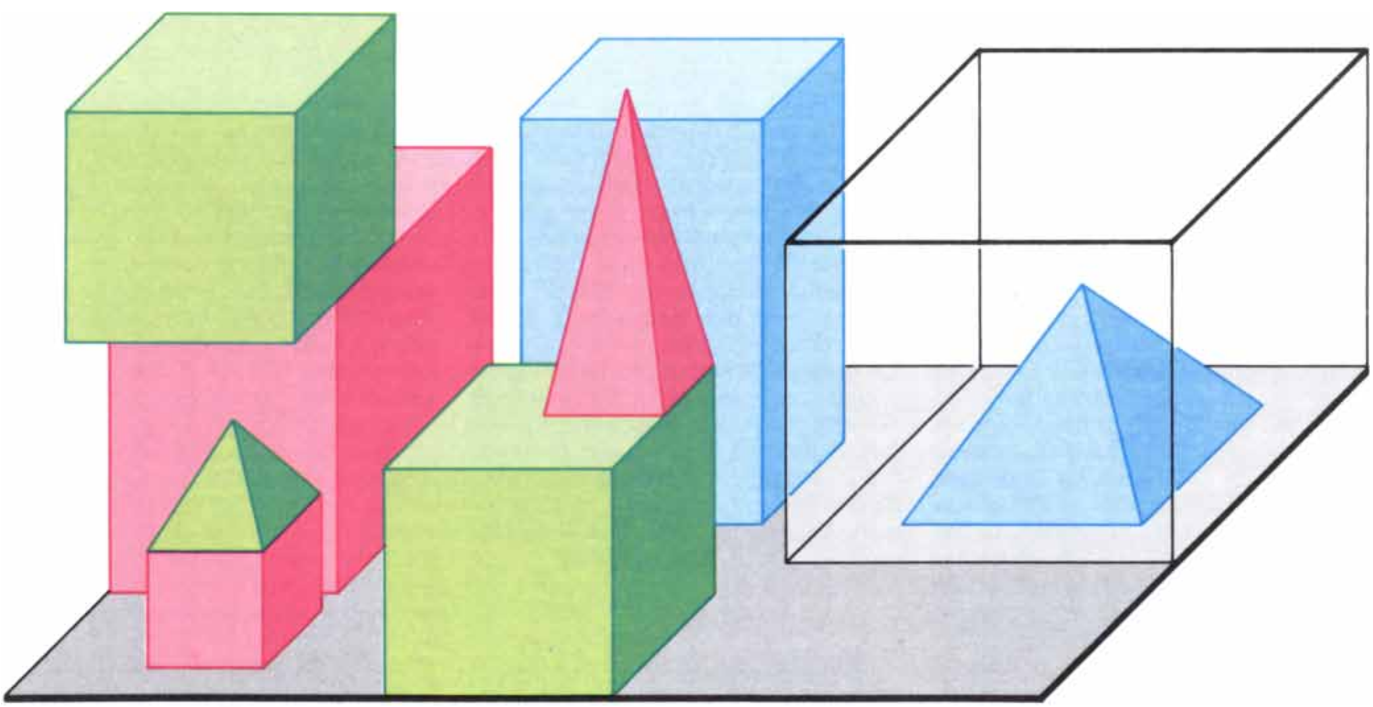
The heuristic principles incorporated into a chess program control the breadth and the depth of the search tree the program explores in looking for the best move. Generally the program calculates the score for every legal move from a given position. It can then consider the opponent's responses to each of these moves and its own range of countermoves for each response. Early in the process, however, some selection of moves must be made or the number of positions to be evaluated will become impractical. It is in making the selection that the heuristic principles are brought to bear. The program might select only the move with the highest score for further study, or it could explore the consequences of, say, the five most promising moves. For the moves still under consideration the program examines the search tree to some specified depth, at which point the heuristic principles determine whether or not to continue the exploration: a line of search is abandoned only if the position reached is reasonably stable. Black then makes the move likely to lead to the highest score, assuming that White makes the best move available at each turn.

In chess the average number of moves that can be made from a given position is 35; an exhaustive search only three moves deep for each player would require the examination of more than 1.8 billion moves. Rigorous methods for pruning the search tree have been devised, but in spite of such methods the essential operation of most computer-chess programs is an extensive exploration of the search tree. The current world champion among programs is Belle, developed by Ken Thompson and Joe Condon of Bell Laboratories. The program runs on a specialized computer with hardware designed specifically to make chess calculations; it examines an average of 160,000 positions per second. Belle plays in tournaments as if it had a chess rating of 2,160 (a rating of 2,000 to 2,199 qualifies a player as an expert), making its level of play higher than that of all but a small fraction of human chess players.

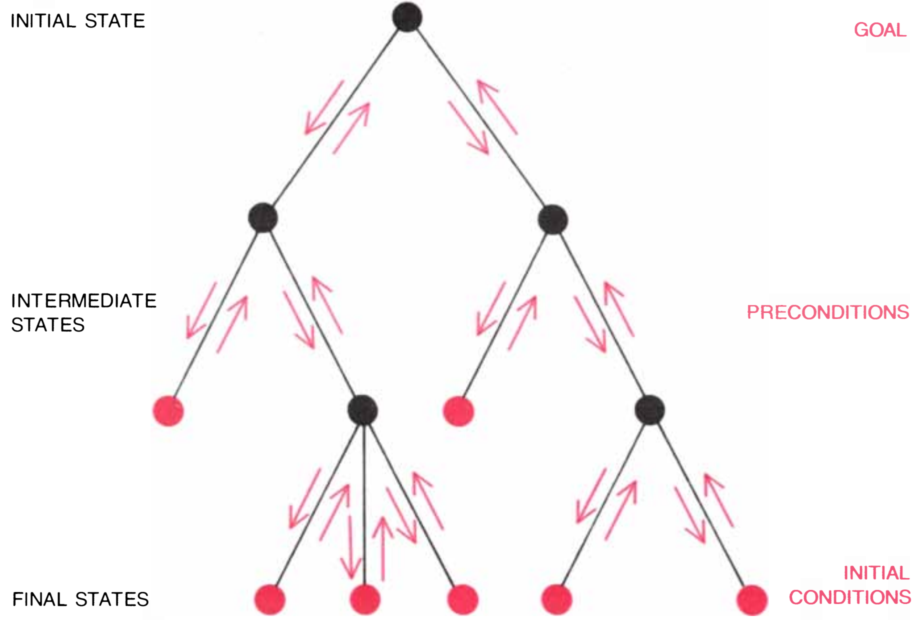
Artificial Intelligence and Psychology

The operation of Belle and similar chess-playing programs illustrates one of the most important questions about the goals of work in artificial intelligence. If one of the primary goals is the simulation of intelligent human behavior, on what level is the simulation to be carried out? In a sense a program that plays chess at the level of an expert chess player, by whatever means, simulates the behavior of the expert human player. On the other hand, if the goal is to emulate the strategic decisions made by a human player, a program such as Belle must be counted a failure.

There is considerable evidence that good human chess players employ processing methods radically different from the ones incorporated into Belle. A human player relies on strategy: he picks a goal, such as capturing a particular piece, then looks for ways to attain the goal. Reaching the goal may in turn call for identifying subgoals and verifying that each of them can be attained. Tests have shown that a good human player examines no more than about 100 board positions before choosing a move; only



"BLOCKS WORLD" is a highly simplified domain that enables workers in artificial intelligence to explore the ways perception, thought and action can be modeled and linked by a computer. The blocks world of the computer is represented by the colored cubes, boxes and pyramids in the illustration. Also part of the blocks world are the standard principles of everyday physical interactions; for example, a pyramid in the blocks world is not allowed to rest balanced on its vertex. The blocks world was devised at the Massachusetts Institute of Technology by Terry A. Winograd, who is now at Stanford University, as a domain of action for a computer program called SHRDLU. SHRDLU simulates the actions and responses of a person with respect to the objects in the world: the program can plan and carry out certain manipulations of the objects according to instructions issued to the computer in ordinary natural language, and it can answer questions about present and past configurations of the world, about actions it has taken or plans to take and about the reasons for its actions. Other investigators have shown how to model the learning of new actions in the blocks world. Scenes from the blocks world have been adapted from Winograd's article "Understanding Natural Language," which first appeared in the journal *Cognitive Psychology*.



SYSTEMATIC SEARCH of an inverted tree structure proceeds from the top node, or root, along the leftmost branch to a terminal node. The search then returns to the first intermediate node from which it can trace another downward branch and goes on to explore all the terminal nodes from left to right (*colored arrows*). An inverted tree can represent the structure of a variety of problems. The root can represent a given situation, such as a position in a chess game, and each node directly below the root can represent the outcome of some action that is possible in the situation, such as the movement of one of the chess pieces. Still lower nodes can represent the results of possible subsequent actions, such as the countermoves available to a chess opponent after a given move. The root of a tree can also represent a goal, and the immediate branching nodes can represent the preconditions of taking some action that will achieve the goal. A precondition can have preconditions of its own, which are represented by the nodes at the next level. Exploring a tree structure is a serial process, that is, the nodes are examined one at a time.

the most promising lines of attack are considered, but they are explored in relatively great depth. Belle, on the other hand, examines some 29 million positions in the average of three minutes it can allot to each move in a tournament game.

Given Belle's success at the chess board it seems clear that intelligent behavior can result from underlying processes quite different from those of human cognition. Some investigators, however, argue that artificial intelligence is, broadly speaking, a branch of human psychology. They submit that whenever a program simulates some aspect of human intellectual performance, the program constitutes an "existence proof" for the computer model of human intelligence. The operation of the program may diverge so far from human mental processes as to preclude any direct contribution to human psychology. Nevertheless, the program may simulate certain other aspects of human intelligence that seem quite important, such as its unpredictability in performance. Chess programs often play better than the people who write them, and so it is highly misleading to assert, as many people have, that computer intelligence is limited because a computer can do only what it is programmed to do. In many cases the programmer does not know what his program can do until it is run on a computer.

A striking example of the phenomenon was reported recently by Donald Michie and Tim Niblett of the University of Edinburgh and Ivan Bratko of the Edward Kardelj University of Ljubljana. A program developed by the three investigators analyzes certain end games in chess, a part of the game in which most computer programs fare considerably less well than they do in the middle of the game. Master chess players generally memorize a set of rules for the end game, but the program has demonstrated that certain end-game positions once thought to end in a draw actually result in a forced win for one side. In order to force the win, however, one must apply a set of rules too complex to be easily remembered by a human player; the rules are well within the capacity of current computers.

Planning

Although goal-directed planning is not a dominant feature of chess-playing programs, it is easy in principle to write a program that identifies goals and finds ways to attain them. Again the operation of the program can be understood as the exploration of an inverted treelike structure.

The root of the tree represents the goal and the nodes that branch away from the root represent preconditions that may have to be met before the goal can be achieved. (Preconditions that, taken together, are sufficient for achieving the goal are linked, but all the sufficient sets of preconditions are represented in the tree.) Each precondition can in turn become a subgoal with preconditions of its own represented by further branching nodes. The terminal nodes of the tree represent conditions already satisfied by the initial state of affairs. An action that can accomplish a goal or a subgoal is also associated with each nonterminal node. The program explores all the branches of the tree until every precondition is satisfied. It then constructs a plan of action by tracing a path through the tree from the lowest node to the highest one.

A program developed in the late 1960's and early 1970's by Terry A. Winograd, then at the Massachusetts Institute of Technology, illustrates the operation of a planning program in a carefully limited domain. Winograd's program is called SHRDLU, after the seventh through the 12th most frequent letters in English; the domain in which it operates is made up of simple three-dimensional objects such as cubes and pyramids that rest on a table. (The program actually has no facilities for manipulating real objects or even for representing them on a video screen, but it keeps track of their properties and positions internally.) SHRDLU replies to questions typed by an operator, executes commands on the objects in its simple environment, called the blocks world, and reports on the results.

To understand how a plan of action is formulated by SHRDLU, suppose the program could carry out only two actions, called STACK and CLEARTOP. The program executes STACK by grasping an object, moving it to a point above a second object and putting it down on top of the second object. CLEARTOP effects the removal of one object from the top of another; it calls for grasping the upper object, moving it over an unoccupied region of the table and putting it down.

Actually both STACK and CLEARTOP represent a hierarchy of actions. At the next level of detail are subactions such as GRASP, MOVE and PUT. Each of the subactions in turn is made up of a sequence of more detailed steps, until ultimately the steps could be expressed as control signals for such activities as closing mechanical "fingers," sensing contact with an object and rotating a joint. Thus the hierarchy of program statements reflects the hierarchical structure of human actions, from high-level, conscious goals to the detailed sequences of motor control. Hereafter I shall omit all such details of control and discuss the plans formulated by the program only at the highest level.

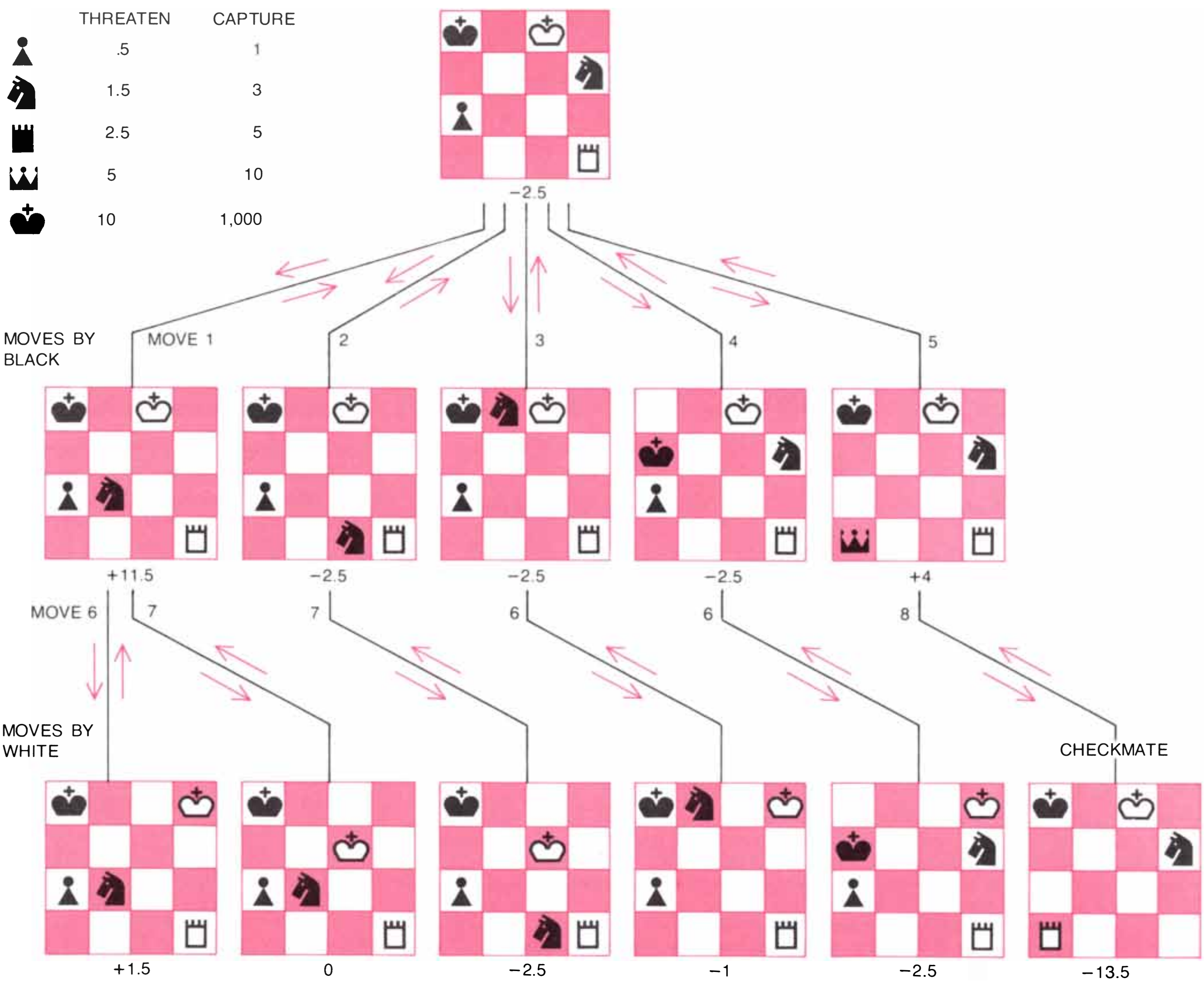
Before the action STACK can be executed two preconditions must be satisfied: the tops of both the objects that will form the stack must be clear. Similarly, there are two preconditions for the action CLEARTOP. The object whose top is to be cleared must have another object resting on it, and the top of the upper object must itself be clear. In order to select an action that can accomplish a certain goal or subgoal, the program compares the goal or subgoal with a set of "postconditions" associated with each action. The postconditions specify the configuration of the blocks world that the action will bring about. If the configuration satisfies the conditions of the goal or subgoal, the action is proposed as part of the plan.

Suppose a red cube is resting on a blue one. How can the program exploit the actions STACK and CLEARTOP to achieve the goal of placing the blue cube on the red one? Beginning with the goal state, the program searches for an action that has postconditions matching the goal and sets up the preconditions of the action as new subgoals. The action STACK could place the blue cube on the red one (the goal) if the tops of both the red and the blue cubes were clear. In the given initial state the top of the red cube is clear but the top of the blue one is not. Hence clearing the top of the blue cube becomes a new subgoal.

Backward Chaining

The program continues by searching for actions that will achieve each subgoal and sets up preconditions of each action in turn. To clear the top of the blue cube the action CLEARTOP could be applied. Its preconditions are satisfied by the initial state of the blocks world and so the search is complete. The entire process of constructing a sequence of actions from the goal state back to the state in which all the preconditions of the actions necessary to initiate the chain are satisfied is called backward chaining.

The program can now devise a plan of action to achieve the goal by reversing the sequence of actions found during the search. The plan is first to apply CLEARTOP in order to remove the red cube from the top of the blue one and second to apply STACK in order to place the blue cube on the red one. One effect of the plan is that the top of the blue cube is cleared, a change brought about by the plan that was not one of its explicit



TREE STRUCTURE of a simplified chess game shows how the computer program, here playing Black, determines its next move. The program examines the available moves and the opponent's countermoves one by one, tracing the game tree (*colored arrows*) to a depth determined by heuristic rules. Each position is assigned a score according to an evaluation procedure, such as the one in the table at the upper left. The initial position, shown at the top of the tree, is scored −2.5: a total of 1,004 for the black pieces on the board, −1,005 for the white pieces and −1.5 for the threat to the black knight. The move selected is the one that leads to the highest score for Black at the greatest depth examined, assuming that White will make the best countermove available. In this case Black plays move 1 and assumes White will respond with move 7. The heuristic rules instruct the program to abandon the search whenever one of White's countermoves leads to a better score for White than the best countermove previously considered. Thus the program abandons its examination of moves 2, 3, 4 and 5 as soon as it finds countermoves that lead to positions worth less than 0, the value of the best position available to White after move 1.

goals. Such a change is called a side effect of the plan.

One of the most intriguing recent applications of backward-chaining search methods has been the development of computer programs that draw on expert knowledge and interact with the user. The programs have proved capable of expert performance and are called expert systems. They have been applied to tasks such as medical diagnosis, prospecting for mineral deposits and advising on tax matters. For example, if the goal is the diagnosis of disease, the program works backward from general information about the patient's condition to increasingly specific facts until the chain of facts can justify a fairly plausible inference about what disease is present. Thus the emphasis of an expert system is to develop a plan for making a plausible inference, based on knowledge supplied by the user, rather than to develop a plan of action. The possible inferences are stored as rules in the program, where they have a role analogous to that of the actions in SHRDLU.

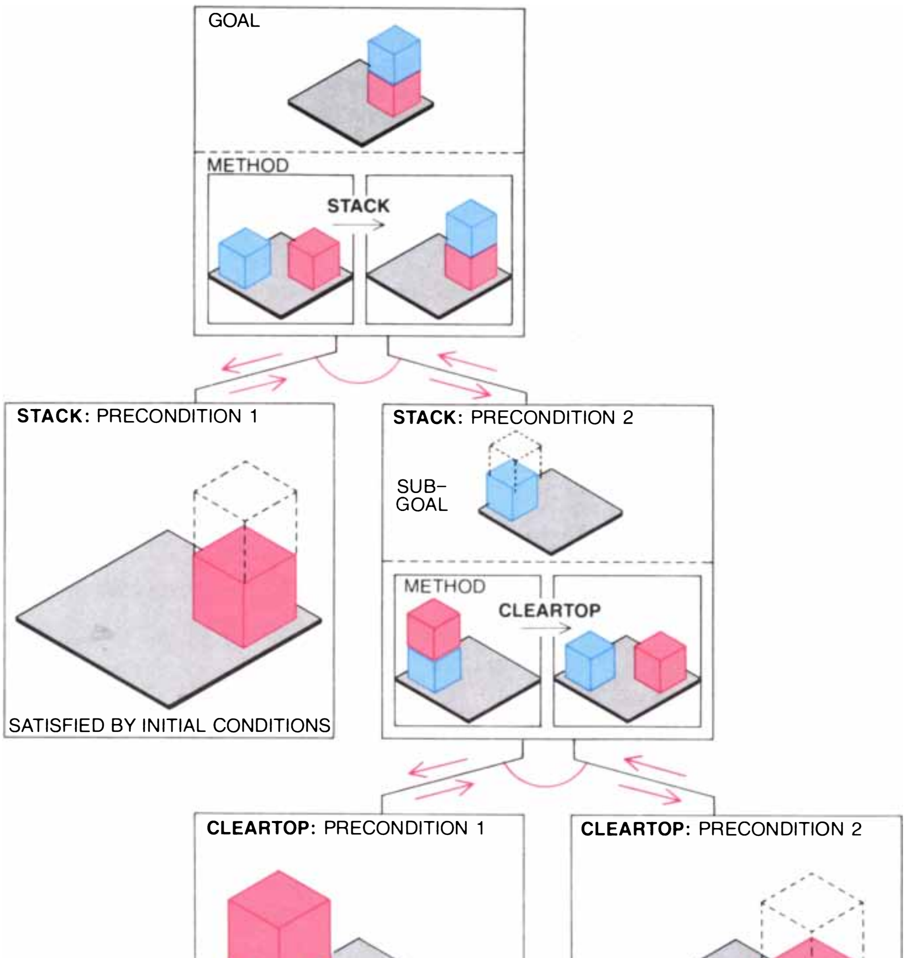
A good example of a rule is found in the medical-diagnosis system called MYCIN, written by Edward H. Shortliffe of Stanford University: "If the infection is primary bacteremia, and the site of the culture is one of the sterile sites, and the suspected portal of entry of the organism is the gastrointestinal tract, then there is suggestive evidence (.7) that the identity of the organism is bacteroides."

Such a rule is quite similar in form to the rules people cite to explain their decisions; rules employed by expert systems have therefore been relatively easy for experts to learn to write. It is important that a numerical confidence rating, such as the number .7 in the MYCIN rule for diagnosing bacteroides infection, be assigned to each rule. The confidence rating is the basis of heuristic controls that guide the program's search for a plausible inference, and it is supplied to the user for assessing the validity of various conclusions the program generates. The program must also explain why it reached the conclusion it did; no one would trust a program that simply printed its conclusion without saying why it rejected some other conclusion that seems reasonable.

Learning in the Blocks World

Unless a computer can expand its own capabilities on the basis of "experience," the performance of a program is limited by the knowledge, foresight and available time of the programmer. Learning, therefore, is a particularly important concern in artificial intelligence. A number of investigators have adopted Winograd's blocks world as a kind of laboratory in which to explore various learning strategies.

One kind of learning is based on the "memorization" of a plan formulated after a search by backward chaining; the idea was first developed by Richard E. Fikes and Nils J. Nilsson, both then at the Stanford Research Institute, for the system STRIPS (for the Stanford Research Institute Problem Solver). Suppose a program is to achieve the goal state in which a brown block is resting on a green one, and it encounters a state in which a green block is resting on a brown one. Having earlier interchanged the blue block and the red one, the program stores a representation of the plan it developed for the exchange, namely to employ first the action CLEARTOP and then the action STACK. The representation of the plan must be a generalized one, however, since otherwise it could be applied only to the blue and



FORMULATING A PLAN in the blocks world calls for identifying a goal and choosing a sequence of actions that will achieve the goal from a given initial state. In the example shown the red block is initially resting on top of the blue one, and the goal is to place the blue block on top of the red one. In order to devise a plan of action the program conducts a search that is equivalent to exploring a tree structure. Each node of the tree represents a state of the blocks world; the root is the goal, that is, the state in which a blue block rests on top of a red one. An action is selected from the repertory of actions available to the program, such as the action STACK. The action selected is the one whose results, or postconditions, match the essential features of the state; the action is also associated with the node. The nodes branching down from a given node represent preconditions, or states in the blocks world that must be attained before the action represented by the given node can be carried out. When multiple preconditions must all be satisfied, they are connected by an arc. Before the action STACK can be undertaken the top of both the red block and the blue block must be clear. The program determines that the first precondition is met by the initial state but the second precondition is not. To clear the top of the blue block, therefore, another action from the repertory must be applied, and its preconditions must in turn be compared with the initial conditions. When all the preconditions of every node that is explored by the program are satisfied by the initial conditions, the sequence of actions encountered during the search is reversed to make a plan that will accomplish the final goal.

the red blocks it first encountered. The generalization is accomplished by storing the names of the blocks to be interchanged as variable quantities. The entire action, here called INTERCHANGE, is added to the program's repertory of possible actions as a fixed sequence of the simpler actions GRASP, MOVE and PUT, together with new preconditions and postconditions for the action INTERCHANGE. Because the initial state of the brown and the green blocks satisfies the preconditions of the action INTERCHANGE the program can immediately achieve the goal state without a search.

A second kind of learning is trial-and-error learning, which is modeled in a program called HACKER written by Gerald J. Sussman of M.I.T. (In the jargon of computer science a hacker is someone who spends much of his time writing computer programs.) HACKER is made up of a planning system similar to the one in SHRDLU, "critics" that monitor the planning and note problems and "debuggers" that formulate new rules for the programs to avoid repeating the errors caught by the critics.

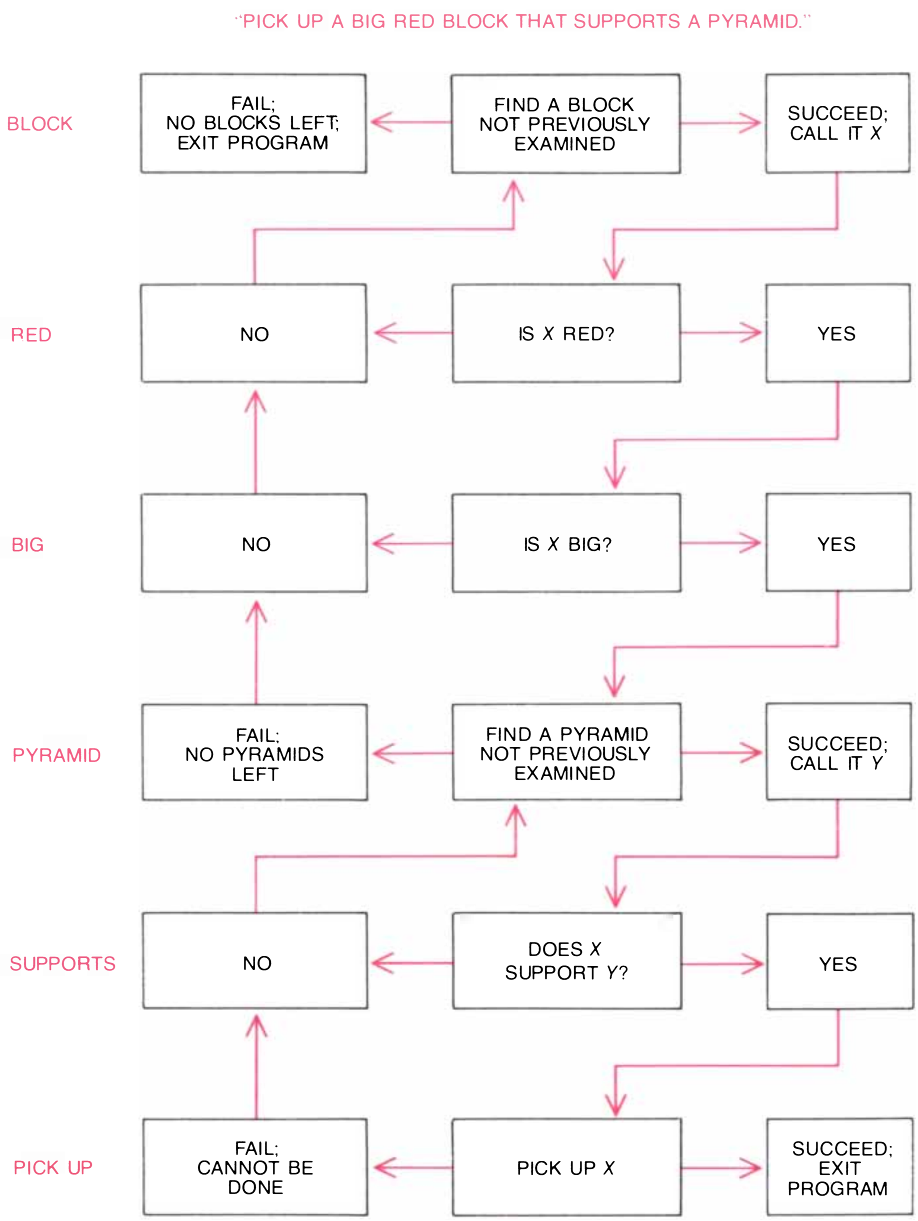
Suppose a red block is resting on a blue one and a green block is placed alone on the table. HACKER is asked to build a three-block stack with the red block on the top, the green one in the middle and the blue one on the bottom; the program has available only the actions STACK, CLEARTOP and INTERCHANGE. In order to plan its actions HACKER breaks down the goal into subgoals: for the first subgoal it can choose either to place the red block on the green one or to place the green block on the blue one. Assume that the program arbitrarily begins with the wrong subgoal, namely placing the red block on the green one. It saves the information that it can achieve the chosen subgoal by applying the action STACK. Its next subgoal must be to place the green block on the blue one. It cannot pick up the green block, however, as long as the green block supports the red block because the operation of STACK requires that the top of the green block be clear. To stack the green block on the blue one the program must remove the red block from the top of the green one, thus "clobbering," or undoing, the subgoal it has just achieved.

The critics in HACKER are programmed to call in the debuggers whenever they detect that a subgoal has been clobbered. The debuggers look for planning options, and in this case they find that if the green block is first put on the blue one, the plan works without clobbering any subgoals. Hence the debuggers can conclude that the clobbering of subgoals can be avoided, at least sometimes, by rearranging them. HACKER saves the conclusion as a new, general debugging method. Sussman's program eventually learned to order the subgoals for a stack made up of any number of blocks.

Concept Learning

A number of programs enable a computer to carry out an action more competently or to form a more adequate representation of a concept. One of the earliest programs to improve its own performance was a checkers-playing program written by Arthur L. Samuel of the International Business Machines Corporation in about 1960. Learning in the program was designed to simulate the mechanisms of evolution and natural selection. One version of the program played against another version in which the scoring system for evaluating the strength of a position had been altered slightly. In this way a given scoring system and a "mutation" of the system were tested in direct competition, and the winning version was retained as the basis for constructing new variations.

In the late 1960's Patrick H. Winston of M.I.T. wrote a program that successively improves its mastery of certain concepts. A teacher presents the program with a carefully graded sequence of scenes arranged so that the first scene is a good example of the concept the program is to learn. The program analyzes the first scene and builds a "hypothesis" about what components of the scene and what relations among them are essential aspects of the concept. For each subsequent scene the teacher tells the program whether or not the scene is an example of the concept. The program



UNDERSTANDING OF COMMANDS is accomplished by the program SHRDLU by following procedures like the one shown in this flow chart. Key words in the command "Pick up a big red block that supports a pyramid" initiate a logical series of steps that determine which block the command refers to. SHRDLU will then develop a plan to carry out the command.

tests and revises its initial hypothesis as each new scene is presented.

Suppose the program is presented with a simple line drawing and told the drawing is an example of an arch [*see illustration on page 131*]. The program notes that the object represented in the drawing is made up of three blocks; two of them stand upright on a table and support the third. Although the program might have guessed many other properties of the example that are true of arches in general, its concept of an arch is kept deliberately impoverished so that the concept can be enriched by analysis of the later scenes.

The second scene is not an arch. The program again constructs a description of the scene and compares the description with the hypothesis it has built up from the first example. If the second scene does not differ from the first one in too many ways, the program can begin to determine which relations are essential to the concept of an arch. If the second scene is identical with the first one except that the block that was on top is now lying on the table, the program infers that in an arch the two standing blocks must support the third block. The program has learned that the relation of support between the lower blocks and the upper block was not an incidental aspect of the first scene, and the hypothesis about the concept of an arch is updated to reflect the finding. As additional scenes are introduced the representation of the concept by the program becomes more adequate for recognizing novel instances of an arch.

Recently a program called AM (for automated mathematics) was developed by Douglas B. Lenat of Stanford that can formulate new concepts and theorems in mathematics from about 100 elementary concepts in set theory. Heuristic principles specify the creation of the new concepts under certain well-defined conditions, and the program then proceeds to investigate the concepts. During one run on a computer AM considered the concept of the divisors of a number. It found several numbers that have four or more divisors (for example, the divisors of 6 are 1, 2, 3 and 6), but the heuristic rules confined the investigation to numbers with only a few divisors.

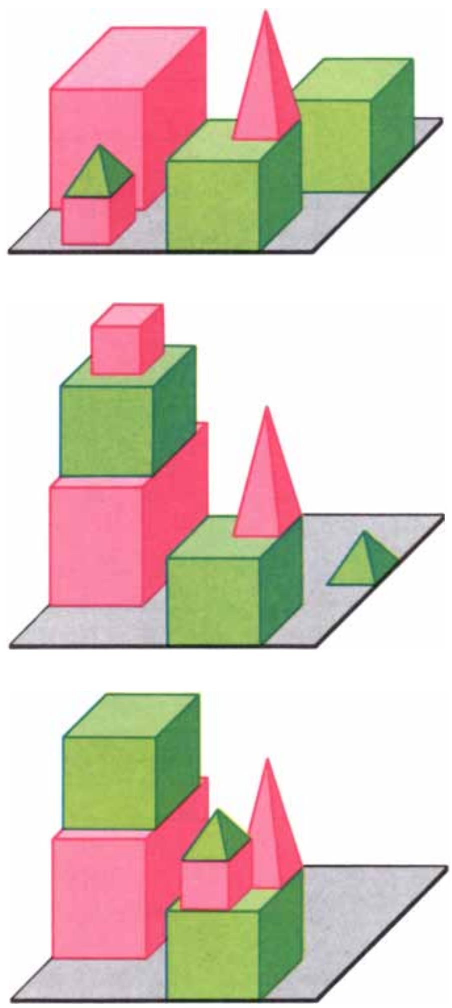
In considering the numbers with exactly three divisors the program found that all the examples are also perfect squares. Moreover, the program found that the square root of a number with three divisors is always a number with exactly two divisors. Because of the coincidence of the concept square-roots-of-numbers-with-three-divisors and the concept numbers-with-two-divisors, the program raised the priority of both concepts on its agenda and determined to investigate numbers with exactly two divisors in considerably more detail. In this way AM began to explore the rich pattern of mathematical relations that develops from the concept of numbers with exactly two divisors, otherwise known as prime numbers. In about an hour of running time on the computer Lenat's program went on to reproduce several well-known conjectures about prime numbers and guessed that every natural number is the product of a unique set of primes.

Constraint Propagation

The kind of search I have described for chess programs and planning programs is a serial process, and so the size of the tree that can be searched is limited by the time needed to evaluate every node in the tree sequentially. In some cases a process that seems to require a serial tree search can be recast as a parallel process, in which independent aspects of the search are done simultaneously. Much of the signal processing carried out by the brain is done in parallel, and many investigators think programs that incorporate parallel processing will be much better able to simulate human sensory abilities than programs that rely on serial processing. A number of programs in artificial intelligence have experimented with a parallel process called constraint propagation, an early example of which I presented in my Ph.D. dissertation at M.I.T. in 1972. My program was written in LISP, a programming language now employed for most work in artificial intelligence.

One of the things people learn to do almost at a glance is to interpret a two-dimensional line drawing as a three-dimensional scene. In my dissertation I wanted to develop a program that could mimic this skill. Earlier independent

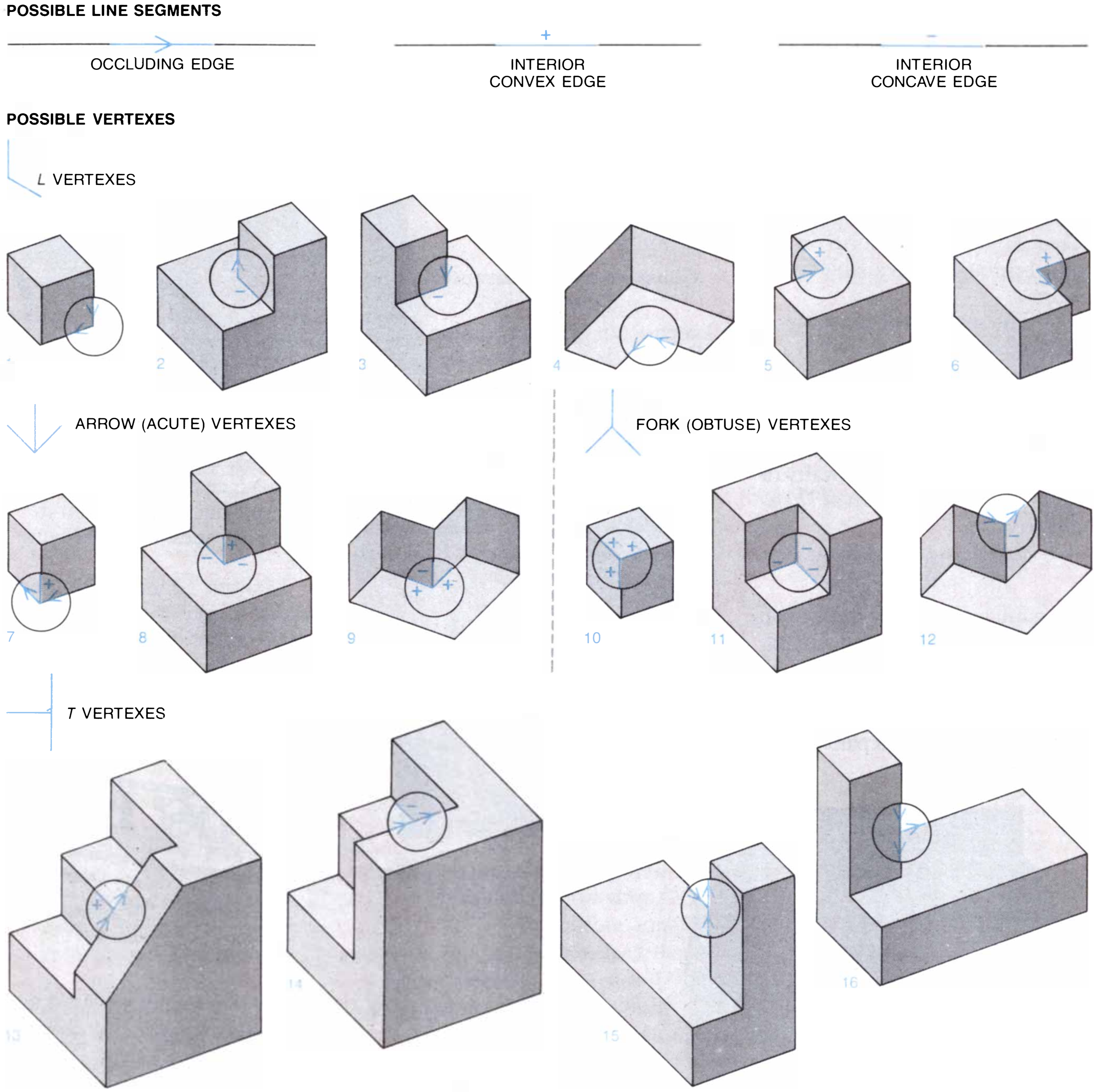
Will you please stack up both of the red blocks and either a green cube or a pyramid?
OK.
Which cube is sitting on the table?
The large green one which supports the red pyramid.
Put a small cube onto the green cube which supports a pyramid.
OK.
Put the littlest pyramid on top of it.
OK.
How many things are on top of green cubes?
I'm not sure what you mean by "on top of" in the phrase "on top of green cubes."
Do you mean:
1—Directly on the surface
2—Anywhere on top of?
2
Three of them.
Had you touched any pyramid before you put the green one on the little cube?
Yes, the green one.
When did you pick it up?
While I was stacking up the red cube, a large red block and a large green cube.
Why?
To get rid of it.
Why did you do that?
To clear off the red cube.
Why did you clear off that cube?
To put it on a large green cube.
Why did you do that?
To stack up the red cube, a large red block and a large green cube.
Why did you do that?
Because you asked me to.



EXTREME SIMPLICITY of the blocks world makes it possible for the program SHRDLU to give a wide and flexible range of responses to questions by the programmer. The program is able to plan and execute the manipulation of the objects in its world on command from the programmer. As the programmer issues the commands listed at the left (*black type*) the initial state of the blocks world (*top right*) is transformed to reflect the results of the actions that carry out the commands (*middle and bottom right*). As the dialogue proceeds the program queries the programmer about an ambiguous question and answers questions about the state of the blocks world, its previous actions and the reasons for its actions (*colored type*). SHRDLU was one of the first programs to combine both language understanding and the simulation of action.

work by David A. Huffman of M.I.T. and the late Max B. Clowes of the University of Sussex had shown, under certain simplifying assumptions, that a line drawing can be interpreted as a three-dimensional scene only if all its vertexes can be labeled consistently from a set of 16 allowable kinds of vertex. The scheme developed by Huffman and Clowes applies only to drawings of objects whose surfaces are all planes and in which no vertex represents the intersection of more than three planes. Each line emanating from a vertex must be labeled as either an occluding edge, a convex edge (pointing toward the viewer) or a concave edge (pointing away from the viewer).

A scene can be labeled consistently if every vertex is labeled and every line is assigned only one label. If two vertexes connected by a single line cannot be labeled in such a way that the line is unambiguously identified, the drawing cannot represent a real three-dimensional object. An example of a drawing that cannot be labeled consistently is the so-



RECOGNITION of a line drawing in two dimensions as a perspective view of a three-dimensional object can be achieved by a computer program that applies a labeling scheme to the vertexes of the drawing. It is assumed that every line segment in the drawing is either an occluding edge of an object in perspective, an edge of the object pointing toward the viewer (a convex edge) or an edge of the object pointing away from the viewer (a concave edge); furthermore, no more than three planes are allowed to meet at any point. An occluding edge is the boundary of an occluding plane, a plane at the exterior of the object or a plane that lies in front of another plane of the object from the viewer's position. By convention an occluding edge is indicated by an arrow, and the occluding plane lies to the right of the edge when it is viewed along the direction of the arrow. There are four categories of vertexes and 16 kinds of vertexes in all; every vertex in the plane figure must be classified as one of the 16 kinds of vertexes before the interpretation of the object as three-dimensional can proceed. The labeling scheme was first proposed by David A. Huffman of M.I.T. and the late Max B. Clowes of the University of Sussex; the scheme was later generalized by the author in order to make possible the recognition of shadows, cracks and other kinds of edges.

Now you can shrink a $600 car repair down to $25.

Everywhere you look these days, it's $600 for a transmission job here, $400 for engine repairs there.

It's gotten to the point where even the remote possibility of major repairs is enough to worry anyone.

Until now.

Because now you can cut those worries down to a more manageable $25. And keep them that way for up to five full years.

Introducing the Ford Extended Service Plan—for purchasers of Ford or Lincoln-Mercury cars or light trucks.

Here's how it protects you.

The Extended Service Plan covers thousands of parts, including major repairs. And promises no parts or labor charges, except a small $25 deductible each time you bring your car in. No matter how many different covered parts need to be fixed.

Which means you can go to any one of 6,300 Ford or Lincoln-Mercury Dealers across the U.S. and Canada, and the most you'll pay is $25 for any covered repairs.

Plus there's much more. With plans ranging up to 5 years/60,000 miles of maximum protection. See your Dealer for all the details on how to shrink car repair worries down to size.

Ask for the Extended Service Plan by name. It's the only one backed by the Ford Motor Company.

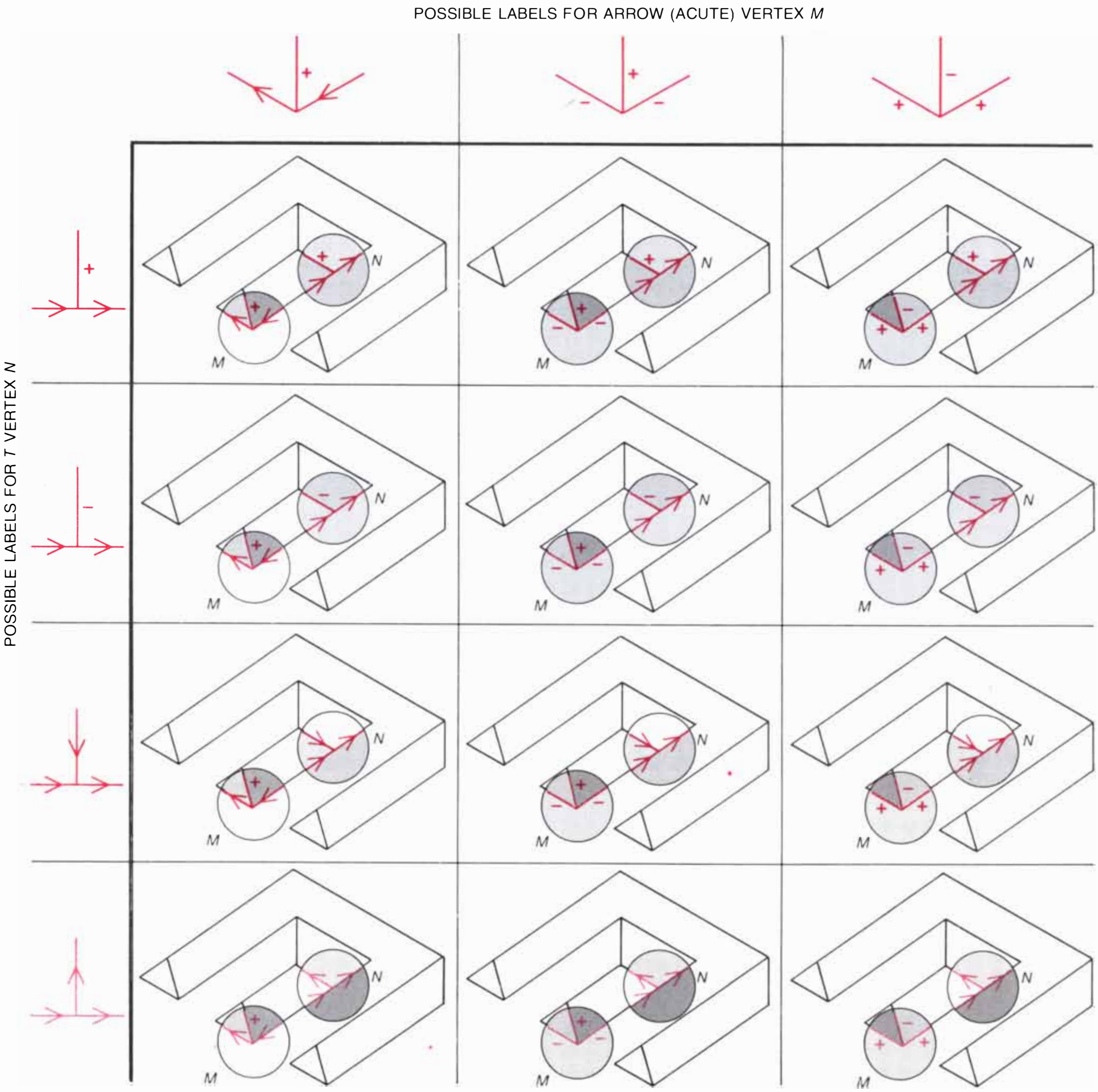
Get it together—buckle up.

Ford Extended Service Plan Ford

called devil's pitchfork, which initially resembles an object but which on closer inspection generates a disconcerting visual paradox. The labeling scheme enables a computer program to rule out the possibility that the drawing represents an object and also helps to dispel the paradox [*see illustration below*].

The first programs for labeling the vertexes of a line drawing ran a serial tree search for consistent labels. Beginning at some arbitrary vertex the programs would assign all possible labels to the vertex. For each label the programs would then proceed to an adjacent vertex and assign all the labels to the second vertex that were consistent with the label chosen for the first vertex. For each consistent label at the second vertex the programs would determine the possible labels for a third vertex adjacent to the second one. The procedure would continue in sequence through all the vertexes until every branch of the tree had been explored. Each branch of the search would end either when no further vertex labeling was possible or when all the vertexes had been labeled consistently. (For some drawings it may happen that consistent labels can be applied to the vertexes in more than one way; such drawings can appear ambiguous to human observers as well.)

Tree-search methods are adequate when the number of vertex labelings is small, but I wanted my program to distinguish cracks, shadows and other kinds of edge that earlier programs could not correctly interpret. When the additional kinds of edge are distinguishable, however, the number of possible labels for each vertex becomes quite



INCONSISTENT LABELING of a line segment, no matter what combination of possible vertex labels is tried, shows that a two-dimensional line drawing cannot be a perspective view of a three-dimensional object. In the drawing of the impossible object called the devil's pitchfork all possible combinations of vertex labels are tried for the ends of the line segment *MN*, but no combination gives the same label to *MN* along its entire length. There are drawings, however, that can be labeled consistently but cannot represent a real object.

large: about 100 for certain kinds of vertex and more than 1,000 for others. For a scene having only four vertexes the number of ways to label the scene can be greater than $1,000^4$, or a trillion. Even though a search of the complete tree for such a scene would not have to consider every one of the labeling schemes (if only because many of them would be found to be inconsistent before the last vertex in the scene was examined), the search would still be unmanageable.

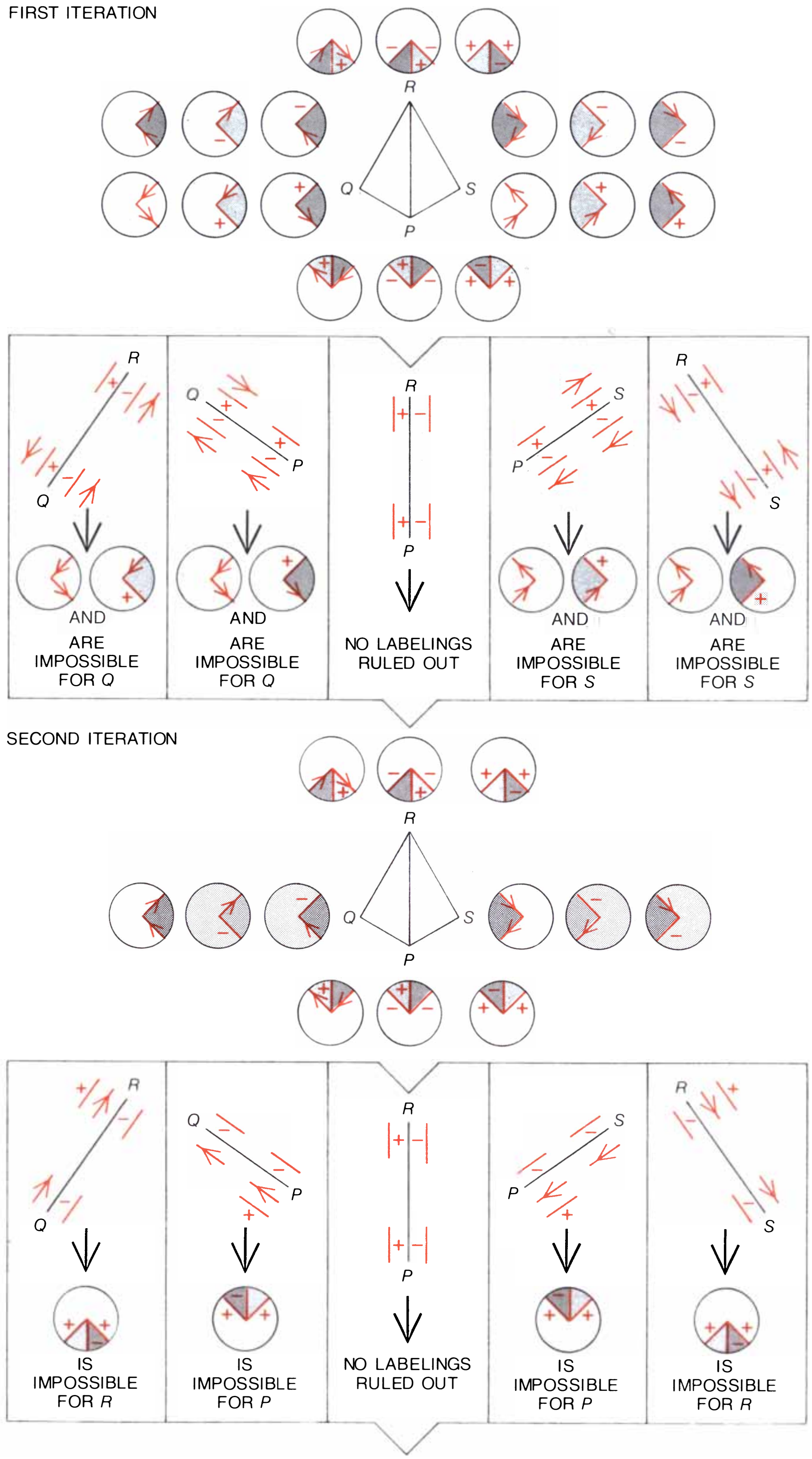
I decided, therefore, that my program should eliminate all the impossible labels for each vertex before undertaking a search. The program assigns all the labels that are possible at each vertex. It then examines every pair of vertexes connected by a line. Because of the requirement that the line have a single label some of the labels for the two adjacent vertexes may be eliminated: it may turn out that a label at one vertex is not compatible with any of the labels at the adjacent vertex. Every pairwise examination of labels is independent of every other one, and so the processing in this stage of the program is done in parallel. Once the first round of label eliminations has been completed all the surviving vertex labels are reassigned and pairwise examinations are done in parallel once again. Thus the constraints identified in the first round are propagated to the second and subsequent rounds [*see illustration at right*].

I had originally meant to use the pairwise elimination of labelings simply to cut the tree search down to a manageable size. On the first scene I tried I found that the process reduced the number of labels for each vertex to one. Hence a tree search was unnecessary.

The discovery was a surprise, and indeed I thought there were errors in the program when it first gave unique labelings to scenes. Moreover, if I had not written the program to be run on a computer but instead had attempted the labeling procedure by hand, the dispensability of the tree search would probably not have been discovered. The simplest manual labeling would have called for examining some 10,000 vertex labels; in such a tedious undertaking error is almost inevitable. The possibility of such unexpected findings is seldom appreciated outside the field of computer science, but it is not at all unusual.

Subsequent attempts to apply my program to more realistic scenes have also shown that experiment is an essential aspect of research in artificial intelligence. Although the program turned out to be an elegant and efficient one for the scenes to which it was originally applied, the work has not generalized well to scenes that include curved objects, textures, shiny surfaces and other features of real objects.

The most impressive recent results in computer vision were achieved by the



PROPAGATION OF CONSTRAINTS on the labeling of vertexes can be analyzed in parallel, in much the same way as some signals are thought to be processed by the brain. Each vertex is assigned all possible labels, and each possible label is compared with all the possible labels of a vertex at the other end of a line segment connecting the two vertexes. If the line segment cannot be labeled consistently (so that it has the same label at both ends), all the vertex labels leading to the inconsistent labeling are eliminated. The constraints on the labeling of the line segments are given only by the colored symbols; the shading has been added for clarity. Because the process can be done independently for each line segment the segments can all be analyzed simultaneously. Once all the constraints have been determined at each vertex they are combined. The labels that remain are then applied to the vertexes and the procedure is repeated.

late David C. Marr of M.I.T. Marr modeled the operations of the visual cortex of the brain that may be needed for dealing with scenes in their full complexity. Techniques related to my work, however, have been useful for analyzing electronic circuits, for carrying out stereoscopic correlations of two views of the same scene, for comparing detailed features of a scene with other features nearby and for reducing the ambiguity of certain linguistic phrases.

Language Understanding

The medium for human interaction with computers is often called a language, but programming computers to understand ordinary natural language is one of the most difficult challenges now facing the discipline of artificial intelligence. Even the simplest programs for understanding language are large and complex, and the most powerful programs are still confined to narrow semantic domains in which they attend primarily to the most superficial meanings. Nevertheless, the area is of great importance, not least because a successful program would simulate processes that seem to be close to the essence of human thought.

The earliest work on the understanding of language by computer programs began in the 1950's and was aimed at mechanizing translation. The general approach was to provide a dictionary equivalent for each word in the material to be translated; simple rules were given to reorder the words in the translation and improve its syntax. The effort was a failure. When the sentence "The spirit is willing but the flesh is weak" was translated into Russian and back into English, it is said to have come out as "The vodka is strong but the meat is rotten." Good translation without understanding proved impossible, and by the middle of the 1960's the work had largely been abandoned.

Another early approach to language understanding is a program called ELIZA, written by Joseph Weizenbaum of M.I.T. in 1966. ELIZA bypasses any real linguistic processing and instead relies on a clever system of rather fixed patterns of response that give an imitation of language understanding many people find convincing. The responses of the program mimic those of a psychiatrist; each response is called up from a set of stored sentences or sentence patterns that are associated with words or patterns of words in statements made by a "patient." Whenever the word "mother" is mentioned, for example, the program replies with one of several stock sentences, such as "Tell me more about your mother." If the patient types "I'm feeling a bit tired," ELIZA can embed part of the patient's sentence in its reply: "Why are you feeling a bit tired?"

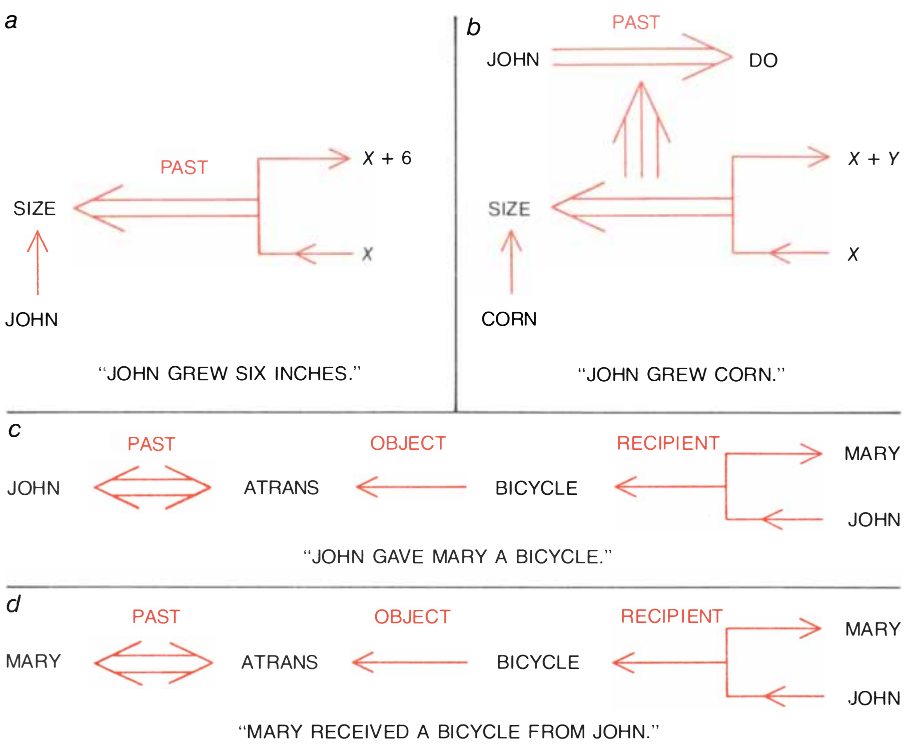
Although many of the input words are ignored by the program, it still requires a large library of meaning patterns to deal with a large number of potential input statements. Weizenbaum later argued that the program demonstrated the shortcomings of making the simulation of human behavior a criterion for ascribing intelligence to computer programs. ELIZA, he pointed out, operated on an extremely simple and superficial level of language understanding, yet many people were lured by its lifelike responses into reciting their personal problems to the machine just as if it were a psychiatrist.

In about 1970 Roger C. Schank of Yale University introduced programs for processing natural-language sentences about human actions. The programs are based on what Schank calls primitives of conceptual dependency. The primitives include MTRANS, the set of all actions in which there is a transfer of mental information, such as telling, hearing, writing and reading; ATRANS, the set of actions in which there is a transfer of possession, such as buying, selling, giving and taking, and ATTEND, for actions that involve sensing, such as watching, listening, smelling and tasting.

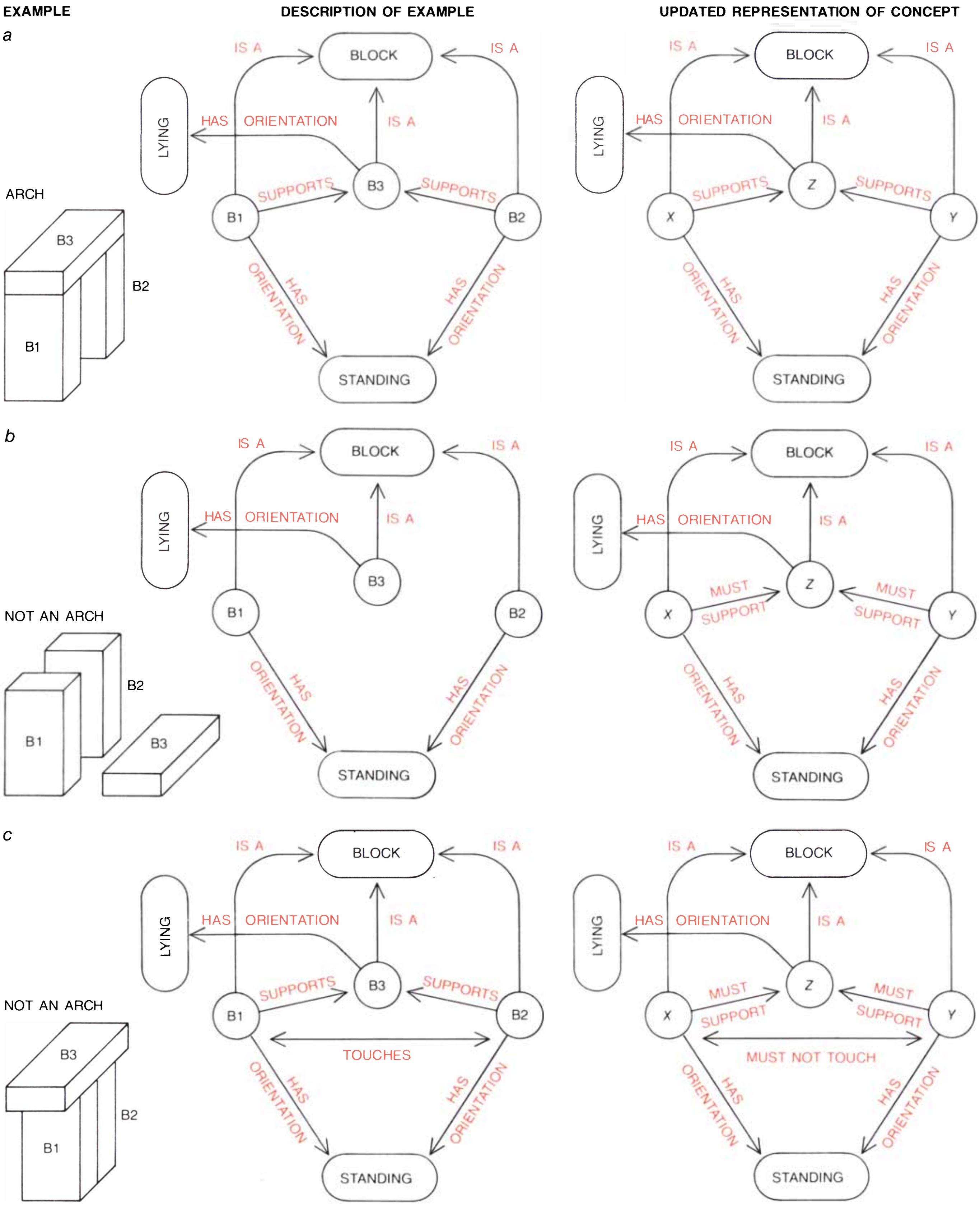
It is possible to represent the meaning of a sentence in a diagram based on conceptual dependencies. It turns out that sentences that are superficially similar but have different meanings have quite different diagrams. Moreover, the diagrams representing the semantic structure of a sentence and a paraphrase of the sentence are similar even though the two sentences may be superficially different. In addition the conceptual dependencies organize the "expectations" a program has after analyzing part of a sentence. For example, after the sentence fragment "John gave" the program expects the sentence to name a recipient of the action as well as an object that was given. The program sets up slots, or blank spaces, in its provisional understanding of the sentence to hold the expected items.

Frames and Scripts

Schank's work was part of a long-term trend toward the view that language understanding is a "top down" process, in which words serve as cues for retrieving expectations from memory and as evi-



DIAGRAMS OF CONCEPTUAL DEPENDENCY represent the semantic structure of a sentence in such a way that sentences that are superficially similar but have different meanings are represented differently (*a, b*). Sentences similar in meaning but superficially different in structure are represented in similar ways (*c, d*). The diagrams show that the sentence "John grew six inches" corresponds roughly to "John's size went from some value *X* inches to some value *X* + 6 inches" (*a*), whereas the sentence "John grew corn" corresponds roughly to "John did something unspecified that caused the size of the corn to increase from some size *X* to some size greater than *X* by the amount *Y*" (*b*). The sentence "John gave Mary a bicycle" corresponds roughly to "John transferred possession of the bicycle from himself to Mary" (*c*). (ATRANS is a class of verbs that have to do with the transfer of possession.) The sentence "Mary received a bicycle from John" is given a similar representation (*d*), except that Mary is listed as the agent who caused the transfer. The system was developed by Roger C. Schank of Yale University.



LEARNING A CONCEPT can be understood as the formulation of a hypothesis about the aspects of the concept that are embodied by a particular example and the revision of the hypothesis to accommodate information gained from further examples. To teach a computer program the concept of an arch the teacher presents the program with a picture of what he wants to call an arch (*a*). The program describes the picture by means of a semantic net, or a network of relations among the concepts that are already incorporated into the program. The first hypothesis is a generalized semantic net in which variables are substituted for the particular components of the picture. The teacher then presents the program with additional pictures that are not arches (*b*, *c*). The program combines a description of each image with the information that it does not represent an arch and thereby revises its original hypothesis. The sequence of examples must be chosen carefully; an example that differs in too many ways from the preceding one will "confuse" the program. It will not be able to discern which aspects of the examples are important and which are incidental. The program was developed by Patrick H. Winston of M.I.T.

dence for or against established expectations. The view is in contrast to an earlier one in which meaning was to be constructed "from the bottom up," that is, from the definitions of words to their meaning in phrases and finally to their likely function in a sentence.

In 1974 Marvin L. Minsky of M.I.T. suggested that not just language but all thinking might depend heavily on processes driven by expected structures of knowledge he called frames. A frame, like a primitive of conceptual dependency, is made up of a core and a set of slots. Each slot corresponds to some aspect of or participant in a concept implicitly defined by the frame. Minsky argued that an important function of a frame is to represent a stereotype. The stereotype is an intuitively plausible model of the process by which people fill in information about a situation that is not explicitly mentioned.

A number of language-understanding programs now exploit this general strategy. For example, a program written by Wendy G. Lehnert of Yale can answer questions about stories such as the following:

John took the bus from New Haven to New York. On the way his pocket was picked. He went to a restaurant and ordered spaghetti. John could not pay the bill and so he washed dishes.

Although the passage states that John merely ordered spaghetti, most people would assume from the story that John also ate the spaghetti. Lehnert's program makes such plausible inferences from a special kind of frame called a script, in which the background expectations most people have about going to restaurants are made explicitly available to the computer.

A program called FRUMP, developed by Gerald F. DeJong II of Yale, can summarize news stories as they appear on the wire service of United Press International. The program includes scripts for many events such as earthquakes, automobile accidents and diplomatic visits. DeJong, who is now at the University of Illinois at Urbana-Champaign, is making additions to the program that will enable it to write its own scripts. The program might be given the text fragment "Mary wanted a new radio; she went to the bank." The program would consult its script for obtaining new objects and its script for going to banks, looking for information that would relate the two parts of the sentence. The program would note that new radios cost money and that people go to banks to withdraw money, and so it would "reason" that Mary probably went to the bank to get the money to buy the radio. Similarly, if the program knew nothing about kidnapping, it could develop a plan to figure out why someone would steal a person. Once the analysis was finished and generalized the plan could be added as a kidnapping script.

Many other issues in language understanding have recently come to the fore. Programs have been developed that attempt to model the understanding of metaphor, the understanding of unstated goals and beliefs of the speaker and the appreciation of the emotions and the underlying motivations of characters in stories. Other programs are intended to deal with inconsistent information and to judge the plausibility of sentences.

The judgment of plausibility is often central to the understanding of ordinary language. Before a program can understand metaphors, humor, lies or exaggerations it has to recognize that the literal meaning of the sentence is implausible for the given situation. Sentences can fail to be plausible for a variety of reasons. For example, the object of the verb may be inappropriate for its action, as in "John ate up the compliments," or a sentence may posit physically implausible events, as in "Mary jumped 10 feet when she heard the news." I regard work on plausibility judgment as an early step toward modeling common sense.

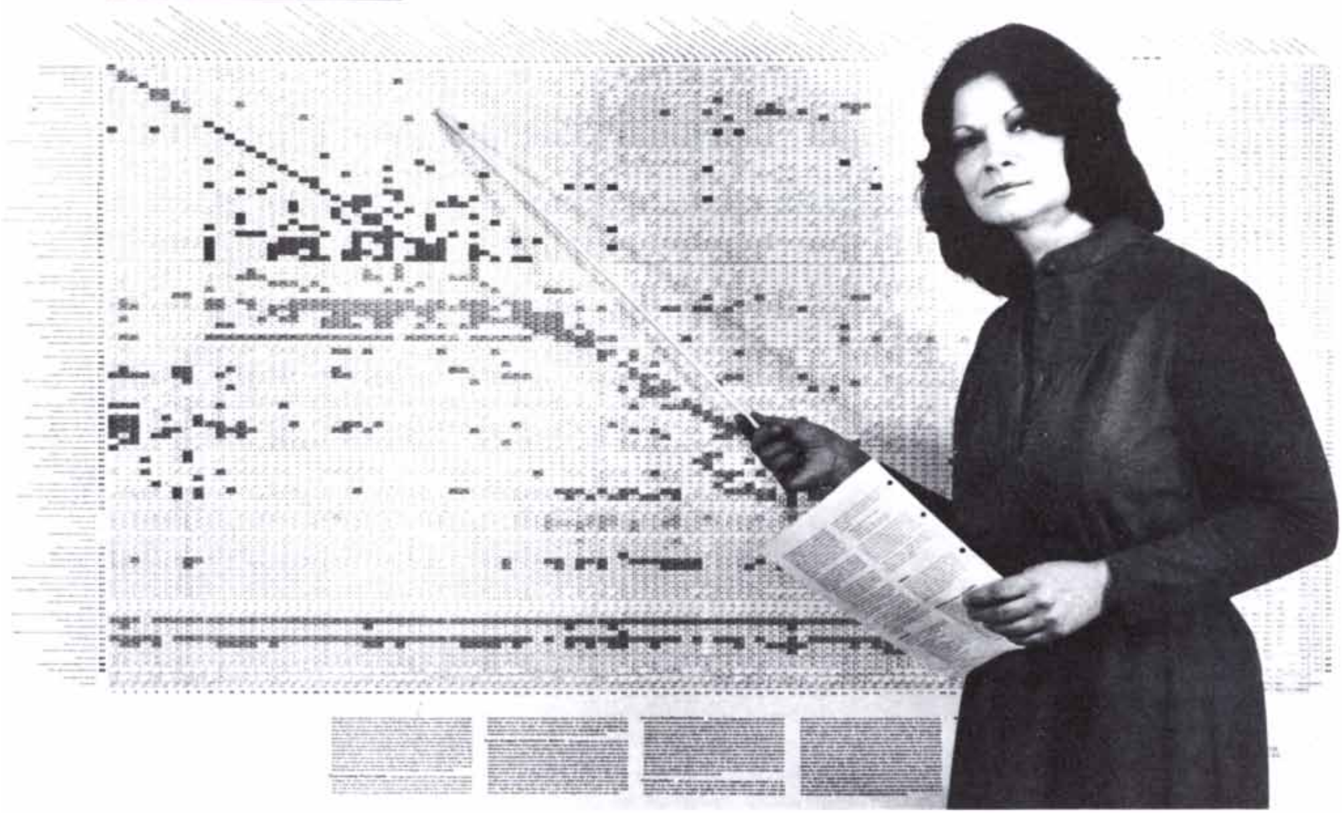
Common Sense

Probably the most telling criticism of current work in artificial intelligence is that it has not yet been successful in modeling what is called common sense. One difficulty in simulating common sense is that a program must link perception, reasoning and action simultaneously, because ultimately the intelligent use of a concept depends on all three domains. The best current programs for language understanding literally do not know what they are talking about; their only contact with the world is through language.

The greatest successes in artificial intelligence have come about where it has been possible to identify narrow, self-contained domains such as the blocks world of SHRDLU. In my opinion, however, experience has shown that such programs cannot simply be enlarged and generalized in a straightforward way. Much conflicting and incompatible knowledge, real scenes and unrestricted processing of language will probably not yield to the engineering techniques that have been applied so far.

This observation does not mean that engineering methods will not be extremely useful in uncovering general principles of intelligence. Nevertheless, substantially better models of human cognition must be developed before systems can be designed that will carry out even simplified versions of common-sense tasks. I expect the development of such models to keep me and many others fascinated for a long time.

THE INPUT/OUTPUT STRUCTURE OF THE UNITED STATES ECONOMY



WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN

415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $_____ for _____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $_____ for _____ chart(s) at $40 each.

Name____________________

Company____________________

Street Address____________________

City, State, Zip Code____________________

*Charts shipped international require one fold. 10/82

Make check payable to SCIENTIFIC AMERICAN. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

Magnetic Fluids

When small ferromagnetic particles are suspended in a liquid carrier, the resulting "ferrofluid" exhibits unique properties that are intuitively baffling as well as technologically useful

by Ronald E. Rosensweig

Imagine a liquid with the properties of a ferromagnetic material: like an ordinary bar of iron it would become spontaneously magnetized in the presence of an external magnetic field, but it would also act as a fluid, capable of assuming the shape of its container and of flowing downhill or around obstacles. The interaction of the magnetic and the fluid properties could give rise to quite unusual phenomena. For example, a solid ferromagnet, such as a compass needle, can be induced to move or to change its orientation by an external field. In a "ferrofluid" the external field could act almost independently on every microscopic volume of the liquid, leading to a response much more complex than the motion of a rigid body. One likely result is that the ferrofluid would alter not only its position and orientation in response to the field but also its shape. A continuously changing field could even cause the magnetic liquid to flow. Indeed, virtually every aspect of hydrodynamics could be considered afresh in the expectation that magnetism would modify fluid properties in remarkable ways.

Ferromagnetism is a property of iron, nickel, cobalt and some compounds and alloys of these elements. One approach to creating a ferrofluid might therefore be to heat a metal until it becomes molten, but this strategy cannot work. The reason is that ferromagnetism disappears above a certain temperature, called the Curie point, which is invariably well below the melting point of the material. It is possible nonetheless to create a stable magnetic fluid. The essential idea is to form a colloid in which minute ferromagnetic particles are suspended in a liquid carrier.

In the past 15 years or so methods of preparing colloidal ferrofluids have been developed and the peculiar physical properties of the liquids have been explored in some detail. A few applications of ferrofluid technology have been introduced and others are clearly in prospect. A ferrofluid can serve as an airtight seal in rotating machinery or as a coolant for the voice coil of a loudspeaker. Ferrofluids are the basis of ingenious new techniques for the separation of materials according to density. There is also the possibility of developing a heat engine or a heat pump whose operation would depend on the magnetic properties of the working fluid.

The idea of creating a colloidal fluid with ferromagnetic properties was put forward independently and almost simultaneously by several investigators. One of the first and most easily prepared colloidal systems was developed by Stephen Papell of the National Aeronautics and Space Administration in the early 1960's. Papell's fluid consisted of finely divided particles of magnetite (a molecular mixture of the iron oxides FeO and Fe_2O_3) suspended in kerosene. To keep the particles from clumping together or settling out of the suspension Papell added oleic acid, an organic substance that served as a surfactant, or dispersing agent.

As it happens a suspension of coarse magnetic particles had been in use since the 1940's in magnetic clutches, but Papell's colloid resembled a clutch fluid only superficially. The most important distinguishing characteristic is the size of the particles. Under the influence of a magnetic field a clutch fluid congeals into a solid mass; hence when it is magnetically active, it is not a fluid, and vice versa. The particles in a ferrofluid are 1,000 times smaller in linear dimensions (a billion times smaller in volume), with the result that a ferrofluid does not congeal when it is subjected to a magnetic field. It becomes magnetized, but it remains a liquid.

I began my own study of magnetic fluids in the 1960's at the Avco Corporation, where, in a systematic program, my colleagues and I succeeded in formulating magnetic fluids that were some 10 times as strong magnetically as Papell's original ferrofluid. In collaboration with Robert Kaiser we achieved considerable flexibility in the choice of magnetic particle, carrier fluid and dispersing agent. At the same time we elucidated the physical principles responsible for the properties of magnetic fluids.

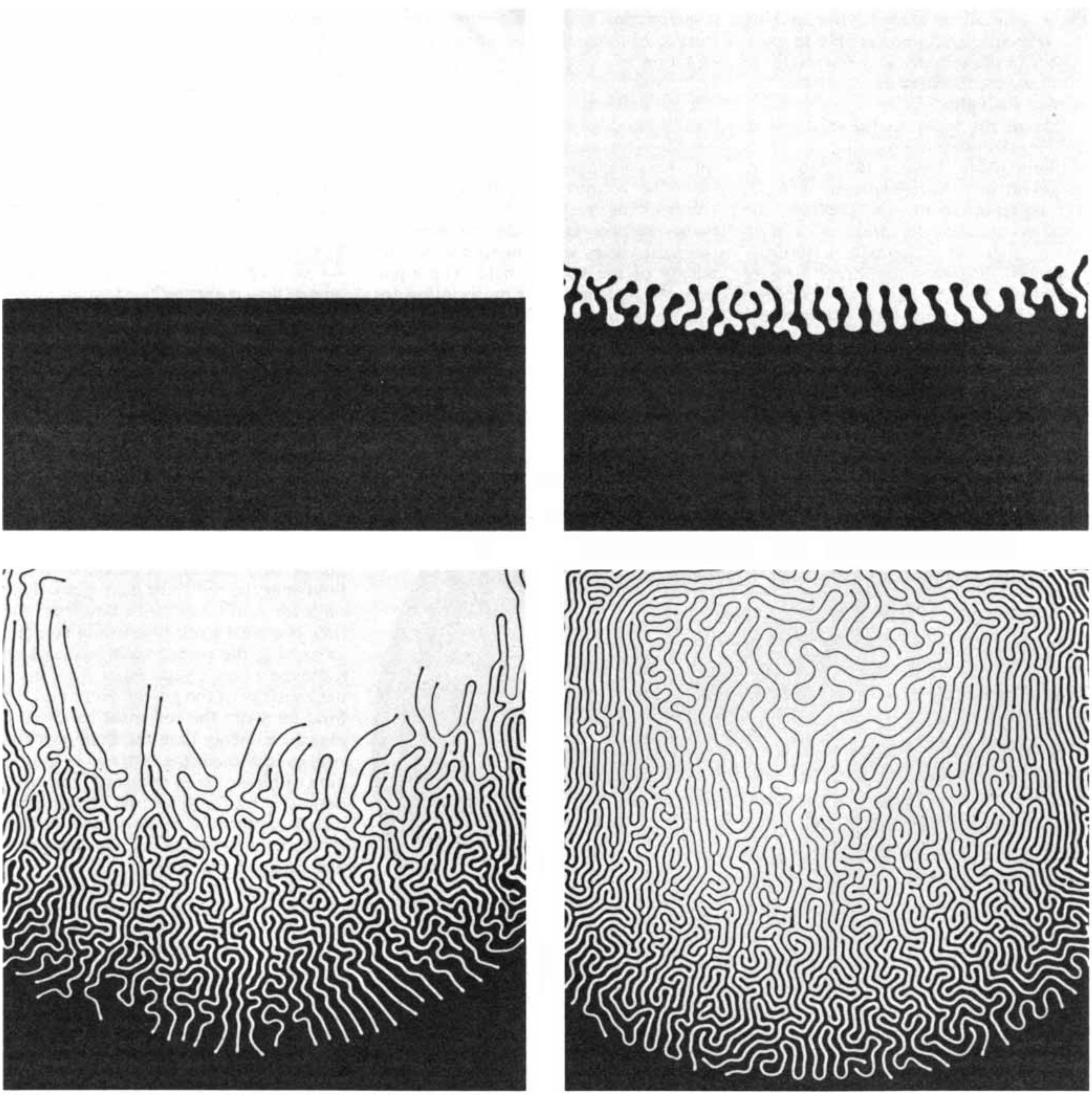
When a magnetic field is applied to a ferrofluid, a body force is developed within the liquid. A body force is simply a force experienced throughout a given volume of the fluid (although the magnitude and the direction of the force are not necessarily the same everywhere). The effects of the magnetic body force can be exhibited by allowing a stream of ferrofluid to fall freely past the poles of a horseshoe magnet. The field generated by the magnet, acting on all the particles in the volume of fluid that lies near the poles at any instant, deflects the stream from a vertical path.

A more striking demonstration of the magnetic body force can be seen when a metal rod is placed vertically in a shallow pool of ferrofluid. If a magnetic field is imposed by passing a steady electric current through the rod, the fluid leaps up and surrounds the rod, forming a symmetrical concave meniscus, wide at the base and tapering at the top. The height of the meniscus accurately maps the decrease in the intensity of the magnetic field with distance from the rod.

In ordinary hydrodynamics the only body force acting from the outside on the entire volume of a fluid is the force of gravity. In magnetohydrodynamics an ionized gas or a liquid metal that is carrying an electric current in the presence of a magnetic field is subject to another body force called the Lorentz force, which represents the interaction of the current with the magnetic field. In electrohydrodynamics ions or other particles that bear an electric charge are acted on by an electrodynamic force when they are exposed to an electric field. In the case of a ferrofluid there are no electric currents or electric charges. The body force originates from the interaction of a magnetic field with the ferromagnetic dipole moment characteristic of each colloidal particle.

A magnetic field is described mathematically as a vector quantity, one that has both a direction and a magnitude at every point in space. The ferromagnetic dipole moment of a ferrofluid is also vectorial. The magnitude measures the strength of magnetization; the direction is determined by the orientation of an imaginary line passing from the south pole to the north-pole through any small region of the ferrofluid. In a uniform applied magnetic field the force on a north pole is in the direction of the local field present at the position of the pole, whereas the force on a south pole has the same magnitude but is in the opposite direction. Each sample of fluid, like any other magnetized body, always has equal numbers of north and south poles. Because the field ordinarily varies at least slightly from point to point the force on the sample depends on the rate of variation of the field in both direction and magnitude.

Vectors have a special calculus of their own that allows this complicated physical effect to be expressed exactly. It turns out that in a Cartesian, or rectangular, coordinate system the equation describing the force on a ferrofluid is made up of nine terms. By introducing some reasonable assumptions about the nature of the ferrofluid itself, however, one can simplify the description. Briefly, one assumes that a magnetic dipole par-

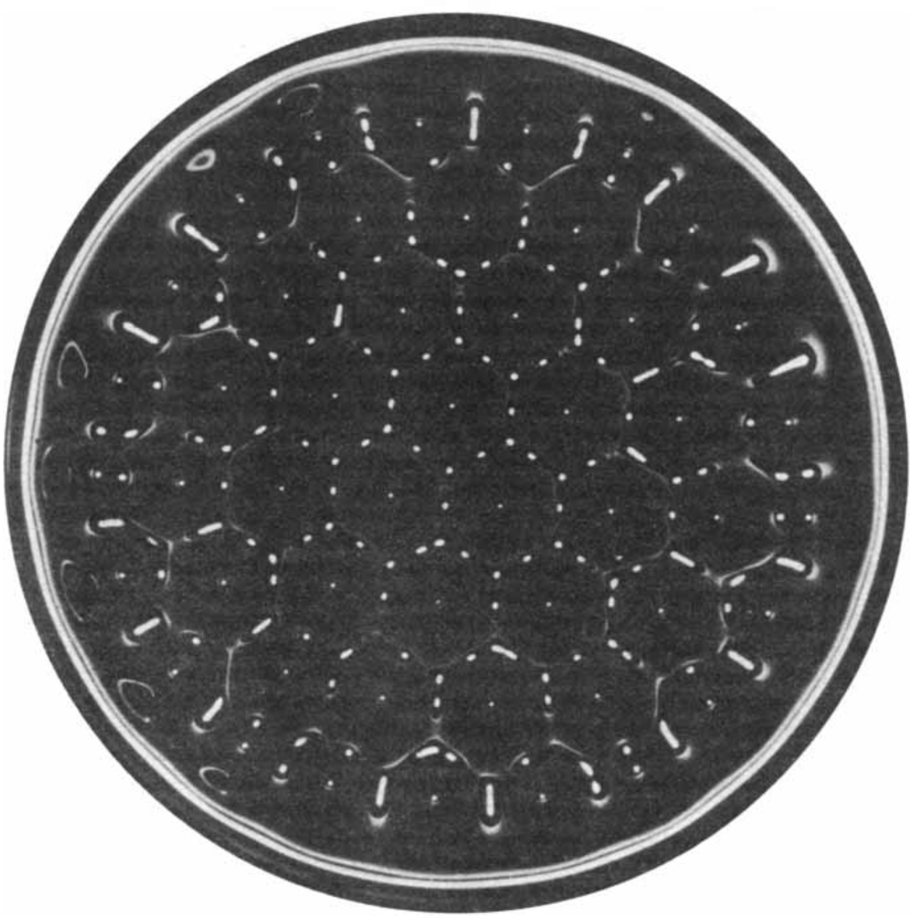


INTRICATE MAZE PATTERN evolves rapidly when a magnetic liquid (*dark*) and an immiscible nonmagnetic liquid (*light*) are subjected to a uniform horizontal magnetic field. The ferrofluid consists of fine particles of magnetite (a form of iron oxide) suspended in kerosene. The two liquids are in a glass cell 75 millimeters on a side with a one-millimeter separation between the walls. The density of the transparent nonmagnetic liquid is 18 percent less than that of the opaque ferrofluid. In the absence of a magnetic field the denser magnetic liquid will remain at the bottom of the cell indefinitely. When the magnetic field is applied, however, small perturbations grow into invasive fingers. The first three photographs are from a sequence of 36 made in the first nine seconds after the magnetic field was turned on; the fourth photograph was made at the end of 90 seconds, when equilibrium had been reached. The labyrinthine pattern formed by the liquids resembles the pattern of magnetic domains seen in thin slices of garnet used in fabricating magnetic-bubble-memory devices.

ticle in the fluid aligns itself parallel with the local magnetic field as if it were a compass needle and that no electric currents are present to alter the field. These assumptions collapse the nine-term equation for the body force into a single relation. The force acting on a colloidal magnetic particle is proportional to the product of two quantities: the absolute value of the magnetic moment and the spatial gradient of the magnitude of the applied field. (The gradient is a measure of how rapidly the field changes from point to point in space.) After making this transformation one is able to work with field magnitudes and essentially to suppress the bothersome vectorial aspect of the fields.

One of the most useful relations in fluid mechanics is the equation the Swiss mathematician Daniel Bernoulli presented in his *Hydrodynamica* of 1738. The equation relates the pressure, the velocity and the elevation of a fluid flowing in a gravitational field. Bernoulli showed that when the three forms of energy inherent in the flow—pressure energy, kinetic energy and gravitational energy—are added, their sum remains constant, provided the effects of friction are negligible. With the Bernoulli equation one can calculate such quantities as the variation of pressure with depth in a reservoir, the lift on the wing of an airplane or the force generated by wind on a sail.

In a ferrofluid magnetic interactions can also contribute to the energy. By taking into account the magnetic body force Joseph L. Neuringer of the University of Lowell and I were able to extend Bernoulli's equation and arrive at a ferrohydrodynamic equation. The influence of magnetization appears in the equation as a negative magnetic-energy term added to the sum of the three ordinary Bernoulli terms. The magnetic term is equal to the product of the magnitude of the applied magnetic field and the average magnetic moment of the fluid in the volume exposed to the field. Determining the average magnetic moment requires knowledge of how the moment varies with changes in the intensity of the applied field. The equation provides an exact prescription for describing or predicting ferrohydrodynamic effects.



HIGHLY SYMMETRICAL INSTABILITY is exhibited when a uniform magnetic field is oriented perpendicular to the horizontal surface of a ferrofluid. For the effect to appear the magnetic fluid must be capable of accepting a high degree of magnetization, which means that it must contain at least about .25 milligram of finely divided magnetite per milliliter. The points and bars that appear white are flat surfaces either above or below the level of the surrounding liquid. The hexagonal symmetry of the pattern, which does not depend on the shape of the container, has not been fully explained, although the spacing of the cells can be predicted.

The interactions of the energies represented by terms in the ordinary Bernoulli equation give rise to broad classes of basic flows. It is the interaction of pressure and kinetic energy, which depends on velocity, that determines the lift on an airplane wing. Pressure and gravity interact in a manometer or a barometer. Kinetic energy and gravity together determine the height reached by a stream of water emerging from a fountain. The pairing of the magnetic term with the other energy terms of the Bernoulli equation should yield new classes of flows peculiar to magnetic fluids. Indeed, this is the case, with the result that the number of basic flow classes is increased from three to six. The interactions of pressure and velocity, pressure and gravity and velocity and gravity are augmented by interactions of magnetization and pressure, magnetization and gravity and magnetization and velocity.

Let us consider first the interaction of magnetization and gravity. The type of flow is exemplified by the experiment described above in which a ferrofluid climbs a vertical rod that is carrying a current. The interaction can be understood by evaluating the magnetic-energy term in the extended Bernoulli equation. The magnetic field lines created by the current in the rod consist of concentric circles; the intensity of the field, or in other words the density of the lines, varies inversely with distance from the center of the rod. Because the magnetic liquid is attracted to regions of higher field intensity a parcel of ferrofluid near the rod must have work done on it for it to be moved away from the rod. The total work that would be done in moving the parcel from some finite radius to a very distant point is the magnetic energy of the parcel. Note that the fluid far from the rod must have more magnetic energy than the fluid near the rod, even though the latter is in the region of highest field intensity. The magnetic energy far from the rod is arbitrarily assigned a value of zero and as a result the magnetic energy anywhere near the rod is negative.

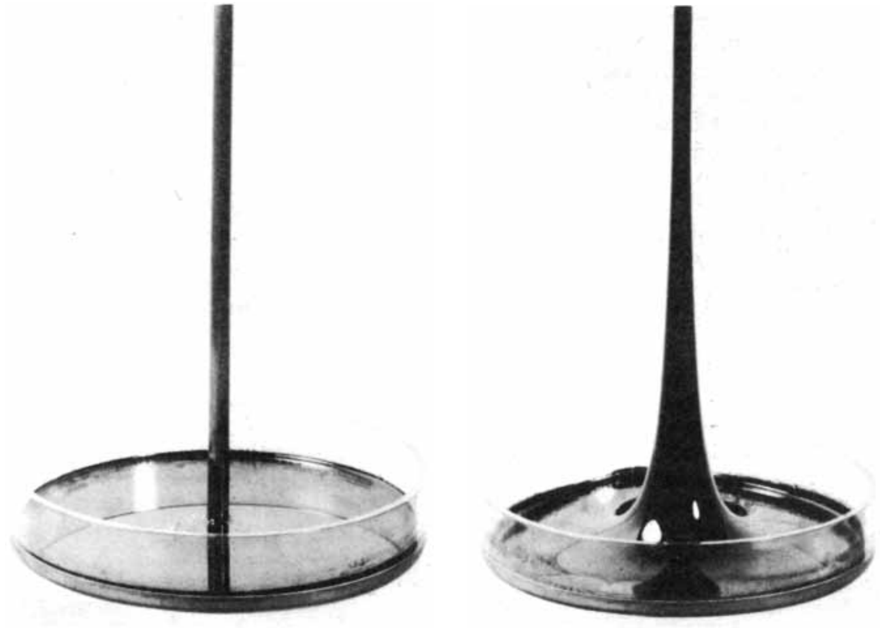
Now consider the gravitational energy of various parcels of fluid. As a parcel is elevated to a higher position along the rod its gravitational energy increases. On the meniscus the pressure is constant everywhere and the velocity is zero. Under these conditions, according to the augmented Bernoulli equation, the sum of the gravitational energy and the magnetic energy must also remain constant. The parcels of fluid nearest the rod, which have the least magnetic energy, must have the most gravitational energy and therefore climb the highest.

The interaction of magnetism and pressure has equally interesting consequences. For example, a plug, or blob, of ferrofluid can be made to form a tight seal in a tube connecting two vessels in which gases are at different pressures. A plug made up of an ordinary fluid would simply be displaced until the pressures were equalized. A ferrofluid plug can be held in place by a focused magnetic field imposed from the outside. When the pressure in the vessels is equal, the plug stays in the middle of the field. If the pressure in one of the vessels is raised, however, the plug moves only a short distance toward the region of lower pressure before it is locked in place by magnetic forces.

The mechanism of this action is also explained by the ferrohydrodynamic Bernoulli relation. The analysis is simplest if the system is horizontal (so that the gravitational energy is constant) and motionless (so that the kinetic energy is zero). The sum of the pressure energy and the magnetic energy is then constant everywhere in the plug of ferrofluid. As an imbalance of pressure displaces the plug, the fluid on the low-pressure side is pushed into the weak fringe of the magnetic field, increasing the fluid's magnetic energy. At the same time the fluid pushed into the strongest part of the field loses magnetic energy. The two changes work together to create a pressure difference in the ferrofluid that arrests the displacement of the plug.

A fluid seal is unlikely to be needed in a simple tube, where a valve or a solid plug would serve as well. The situation is different, however, where a rotating shaft must pass between gastight compartments. Indeed, pressure seals for rotating shafts were the first commercial application of ferrofluids. An optimized single-stage seal can withstand a pressure differential of one atmosphere, or 14.7 pounds per square inch.

A somewhat unexpected observation is that when a droplet of ferrofluid is exposed to a uniform magnetic field, the droplet elongates in the direction of the field. Although this effect is consistent with the expanded Bernoulli equation, it is not predicted by it. The equation states only that magnetic energy is constant for a field that is uniform within the droplet. The elongation arises from variations in magnetic stress at the surface of the droplet. The effect can be explained mathematically by evaluating differences in magnetic stress across the droplet surface with the aid of a concept called the magnetic-stress tensor. The calculation provides a boundary condition that must be satisfied along with the Bernoulli equation wherever a ferrofluid has a free surface.



BALANCE OF MAGNETIC AND GRAVITATIONAL FORCES is demonstrated when an electric current is passed through a metal rod placed upright in a pool of ferrofluid. The photograph at the left shows the experimental arrangement in the absence of current. When the current is turned on, the fluid leaps upward, as shown in the photograph at the right. At each point in the ferrofluid the sum of the magnetic, gravitational and pressure energies is constant.

Let us now consider in more detail the nature of a colloidal magnetic fluid. The kinetic theory of matter suggests that a sufficiently fine particle of matter can remain suspended indefinitely in a liquid even though the particle's density is much higher than the liquid's. The particle is kept from settling because it collides continually with molecules of the liquid, which are in random thermal motion. The molecules share their kinetic energy with the suspended particle. The energy carried by the moving particles is given in units of Boltzmann's constant k, which has a numerical value of 1.38×10^{-16} erg per degree Kelvin per particle. The energy of an individual particle is equal to kT, where T is the temperature in degrees Kelvin. For a particle to remain in suspension indefinitely this thermal energy must be equal to or greater than the gravitational energy needed to raise the particle to the height of the container holding the fluid.

When the calculation is done for a magnetite particle with a density of 5.2 grams per cubic centimeter suspended in a bottle filled to a depth of 10 centimeters with a typical oil, it turns out that the particle can be no larger than 100 angstrom units, or a millionth of a centimeter. This is the largest particle that will spend nearly equal times in all regions of the fluid.

When many particles are present in the same volume of liquid, a new possibility arises. If two or more particles collide, they may stick together, effectively creating a single heavier particle. The ratio of thermal energy to gravitational energy is thereby reduced, which can again lead to settling.

The particles in a ferrofluid are attracted to one another by two forces: the magnetic force and the van der Waals force. The magnetic forces are present even when the fluid has not been magnetized by an external field. In a bulk solid no magnetization is seen under these conditions. The reason is that such a solid is made up of many individual magnetic domains, whose magnetization vectors are oriented randomly and therefore cancel. The size of a domain is on the order of a micrometer, or about 100 times the size of a 100-angstrom particle. Each ferrofluid particle is a single domain, or subdomain, and is inherently magnetized to saturation. Although the spontaneous magnetization of the ferrofluid particles gives rise to an attractive force between them, it generally does not cause agglomeration. Precisely because the particles are so small the magnetic force is quite weak in a properly prepared fluid. The van der Waals force, however, is significant and must be dealt with.

The origin of the van der Waals force lies in the electronic structure of atoms and molecules. A molecule with an electric dipole moment (that is, an excess of electrons in one region and a deficiency elsewhere) can induce a similar distribution of electric charge in nearby molecules. The van der Waals force is the attraction that arises between dipoles, which in turn are created by fluctuations in electronic structure. According to a model devised by Fritz London, the energy needed to overcome the van der Waals force between two particles is inversely proportional to the sixth power of their center-to-center separation. For spheres of equal size the van der Waals energy is equal to the thermal energy kT when the two spheres are separated by about the radius of one sphere, a result that holds for spheres of any size. As the spheres move closer the energy needed to overcome the van der Waals attraction increases rapidly. Hence to avoid agglomeration in a colloid the particles must be kept well apart.

In preparing a magnetic fluid the necessary separation can be achieved by coating each particle with a molecular film that acts as an elastic cushion. The particles themselves can be made in one of two general ways: by reducing the size of coarse particles or by precipitating particles of the wanted size from a solution. Grinding is the most convenient method for reducing the size of large particles. Precipitation, however, is preferred for large-scale production. Whatever the method, the surfactant coating needed to ensure adequate separation is either present when the particles are made or is added to them soon thereafter.

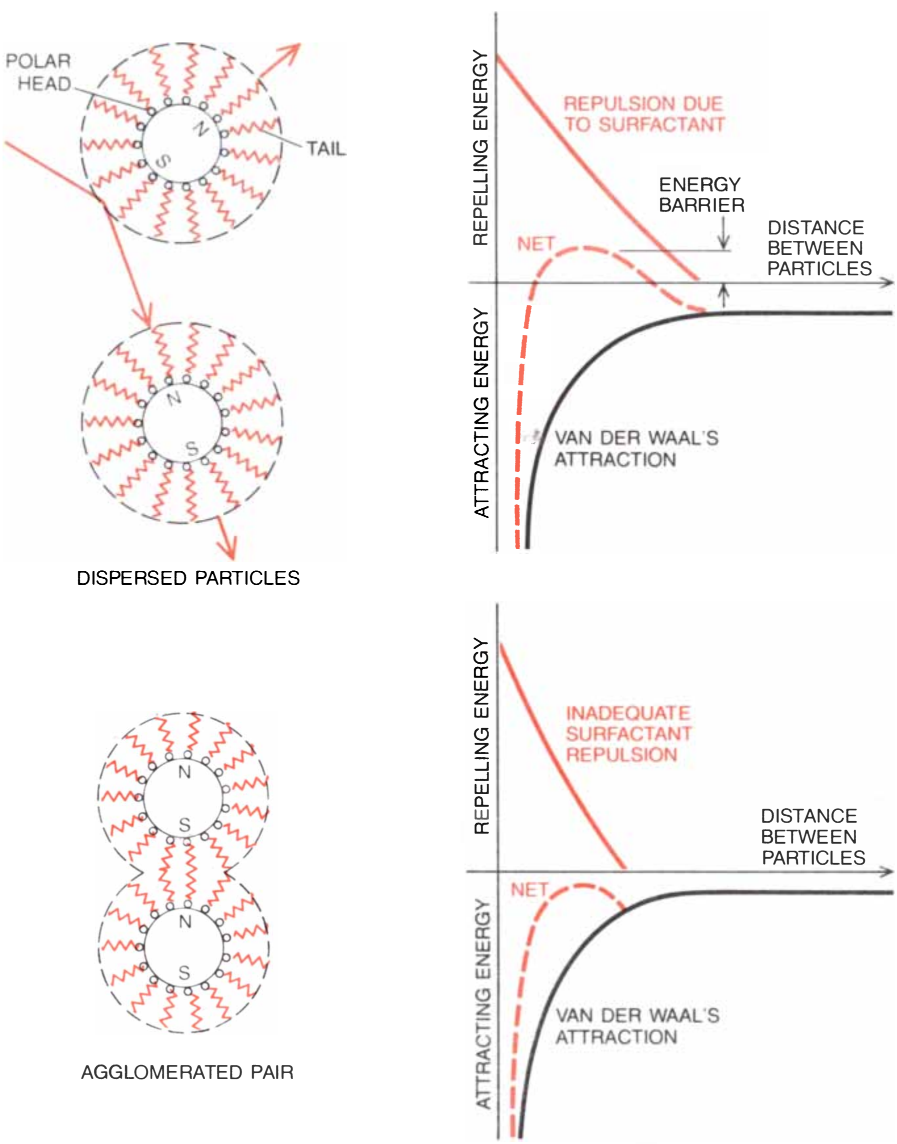
The surfactant molecule has a polar chemical group, or head, that adheres to the surface of the particle and a tail at least 10 to 20 angstroms long that has an affinity for the carrier fluid. The tails are in constant wagging motion, creating a repulsion that tends to separate the particles where the projecting tails overlap. Oleic acid, which stabilizes a suspension of magnetite particles in a kerosene carrier, is an organic molecule whose tail is a chain of 18 carbon atoms and whose head is a polar carboxyl group (COOH). Apparently an important structural requirement for an effective surfactant is that the tail have either a bulge or a kink, which seems to inhibit neighboring molecules from nesting together and forming crystals. In oleic acid a kink is supplied by a double bond in the middle of the 18-carbon chain. Stearic acid, which differs from oleic acid only in that it lacks the double bond, is ineffective as a surfactant.

Polymers that have several head groups with an affinity for magnetic particles are the most tenacious stabilizers. If one head group becomes detached from the particle, the chances are that others will hang on, thereby preserving the integrity of the elastic layer. Depending on the choice of surfactant and suspension medium, ferrofluids have densities beginning at about the density of water and ranging up to twice that value.

The energy relations that determine the properties of a colloid are known at least approximately. The elastic bumpers that enclose a colloid particle give rise to a repulsive force provided the distance between the surfaces of two adjacent particles is somewhat less than twice the thickness of a single bumper layer. The particles can approach closer than this distance only if their thermal energy is sufficient to overcome the repulsion of the surfactant layers. At ordinary temperatures the surfactant layer ensures the long-term stability of the colloid.

One of the most important properties of a stable ferrofluid is the way it responds to an applied magnetizing field. I mentioned above that the magnetite particles in a ferrofluid are smaller than a single magnetic domain and are therefore fully saturated spontaneously. Nevertheless, when a magnetic field of low to moderate strength is applied to a ferrofluid, not all the particles immediately fall into alignment with the imposed field. The field must be increased considerably before full saturation, or maximum magnetic alignment of particles, is achieved.

The reason is that random thermal agitation keeps knocking the magnetic moments out of alignment. There are two possible mechanisms that might be responsible for this effect: either the magnetic moment is locked to the crystal structure of the particle and the particle



STABILITY OF A FERROFLUID depends on making the magnetic particles very small and on preventing them from agglomerating. The particles are about 100 angstrom units in diameter and are kept from settling by the random thermal motion of molecules in the surrounding liquid. The particles are kept from sticking together by coating them with a substance called a surfactant in a layer one molecule thick. Although each particle resembles a small bar magnet, magnetic attraction turns out to be negligible for very small particles. The major attractive force is the van der Waals force, which arises between any particles having fluctuating electric charges that are not uniformly distributed. Because the van der Waals force decreases rapidly with distance it can be overcome by surrounding each particle with chainlike surfactant molecules, which serve as bumpers. Each molecule of a surfactant such as oleic acid has a polar "head" that can be adsorbed on the particle surface and a "tail" with an affinity for the surrounding liquid (*upper left*). The bumpers give rise to a force of repulsion and hence to an energy barrier that can keep the particles separate if the speed with which they approach each other is not too great (*curves at upper right*). If the bumper molecules are too short or have an unsatisfactory configuration, the particles can agglomerate (*diagram and curves at bottom*).

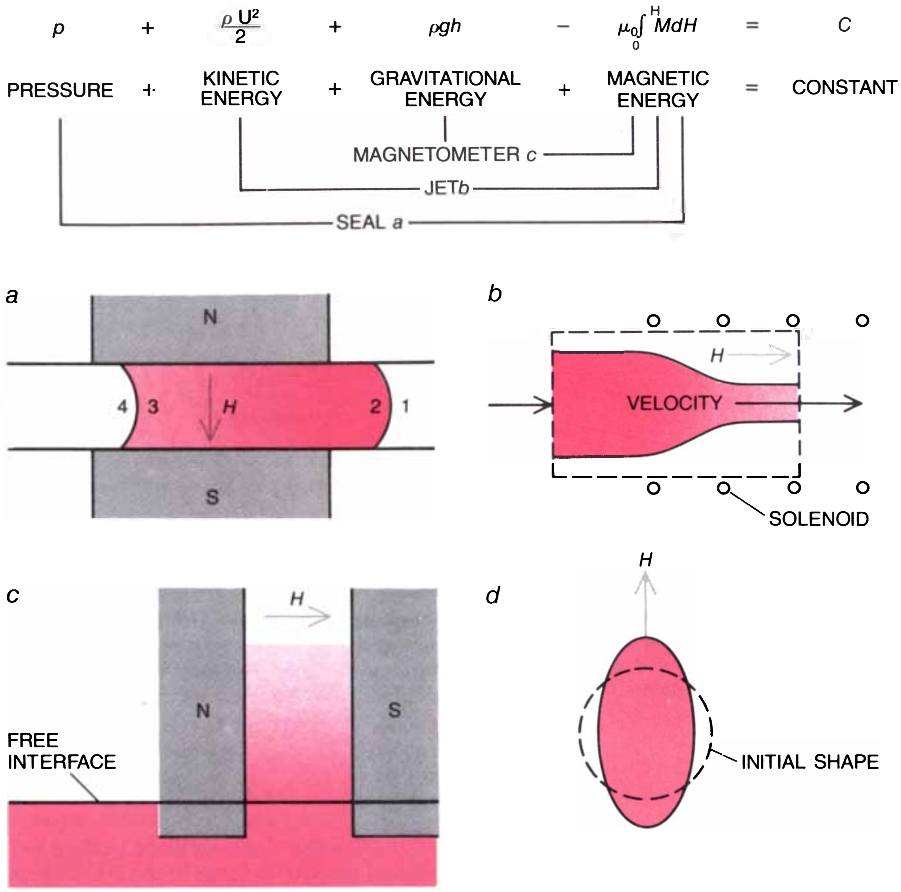
rotates with the moment, or the direction of the magnetic moment rotates with respect to the crystal structure. The difference between the two mechanisms is of no practical importance because their statistical description is the same. The response of the particles to an imposed field is equivalent to the phenomenon of paramagnetism observed in the polarization of molecules that have a dipole moment. Since a ferrofluid particle is much larger than a single molecule, the moment per particle is also much larger and the response is known as superparamagnetism. A similar form of magnetism is charactistic of the particles used for magnetic recording, in geological rock specimens and in finely divided ferromagnetic catalysts.

One of the macroscopic properties of ferrofluids that I believe I was the first to observe is an unexpected kind of levitation. If one were to place a nonmagnetic sphere such as a glass marble in a sealed vessel holding a ferrofluid whose specific gravity is lower than that of the sphere, one would expect the sphere to sink to the bottom, and it does. If the vessel is placed between the like poles of two magnets of equal strength, one directly above the ferrofluid and one below it, the sphere rises from the bottom of the vessel and moves toward the center of the fluid. There it remains floating as long as the field is applied.

What is the explanation? Since the opposing magnets are equally strong, the field in the center of the fluid must be zero and it must increase symmetrically outward. As I showed above, however, the magnetic energy of the fluid is inversely proportional to field strength, which means the magnetic energy is highest in the center of the fluid and decreases steadily with distance from the center. According to the ferrohydrodynamic Bernoulli equation, the sum of the magnetic energy and the pressure must be constant everywhere in the fluid, and so the pressure must be lowest in the center and increase with distance as the magnetic energy decreases. Hence the glass sphere experiences nonuniform pressure forces that tend to raise it from the bottom of the vessel, propel it to the center of the fluid and hold it there in stable equilibrium.

In a related discovery I found that the levitation system can be turned inside out, so to speak. The source of the external magnetic field can be the levitated object itself. A ferrofluid will levitate a solid magnet that is almost four times as dense as the fluid. Recent applications of ferrofluid buoyancy include the sorting of diamonds from beach sand by selective flotation and the guidance of a petroleum drilling bit with an underground accelerometer in which the sensitive mass is levitated in a ferrofluid.

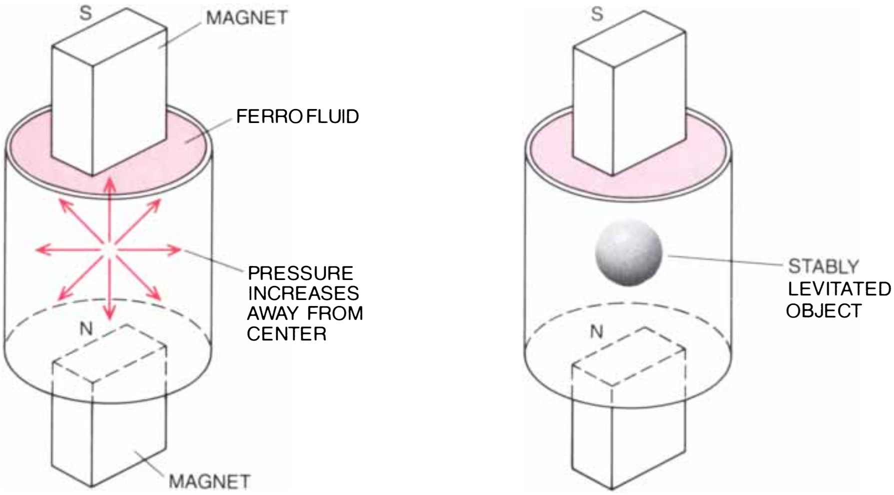
The flowing of an ordinary fluid can become unstable in a number of ways. At low speed, for example, a liquid flows smoothly in a tube, but at a critical velocity the smooth laminar flow becomes turbulent. Water flows smoothly past a sphere at low speed, but as the velocity is increased eddies are shed periodically downstream. Magnetic liquids exhibit new instabilities that have not been observed in any other fluids.



BASIC RESPONSES OF A FERROFLUID to forces of pressure, motion and gravity in the presence of a magnetic field can be understood by expanding the fluid-dynamics equation of Daniel Bernoulli to include magnetism as a fourth force. Bernoulli showed that in an ordinary liquid (disregarding friction) the sum of the energies associated with pressure, motion and gravitation is constant. For a ferrofluid, magnetic energy must be included as a negative term in the equation, and it is the sum of the four quantities that is constant (*top*). Regions of the ferrofluid with higher magnetic energy are shown in darker color. The augmented equation predicts how a plug of ferrofluid in a magnetic field (*a*) responds to a change in pressure if kinetic energy and gravitational energy are held constant. When the pressure in region 4 is higher than that in region 1, the ferrofluid plug is pushed to the right until the pressure in region 3 matches that in region 4, with a compensating loss of magnetic energy; the pressure in region 2 falls until it matches the pressure in region 1, with a compensating gain in magnetic energy. The coupling of kinetic energy and magnetic energy can be seen in a free-flowing horizontal ferrofluid jet (*b*). Work done on the fluid by the magnetic field reduces both the cross section of the stream and the fluid's content of magnetic energy while increasing its velocity. There is no change in the fluid's internal pressure or its gravitational energy. The coupling of gravitational and magnetic energy forms a magnetometer (*c*). A change in the intensity of the field generated by a magnet alters the height of the ferrofluid between the poles of the magnet. When a ferrofluid droplet is exposed to a uniform field, the droplet responds to unequal surface stresses by elongating (*d*), an effect that is consistent with the expanded equation but is not predicted by it.

The most characteristic instability of a ferrofluid arises even when the fluid is motionless, provided it is capable of accepting sufficient magnetization. One exposes a pool of ferrofluid to a magnetic field oriented perpendicular to the surface. As the intensity of the field is raised the flat surface suddenly breaks up into a hexagonal pattern of spikes. My colleagues and I first observed the phenomenon soon after we had succeeded in preparing a ferrofluid 10 times more magnetic than those originally available.

M. D. Cowley of the University of Cambridge and I have done a mathematical analysis that predicts the onset of the instability and the spacing of the pattern. According to our findings, small random waves that are always present on the surface become amplified because they focus the magnetic field and with it the magnetic stress to which they themselves are subjected. When the magnetic force exceeds the restor-



LEVITATION OF A NONMAGNETIC OBJECT, an effect discovered by the author, is another demonstration of the expanded Bernoulli relation. When a container filled with a ferrofluid is placed between the like poles of two bar magnets of equal strength (*left*), the magnetic energy is highest at the center of the fluid and decreases symmetrically outward. Since the sum of the magnetic energy and the pressure must be constant everywhere, the pressure is lowest at the center and increases with distance from it. When a nonmagnetic object such as a glass marble is placed in the container (*right*), it moves to the center and remains there in equilibrium.

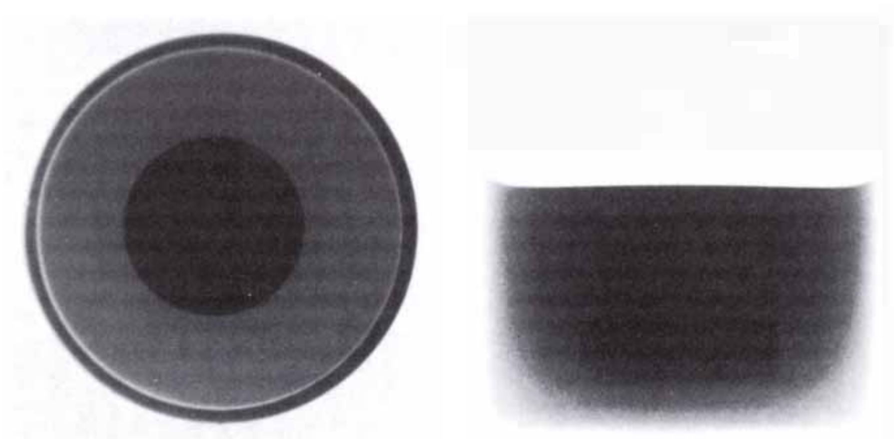
ing forces of surface tension and gravity, numerous small peaks begin to appear on the surface. The peaks that grow fastest either suppress or subsume their neighbors, so that the surviving peaks tend to be uniformly spaced at the maximum possible distance. The spacing is most efficiently accommodated in a hexagonal pattern.

Another instability can be seen when a magnetic fluid and a transparent immiscible fluid are sandwiched between two closely spaced glass plates. In the absence of an imposed field the interface is a horizontal line with the denser magnetic fluid resting stably below. When a magnetic field is applied perpendicular to the plates, however, the interface develops a comblike appearance as fingers of ferrofluid begin pushing upward, leaving gaps for the nonmagnetic fluid to move downward. In a matter of seconds a complex labyrinthine pattern develops, similar in form to the transitions between magnetic and nonmagnetic regions in films of certain superconductors and also to the wiggly domain patterns that can appear in ferromagnetic garnets used in magnetic-bubble-memory devices. At the Massachusetts Institute of Technology, James R. Melcher and his students have studied a number of systems in which a tangentially oriented magnetic field has the opposite effect: it suppresses instabilities that would otherwise develop in a ferrofluid.

One striking ferrofluid instability has been investigated in the U.S.S.R. by B. M. Berkovsky and V. G. Bashtovi of the A. V. Luikov Heat and Transfer Institute in Minsk. The experiment calls for placing a ferrofluid in a liquid bath of the same density and introducing a horizontal current-carrying wire. At a sufficiently high current (and hence magnetic field) the ferrofluid is attracted to the wire and forms a cylinder around it. As the current is reduced again the cylinder breaks up into a number of evenly spaced bulges resembling beads on a string. The instability is a well-known one in which a disturbance conserves the volume of a fluid cylinder while decreasing its surface area; if the process is continued, the cylinder breaks up into droplets. The phenomenon can be seen in a jet of water emerging from a faucet: the water emerges as a smooth cylinder, which then becomes wavy and ultimately disintegrates. In a ferrofluid the instability is controlled by the ratio of the magnetic energy to the surface-tension energy. The ability to establish, control and even prevent a variety of fluid instabilities by magnetic means suggests a number of potential applications that are under investigation.

One further peculiarity of the flow of a ferrofluid can be demonstrated by considering what happens when a rotary motion is imparted to the ferrofluid in the presence of a magnetic field. Suppose an ordinary fluid and a ferrofluid are placed on a rotating platform. In the nonmagnetic fluid the angular momentum induced by rotation varies from one point to another, but at any one point it has a single value, or component. The situation can be analyzed by considering the shear stress—the force that tends to make one parcel of fluid slide past an adjacent one—on a small cubical volume of the fluid. As a consequence of the law of conservation of momentum the opposing components of the shear stress acting on the cube must be equal on the four sides perpendicular to the plane of rotation.

When the rotating fluid is a ferrofluid polarized by a nonrotating horizontal magnetic field, the situation is quite different. As the magnetic particles are carried around by the rotation of the platform they try to stay aligned with the field, but because of friction there is always some angle of lag between the particle orientation and the direction of the field. Furthermore, since each particle rotates in trying to maintain its alignment, it has an "internal" angular momentum that is distinct from the "external" angular momentum of the surrounding liquid carrier. As a result the opposing components of surface shear on the cube of ferrofluid are no longer equal; the stress then includes elements that are described as antisymmetrical. Because the rotating magnetized particles can diffuse through the fluid they



SELF-LEVITATION OF A MAGNETIC OBJECT represents an inversion of the phenomenon shown in the preceding illustration. The two X-ray images show a disk magnet stably suspending itself in a beaker of ferrofluid. In the side view the magnet, which is almost four times as dense as the fluid, is seen hovering above the bottom of the beaker. The phenomenon, discovered by the author, has been applied commercially in low-friction bearings for instruments.

can transport angular momentum to neighboring volumes of the fluid. Thus the rotation of the individual particles can become coupled to the motion of the fluid as a whole.

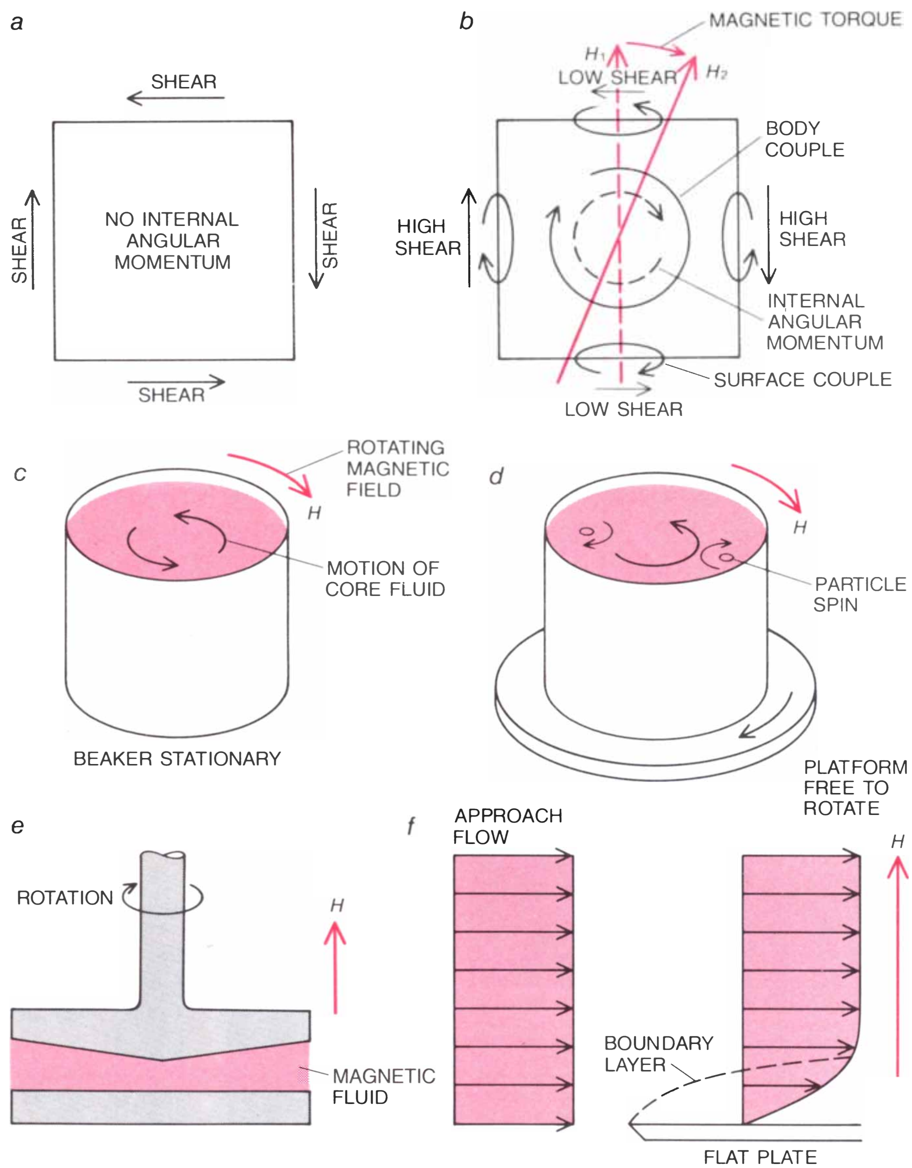
What I have just described is the response of a rotating ferrofluid to a nonrotating field. In our early experiments at Avco my colleague Ronald Moskowitz and I were curious to see what would happen if the ferrofluid in a stationary beaker were subjected to a rotating magnetic field. A rotating field is easily generated with an arrangement of coils such as the one in an alternating-current electric motor. We found that the ferrofluid rotates vigorously as the particles try to remain oriented to the field. In a field rotating at 1,000 revolutions per second the fluid is soon swirling at 200 revolutions per minute.

Two British investigators, R. Brown of the University of Sheffield and T. S. Horsnell of the British Steel Corporation, repeated the experiment with the beaker resting on a turntable that was free to rotate. They were astonished to see the fluid rotating in one direction and the beaker and turntable in the other. Moreover, the fluid's direction of rotation was opposite to that of the field! The phenomenon has not yet been satisfactorily explained, but it is probably associated with the boundary conditions induced by antisymmetrical stress. The existence of antisymmetrical stress is not taken into account in the ferrohydrodynamic Bernoulli equation. In most cases the effects of antisymmetrical stress can be neglected, but it is apparent that such stress can have major consequences in certain circumstances.

Commercial applications of magnetic fluids have been pioneered by the Ferrofluidics Corporation of Nashua, N.H., a firm Moskowitz and I founded in 1968. I have already mentioned that the earliest application was in seals that prevent leakage in the gap that is always present between a rotating shaft and its supporting structure. The seal needs only a few drops of a ferrofluid in the gap between the shaft and a cylindrical permanent magnet that forms a collar around it. The fluid creates an impermeable ring around the shaft while allowing it to rotate with negligible friction. Seals of this kind provide a barrier between the atmosphere and the internal vacuum of furnaces for growing silicon crystals. They have also been applied to the sealing of gas lasers, motors, blowers and many other devices. In the disk memory units of a digital computer the "heads" for reading and writing information hover just millionths of an inch over a spinning disk. A particle of airborne dust that lodges between the head and the disk can cause a head "crash," destroying the recording surface. Wear particles from the disk's own bearings can also lead to a crash. A ferrofluid seal can eliminate both sources of contamination.

In order to cope with higher pressure differentials seals with multiple stages have been designed. When the seal is pressurized, each stage momentarily develops a bubbly leak that pressurizes the interstage chambers. The total pressure that can be contained is the sum of the pressure differences across all the stages. One rotary seal with 160 stages has withstood a pressure differential of 66 atmospheres, or nearly 1,000 pounds per square inch.

A variant of the ferrofluid pressure seal has proved useful in the design of loudspeakers. The major working element of most loudspeakers is a cylindrical voice coil fitted to a cylindrical permanent magnet with a small gap that allows the coil to move. The heat developed in the voice coil can be dissipated more readily when the air in the gap is



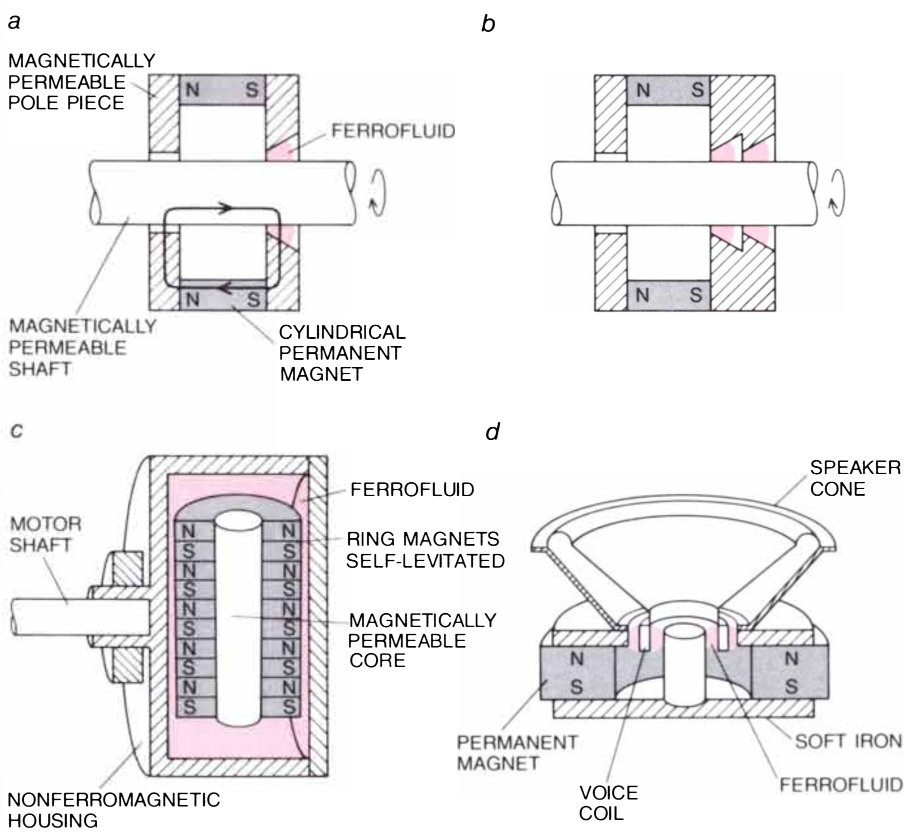
ANTISYMMETRICAL STRESSES are set up in a ferrofluid when an imposed magnetic field is shifted from one orientation to another. When a nonmagnetic fluid is suddenly rotated, a cubic parcel of fluid (*a*) experiences shear stresses that are symmetrical. When a ferrofluid is rotated with respect to a magnetic field (*b*), the magnetic particles experience a torque that sets up unbalanced shear stresses. The change in relative field direction creates a body couple, or distribution of torque, within small parcels of fluid. Because the body couple varies from point to point surface couples are also generated that transport angular momentum throughout the fluid. If a rotating magnetic field is created around a ferrofluid in a beaker (*c*), the bulk fluid, in response to antisymmetrical stresses, rotates counter to the field. If the beaker is placed on a freely rotating platform (*d*), the fluid and the beaker rotate in opposite directions. If a ferrofluid is placed between a stationary lower disk and a rotating upper disk and subjected to a vertical magnetic field (*e*), the apparent viscosity of the ferrofluid increases, making it more difficult to turn the upper disk. If a vertical magnetic field is imposed on a ferrofluid flowing across a stationary flat plate (*f*), the velocity of the fluid in the boundary layer near the plate is reduced not only by friction but also by a magnetic coupling between the fluid and the applied field.

replaced by a liquid. An ordinary liquid would drip out of the gap, but a ferrofluid is retained by the magnetic field already present. There is the promising possibility of designing acoustic sources in which sounds are produced directly by vibrations of magnetically excited ferrofluids.

Magnetic levitation in a ferrofluid is the basis of several processes for the separation of materials that differ in density. Ordinary methods of separation by flotation rely on heavy liquids or slurries, but they are incapable of floating substances with a specific gravity higher than 5 or so. Moreover, many of the liquids are quite toxic. The magnetic-levitation forces that can be established in a ferrofluid are strong enough to float a material of any density. The one limitation is that the material itself must be nonmagnetic.

In the U.S. a ferrofluid pilot plant has been designed to separate industrial scrap metals, including nonferrous scrap from automobiles. The Bureau of Mines has demonstrated that ferrofluids can separate various metals left from the incineration of solid wastes. In Japan, Hitachi Ltd. has sorted components of household electrical appliances made out of aluminum, zinc and copper. It is reported that paramagnetic salt solutions are being used in the U.S.S.R. for laboratory mineralogical analysis and that a similar process is sorting diamonds from rock in South Africa. Ferrofluids have also found application in ink-jet printing, in the control of alphanumeric displays and in the detection of magnetic domains and other metallurgical structures.

One important class of ferromagnetic phenomena has not yet been exploited commercially in ferrofluids, namely the change in magnetic strength that accompanies a change in temperature. As the Curie point is approached a small increase in temperature causes a large decrease in magnetization. The reason is that thermal agitation disrupts the parallel alignment of the individual magnetic dipole moments within each magnetic domain. Magnetism and temperature are connected in another way as well. Because temperature is a measure of disorder a sample of material



APPLICATIONS OF FERROFLUIDS have grown steadily since 1968. The first useful devices (*a*) were wear-free seals for rotary shafts in which a liquid ring of ferrofluid is gripped in a focused magnetic field of high intensity. The seals prevent leakage or mixing of gases. Multiple sealing stages (*b*) can withstand a pressure difference of as much as 50 atmospheres per 10 centimeters of shaft length. Another device (*c*) damps unwanted oscillations in systems such as numerically controlled machine tools or graphics plotters, where smooth incremental motion is essential. The damping is done by an inertial mass composed of magnets self-levitated in a ferrofluid sealed in a chamber that rotates with the shaft. In a loudspeaker (*d*) a ferrofluid is held in place by the ring magnet that surrounds the voice coil. The ferrofluid not only increases power handling by dissipating heat but also flattens the frequency response of the loudspeaker.

that is thermally insulated from its surroundings must become cooler when it is magnetized. Conversely, a sample that is magnetized while it is held at a constant temperature must transfer heat to its surroundings as it becomes more ordered magnetically.

A "magnetocaloric" engine that employs a magnetic colloid as a working fluid has been analyzed by E. L. Resler, Jr., of Cornell University and me. The engine consists of a magnet, a heat source, a heat sink and a loop of tubing filled with ferrofluid. Once the fluid is set in motion it circulates through the loop transferring heat from the source to the sink, with the capacity for doing work. As the fluid leaves the heat sink it is attracted to the magnet, where its temperature is raised by the heat source, which lies within the magnet's field. With rising temperature the fluid's magnetization diminishes and it leaves the region of the magnetic field. The heat sink cools the fluid to its original temperature in preparation for another cycle. The engine could be designed so that the fluid would drive a turbine or pass through a magnetohydrodynamic generator for direct production of electricity. It can be shown that the thermal efficiency of the ferrofluid engine should approach that of the Carnot cycle, a limit no heat engine can exceed. It is possible the ferrofluid cycle will find application in pumps with few moving parts, in cooling devices and even in the production of electric power from sources of low-grade heat. A solar heating system with a magnetic working fluid could drive its own circulation.

Many industrial devices rely on thermal convection for the transfer of heat, and here too ferrofluids may find a new and interesting role. Convection takes place whenever a fluid is heated nonuniformly, typically when an enclosed fluid is heated at the bottom and cooled at the top. The heated lower layers become less dense than the cooler upper layers, which is an unstable configuration. When the temperature difference between the bottom and the top exceeds a critical value, an organized pattern of fluid circulation develops as heavier and lighter regions of the fluid respond to gravitation.

When a gradient magnetic field (stronger at the bottom of the vessel than at the top) is applied to a heated ferrofluid, it induces convection that can be much more vigorous than simple gravity convection. The circulation is driven by changes in the magnetization of the fluid with changes in temperature rather than by changes in density. The effect may prove useful in removing heat from motors, transformers and other electrical equipment where magnetic fields are present. I am confident that ferrofluids will find many other applications no one has yet thought of.

Tidal Bores

A bore is the hydraulic analogue of a sonic boom: a moving wall of water that carries the tide up some rivers that empty into the sea. For several hours after it passes the river flows upstream

by David K. Lynch

We huddled together on the bank of the river in the cool spring night, drinking tea and warming our hands at the fire. The full moon shone intermittently through the clouds, illuminating the river below. Suddenly from downstream came the cry "Flood—O!" We ran higher up the bank with a low, rumbling sound in our ears. As we reached the top of the bank a wall of water four feet high passed below us, moving faster than we could run. The wave was moving upstream, carrying with it pieces of wood and other debris. In the few minutes it took for the wave to move out of range of our hearing the water level rose another two feet. Gradually the turbulence subsided into gentle eddies, and after an hour the river began flowing toward the sea again.

Thus in 1978 did I get my first sight of the tidal bore on the River Severn in southwestern Britain. A tidal bore is a remarkable hydrodynamic phenomenon: it is the incoming tide in the form of a wave going up a river that empties into the sea. After the bore passes, the river flows upstream for some time before the current reverses again. For a river to have a bore two conditions must be met. The tides in the adjoining tidal body must be exceptionally high; a difference of more than 20 feet between high and low water is generally required. In addition the river must be shallow with a gently sloping bottom and a broad, funnel-shaped estuary. As the incoming tide is forced into the narrowing river mouth it builds in height until a single wave materializes and moves upstream.

A bore is a member of the class of solitary waves. Such waves appear in many forms and in many mediums besides water; they are even observed in the atmosphere of Mars. Their most intriguing property is that they do not disperse. Like all waves, a tidal bore is made up of many sinusoidal waves with many different wavelengths. In deep water the component waves move with different speeds and separate from one another. Because of the way shallow-water waves interact with the bottom in the estuary, the component waves of the bore move at the same speed. Hence instead of dispersing they travel upriver together as a single wall of water. Such solitary waves are curious and attractive for their own sake, but they are of particular current interest because they represent the solution to problems in several scientific disciplines.

Bores are elusive. They vary from river to river and even from tide to tide. Moreover, they have not been much studied in the 20th century. For these reasons any generalization about bores must be carefully qualified. As we have seen, a bore is a solitary wave that carries a tide upstream. The bore can be as low as a few inches or as high as 25 feet. Most bores are higher near the banks than they are in mid-channel. The formation of the bore represents the turn of the tide. Bores are largest and most powerful around the time of the new moon or the full moon, when the tides are highest.

Since the river flows upstream after the bore passes, the wave front divides the river into parts with opposite directions of flow. It should be noted that the river does not flow uphill as the bore passes. Since the water level behind the bore is higher than the level in front of it, the water flows upstream but downhill.

As the bore moves upriver it may have a breaking crest like the crest of a wave on a shelving beach or it may be a smooth, gently rounded wave. A breaking bore is a foaming wall of turbulent water that advances noisily upstream. Bores that do not break are called undular bores. Several smaller waves usually follow an undular bore. Along the River Trent in Britain the smaller waves are known as whelps; along the Seine they are known as *etueles,* or stubble. I have seen as many as 55 of the smaller waves following a bore on one of the tributaries of the Amazon.

Whether a particular bore is breaking or undular depends on the height of the tide, the depth of the river, the shape of the bottom and the speed and the direction of the wind. Many bores are undular in mid-channel and breaking on the bank. In some rivers the bore can be undular on one tide and breaking on the next. A bore can even change from one form to the other as the depth of the river changes.

Bores are distributed widely over the globe. They are not concentrated in any one region, although they do tend to form in places where the tides are greatest, such as on the coasts of Britain and France, in rivers flowing into the Indian Ocean and at the mouth of the Amazon. Whether a bore forms is determined by the local conditions at the mouth of a river. Since little attention has been paid to bores in recent decades, precisely how many rivers have bores is not known. With Susan B. Bartsch-Winkler of the U.S. Geological Survey I am compiling a catalogue of the major bores of the world.

North America has several notable tidal bores. The most striking is the one on the Petitcodiac River in Nova Scotia, which empties into the Bay of Fundy. There are also dramatic bores on Turnagain Arm and Knik Arm, which flow into Cook Inlet near Anchorage, Alaska. On Turnagain Arm and Knik Arm bald eagles hunt fish by following the bores moving inland from the Pacific; the fish are brought to the surface by the turbulence of the bore.

Perhaps the most celebrated of all bores is the one on the lower Seine known as the Mascaret. The Mascaret has been greatly reduced by the silting of the Seine near its mouth, but the bore was once notorious for its destructiveness. In 1843 Victor Hugo's daughter Léopoldine and her husband Charles Vacquerie were boating near Rouen when the Mascaret rolled upriver. The bore knocked Léopoldine out of the boat. Charles jumped in to save her, but neither of them could swim and both drowned. Hugo memorialized them in the poem "À Villequier" but did not mention the bore.

Certain tidal bores are associated with myths or ceremonies. The passage at

BORE MOVES UP THE SEINE through the town of Caudebec in a photograph made in 1963. Since that time harbor-control measures at Le Havre have greatly reduced the size of the bore. Known as the Mascaret, the Seine bore was once feared for its destructive force.

BORE MOVES UP THE ARAGUARÍ, a river in the Amazon basin, in this aerial photograph made by the author. Known as the pororoca, the bore of the Amazon and its tributaries reaches a height of 25 feet. Here it is about five feet high and moving at about 10 miles per hour. Bores can be breaking, like the surf on a beach, or smooth and rounded; smooth bores are called undular. Many bores are undular in mid-channel and breaking near the bank where the shallow water slows the wave. The pororoca shows such a configuration here. Immediately in front of an undular bore the water is smooth, but a series of smaller crests often follow the main wave up the river, as they do here.

BORE ON TURNAGAIN ARM, near Anchorage, Alaska, is one of the few in North America. It is usually from two to five feet high and, as shown, its leading edge is not straight or uniform. The depth of Turnagain Arm varies greatly. The parts of the bore in deep water move faster than those in shallow water. The faster parts are undular and the slower parts breaking.

BORE ON THE SEVERN is the largest in Britain. It is a powerful breaking bore with a crest large enough for surfers to ride upstream for miles. The bore forms near the Severn bridge above Sharpness and extends to Gloucester. Two conditions must be met for a bore to form. First, there must be a broad estuary that narrows toward the river mouth and has a shallow, gently sloping bottom. Such a configuration funnels the incoming tide, increasing its height. The Severn estuary from Cardiff to Bristol has the required narrowing shape. Second, the tides in the adjoining tidal basin must be very high. A difference of more than 30 feet between high water and low water at Sharpness is generally necessary for the Severn bore to form.

the beginning of this article describes my experience at the annual celebration held in the spring on the Severn near Gloucester. The transparent baby eels, known locally as elvers, that live on the bottom of the Severn are the traditional Good Friday meal. When the bore comes upriver, it disturbs the eels. When the Severn is flowing downstream again, the inhabitants of Gloucester gather the eels with nets. Those that are not eaten or taken to the local market are sold to merchants who fly them to eel farms, many in the Orient. By day the Severn bore is caught by surfers, who ride it upstream for miles.

The broad Amazon basin has a bore known as the pororoca that sweeps up the river and its tributaries. The pororoca can be 25 feet high and many miles across. Its size and force render the Amazon completely unnavigable during the bore's passage. Because of the Amazon's gradual slope, tidal effects can be detected 500 miles upstream from the mouth of the river.

The mightiest tidal bore of all is the one on the Tsientang River in China. The Tsientang bore reaches a speed of about 15 knots and was once exploited by local fishermen as a means of propulsion. The bore is largest near Hangchow, which was at one time an imperial capital. According to a Chinese myth, the Tsientang bore was unleashed on Hangchow to punish an emperor who had arranged the assassination of a popular general who was a potential rival. When ceremonial paper burnings and offerings of food failed to conciliate the angry spirits, the emperor had a large pagoda put up as a grander propitiatory gesture. The Bhota pagoda still stands near Hangchow, and although the bore did not disappear after the building of the pagoda, it is said to have become somewhat less violent.

The elusiveness of bores is suggested by the fact that Marco Polo, who spent 18 months in Hangchow in the 13th century and wrote a detailed account of his stay, does not mention the bore. The Tsientang bore is so impressive that it is difficult to believe the Italian adventurer could have seen it and not remarked on it. It is more likely that the bore was much smaller in the 13th century than it is today. Changes in local terrain caused by earthquakes and slower changes in the river bed make some bores intermittent. Contemporaneous accounts of the Taiping rebellion in the 19th century indicate that the Tsientang bore was much smaller a century ago than it is now.

Although the large majority of bores form near the mouth of a river, bores can form under other circumstances. In coastal regions where broad sandy flats lie between the high-tide level and the low-tide level the rising tide can spill onto the flats and rush toward the land as a single well-defined wave. The region of Mont-St.-Michel in France is celebrated for a tide that rushes across the flats faster than a horse can gallop. A tsunami, an ordinary water wave with a long wavelength generated by a seismic movement of the ocean floor, can cause a bore to move up a river that does not ordinarily have one.

Certain other conditions can also cause bores. In February, 1976, a rare combination of a high tide, heavy rain, 115-mile-per-hour winds and low atmospheric pressure sent a tidal surge up the Penobscot River in Maine. At Bangor the river rose 20 feet in 15 minutes and caused the only flood in the history of Bangor that came from downstream. As an example of a bore in a medium other than water, a borelike wave forms in the atmosphere of Mars near Tharsis Ridge during the morning hours in the late spring and early summer. These solitary waves appear as local changes in the refractive properties of the Martian atmosphere over a relatively short distance.

To move beyond the drama visible from the river bank (or on Mars) to an understanding of the physics of bores calls for an appreciation of the tides, of wave motion and of the topography of the mouth of rivers. The tides are themselves due to gravity and inertia. The gravitational force is mainly that of the moon. Since the force of gravity is inversely proportional to the square of the distance between the two objects that are attracted, the moon exerts a stronger pull on the near side of the earth than it does on the far side. The difference between the attraction on the two sides is about 13 percent. The difference tends to elongate the earth along the axis between its center and that of the moon. The inertia of the earth as it orbits the center of the system made up of the earth and the moon tends to stretch the planet out of its spherical shape.

The oceans naturally respond more to such forces than the continents. Hence the oceans flow into two tidal bulges. One bulge faces the moon and the other faces away from the moon; between them the water is shallower. Since the moon takes about 28 days to orbit the earth, the tidal bulges take the same period to travel around the earth as they follow the moon. The daily revolution of the earth on its axis causes each point on the surface of the earth to go into and out of each tidal bulge about once per day. To an observer on the earth the tides appear to rise and fall twice a day. An observer suspended above the earth, however, would see that each point on the earth actually spins into and out of the tidal bulge.

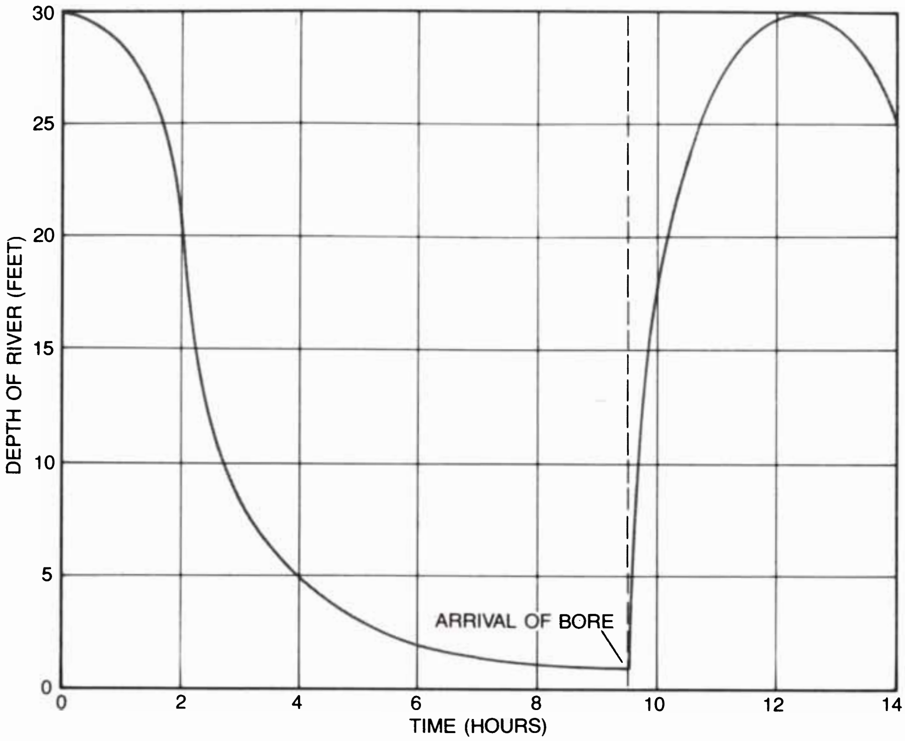
The moon orbits the earth in the same direction in which the earth revolves. By the time a point on the earth has revolved halfway the tidal bulges have moved forward with the moon and the point must move forward to enter the bulge again. Hence there is slightly more than 12 hours between high tides: the period is 12 hours 25 minutes.

If there were no continents to impede the motion of the tides, every point on all coasts would have two tides a day. Many places do have such tides, which are designated semidiurnal. Other places, however, have a different tidal pattern. The wave of the tide is reflected and broken up by the continents and forms a complicated series of crests and troughs thousands of miles apart. Moreover, in some regions the tides are coupled with the motion of large nearby bodies of water. As a result some places have only one tide a day, which is called a diurnal tide. A few places (including Tahiti) have virtually no tide at all because they lie on a tidal node. Such a node is a stationary point about which the standing wave of the tide oscillates.

The sun also raises a tide. Since the sun is so much farther from the earth than the moon is, its tide is only about half the magnitude of the lunar tide. The overall tidal amplitude (the difference between the high-water level and the low-water level) depends on the relation of the solar tide to the lunar tide and therefore on the positions of the sun, the moon and the earth. The tidal amplitude is at its maximum twice a month at the time of the new moon and the full moon, when the three bodies are aligned in the configuration known as syzygy; this is the time of the "spring" tides. The am-

COUNTRY	RIVER OR CHANNEL	TIDAL BODY	LOCATION OF BORE (WHERE KNOWN)
BANGLADESH	GANGES	BAY OF BENGAL	——
BRAZIL	AMAZON	ATLANTIC OCEAN	——
	CAPIM	ATLANTIC OCEAN	CAPIM
	CANAL DO NORTE	ATLANTIC OCEAN	——
	GUAMA	ATLANTIC OCEAN	CAPIM
	TOCANTINS	ATLANTIC OCEAN	——
	ARAGUARÍ	ATLANTIC OCEAN	——
CANADA	PETITCODIAC	BAY OF FUNDY	MONCTON
	SALMON	BAY OF FUNDY	TRURO
CHINA	TSIENTANG	EAST CHINA SEA	HAINING TO HANGCHOW
ENGLAND	SEVERN	BRISTOL CHANNEL	FRAMILODE TO GLOUCESTER
	PARRETT	BRISTOL CHANNEL	BRIDGWATER
	WYE	BRISTOL CHANNEL	——
	MERSEY	IRISH SEA	LIVERPOOL TO WARRINGTON
	DEE	IRISH SEA	——
	TRENT	NORTH SEA	GUNNESS TO GAINSBOROUGH
FRANCE	SEINE	ENGLISH CHANNEL	CAUDEBEC
	ORNE	ENGLISH CHANNEL	——
	COUESNON	GULF OF ST.-MALO	——
	VILAINE	BAY OF BISCAY	——
	LOIRE	BAY OF BISCAY	——
	GIRONDE	BAY OF BISCAY	ÎLES DE MARGAUX
	DORDOGNE	BAY OF BISCAY	LA CAUNE TO BRUNNE
	GARONNE	BAY OF BISCAY	BORDEAUX TO CADILLAC
INDIA	NARMADA	ARABIAN SEA	——
	HOOGHLY	BAY OF BENGAL	HOOGHLY POINT TO CALCUTTA
MEXICO	COLORADO	GULF OF CALIFORNIA	COLORADO DELTA
PAKISTAN	INDUS	ARABIAN SEA	——
SCOTLAND	SOLWAY FIRTH	IRISH SEA	——
	FORTH	IRISH SEA	——
U.S.	TURNAGAIN ARM	COOK INLET	ANCHORAGE TO PORTAGE
	KNIK ARM	COOK INLET	——

CATALOGUE OF TIDAL BORES shows that bores are distributed widely throughout the world. Little systematic attention has been given to tidal bores in the 20th century; therefore the catalogue is incomplete. The author is currently compiling a list of major bores that is expected to include about twice as many as there are in the table. Bores are not concentrated in any one region of the world because the existence of a bore depends on local conditions: the tides and topography at the mouth of the river. Such conditions can be changed by harbor-control measures, damming or natural changes in the river bed. For this reason the extent of the bore given in the table is approximate. The bore on the Colorado River has been greatly reduced by silting near the mouth of the river. The largest of all bores is on the Tsientang in China.



DEPTH OF THE PETITCODIAC shows the sudden increase associated with the arrival of the bore. The curve indicates the depth of the river near Moncton. At the time of low water the Petitcodiac is only a few feet deep. The arrival of the bore immediately increases the depth by from two to three feet. After the bore passes, the river continues to rise for more than two hours, ultimately reaching a depth of about 30 feet. The graph shows the tidal asymmetry that is common on rivers with bores. More than nine hours elapse between high water and low water because the tide goes out gradually with the current. It is only about two and a half hours between low water and high water, however, because the tide comes in suddenly as a bore.

PETITCODIAC BORE goes up the Petitcodiac River through the city of Moncton, New Brunswick. Chignecto Bay is the narrowing body that serves to confine the incoming tide. The adjoining tidal basin is the Bay of Fundy which is known for its great tidal range: the difference between high water and low water can be more than 50 feet. The range is greatest during the spring tides around the time of the new moon and the full moon. Most bores form on the spring tides. The Salmon River has a smaller bore extending from Cobequid Bay to Truro, Nova Scotia.

plitude is at its minimum at the time of the first quarter and the third quarter of the moon, when the sun, the moon and the earth form a right angle and the solar and lunar tides oppose each other; this is the time of the neap tides.

It therefore becomes evident why bores generally form at about the time of the new moon and the full moon: it is then that the tidal amplitude is greatest. The amplitude can be increased further by resonance. Every container of liquid has a natural frequency of oscillation. If the liquid in the container is set in motion, it can move back and forth in a single coherent wave with the natural frequency. On the scale of a coffee cup the oscillation is seen as a sloshing motion. There can also be waves with a number of different frequencies in the motion of the liquid. The lowest frequency, the fundamental, is the most important with respect to resonance.

The amount of time it takes a wave with the fundamental frequency to move from one side of the container to the other and back is called the fundamental period. That period is determined by the depth of the water and the size of the container. For a closed basin the fundamental period is about twice the longest dimension of the basin divided by the speed of the waves. In a lake, which constitutes such a closed basin, large-scale natural oscillations are called seiches. Seiches are common on Lake Michigan. They are often caused by a low-pressure atmospheric system at one end of the lake; on occasion they can be quite destructive.

In a basin open at one end to a larger body of water the fundamental period is about four times the longest dimension divided by the speed of the wave. The tidal basin near the mouth of a river constitutes such a basin, with the ocean providing the tidal impulse. The effect of the natural oscillation is to make the water at one end of the basin high at the beginning and the end of the fundamental period and low in the middle of the period as the water sloshes back and forth in a crest-and-trough pattern. For example, if the fundamental period is 12 hours, the water at one end will be high at, say, noon, low at 6:00 P.M., high at midnight and so on.

If the fundamental period is close to the tidal interval of 12 hours 25 minutes, the basin will resonate with the incoming tide. The incoming tide sets the water of the basin oscillating. The motion of the tide coming in toward the mouth of the river and the motion of the oscillation are synchronized. Thus the oscillation reinforces the tide in the bay and makes the high tide higher than it would otherwise be and the low tide lower than it would otherwise be.

The resonance can also result in an asymmetry between the tide at the ocean end of the basin and the tide near the mouth of the river. In some cases the mouth of the basin is at a node of the wave that represents the oscillatory motion; as we have seen, a node is a fixed point about which the water oscillates. The tide at the node is much smaller than the one near the river mouth.

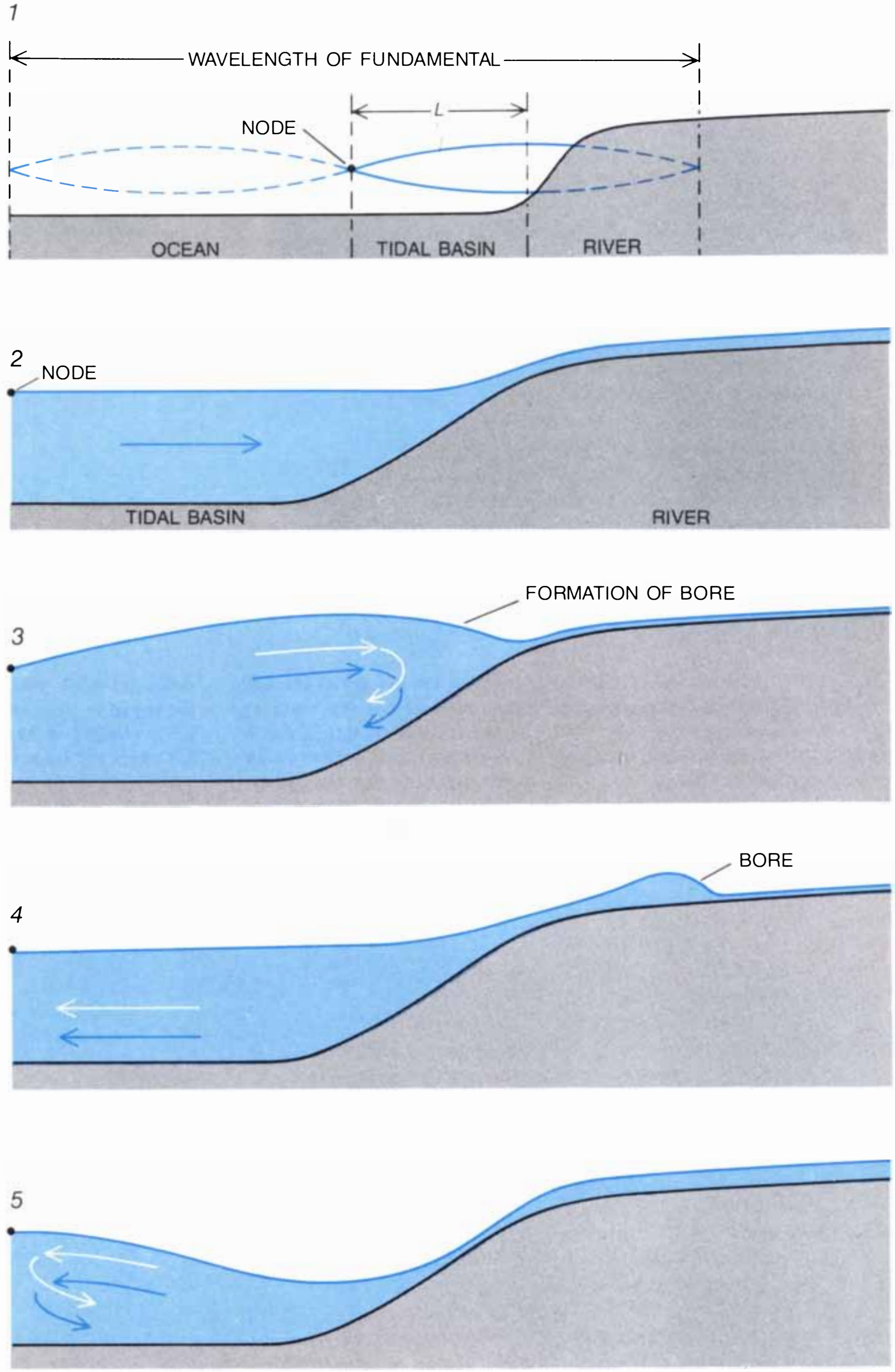
Although any body of water with high tides can generate a bore, about half of the known bores are associated with resonance in a tidal basin. An excellent example of such resonance is in the Bay of Fundy, which is well known for its tidal range of more than 50 feet. The fundamental period of the Bay of Fundy is 11 hours, which is close enough to the semidiurnal tidal interval for the tides to force the bay to resonate and send a tidal bore up the Petitcodiac. Cook Inlet in Alaska resonates so strongly that the 14-foot tide at its mouth is amplified into a tide of more than 30 feet at Anchorage.

The seaward ends of many rivers have a tide. At the mouth of the river the tides are symmetrical: ebb and flood last about six hours each. Upstream the tides become increasingly asymmetric: less time elapses between low water and high water than elapses between high water and low water because the tide comes in quickly but goes out gradually with the current. The presence of a bore tends to exaggerate the asymmetry because the tide comes up the river very rapidly as a solitary wave.

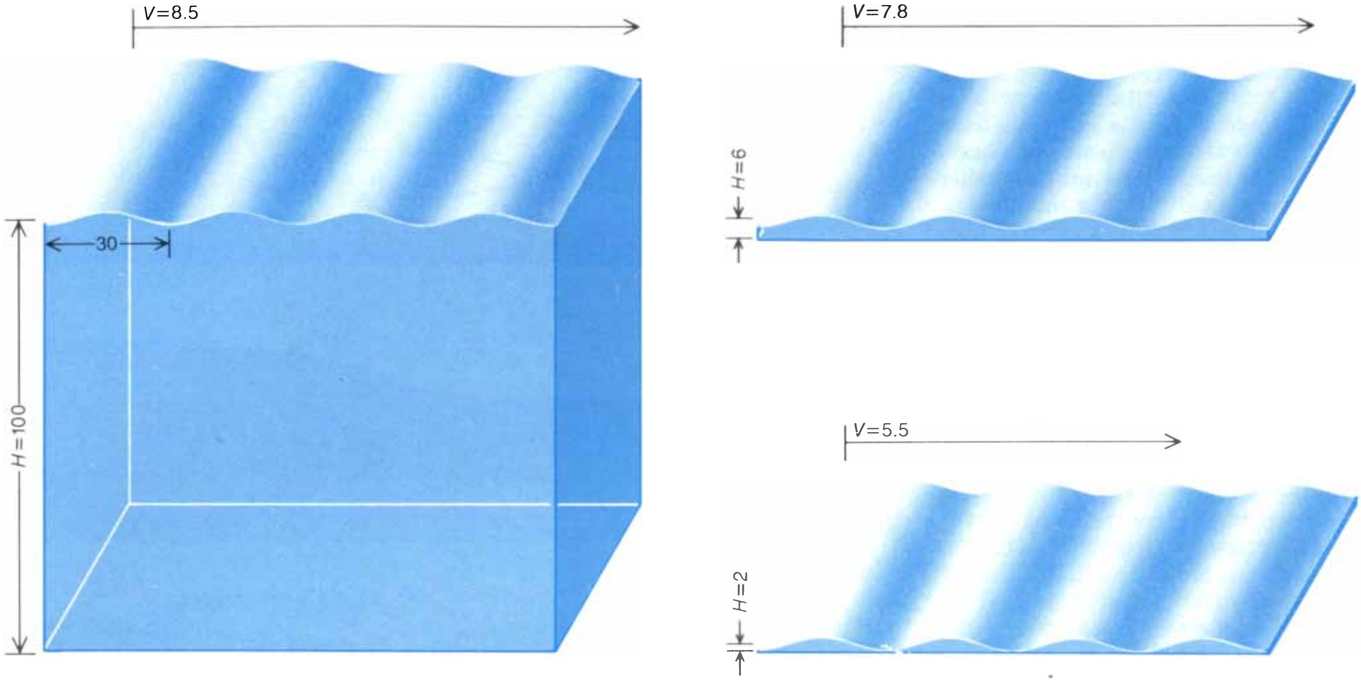
Therefore the tides and their resonance with the oscillation in a tidal basin provide the energy for the tidal bore. The form of the bore, however, is influenced mainly by the dynamics of waves. There are two classes of waves in water: capillary waves and gravity waves. Capillary waves have a wavelength of a few millimeters or less, and the force that restores the surface of the water to its undisturbed state as they pass is surface tension. Such waves are of little interest in connection with tidal bores.

With gravity waves the force that restores the surface to its undisturbed state is gravity. Such waves can be divided into two types on the basis of the relation between the wavelength and the depth of the water through which the wave propagates. Gravity waves with a wavelength much shorter than the depth are known as deep-water waves. Such waves move at a speed that is proportional to the square root of the wavelength. A deep-water wave with a long wavelength moves faster than one with a shorter wavelength.

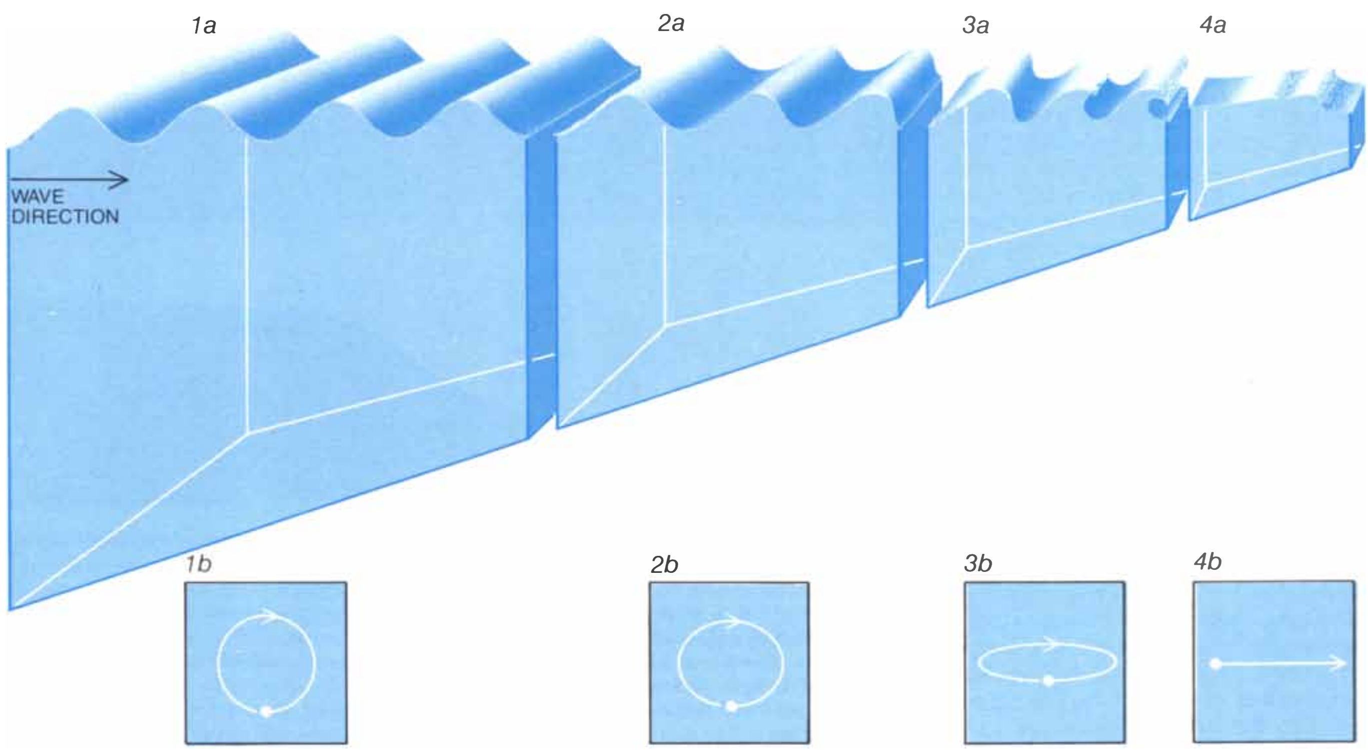
This property of deep-water waves causes them to disperse. Actual waves in water are not simple sinusoidal waves with a single wavelength; they are made up of many sinusoidal components each with a different wavelength. Since each component travels at a different speed, the shape of the deep-water wave is con-



OSCILLATION OF THE WATER in the tidal basin near the mouth of a river can contribute to the formation of the bore by increasing the range of the tide. The diagrams show the tidal basin in schematic form with the ocean at the left and the mouth of the river at the right. If the liquid in a container is disturbed, it will move back and forth in a single coherent wave with a motion similar to that of coffee sloshing in a cup. The time it takes the wave to go from one side of the container to the other and back is the fundamental period. In a tidal basin the fundamental period is roughly equal to four times the length (L) divided by the speed of the wave. The speed is in turn proportional to the square root of the depth. A quarter of the wavelength fits into the basin at one time (*1*). In some instances there is a node, or a fixed point of the wave, at the ocean end of the basin. The tides at the node are small; at the river mouth they are large. If the fundamental period is approximately equal to the 12-hour-25-minute period between high tides, the oscillation will reinforce the high tide near the river mouth. The effect is resonance, and it is shown in panels *2* through *5* of the illustration. The period between panels is one-fourth of the fundamental period. The tide (*blue arrow*) comes into the basin from the ocean as a wave moving toward the mouth of the river (*2*). The incoming tide is synchronized with the oscillation (*white arrow*) of the water in the basin (*3*). The wave propagates through the basin toward the mouth of the river. If the tide rises fast enough, a bore is formed at the river mouth. The tide turns and the sloshing motion reverses its direction. Water is carried toward the ocean end of the basin (*4*). The bore is moving upstream and the tide is still rising upriver. The water level near the river mouth falls as the slosh returns to the ocean end of the basin to join the incoming tide. The cycle is repeated (*5*). About half of the known bores are associated with resonance.



SPEED OF WAVES IN WATER depends on the relation between the wavelength (the distance from trough to trough) and the depth of the water. For deep-water waves (*left*), where the wavelength is much smaller than the depth (H), the speed (V) is proportional to the square root of the wavelength. (In the diagram the height of the wave is assumed to be small compared with the water depth.) A wave with a wavelength of 30 feet in water 100 feet deep moves at a speed of 8.5 miles per hour. When the wavelength is much greater than the depth, the speed is proportional to the square root of the depth. A wave with a wavelength of 30 feet in water six feet deep moves with a speed of 7.8 miles per hour (*top right*). In water two feet deep a wave with the same wavelength will move at 5.5 miles per hour (*bottom right*). Since the speed of shallow-water waves depends only on the depth, waves with different wavelengths have the same speed and do not disperse.



FORMATION OF A BORE, shown schematically, is similar to the development of a breaking wave on a beach. The process by which the bore forms is the conversion of the fast-moving deep-water waves carrying the incoming tide into slower shallow-water waves with a shorter wavelength. The conversion occurs in the estuary. In deep water offshore the waves are symmetrical (*1a*). Both sides of the wave have the same slope and the crests and troughs travel at the same speed. Within each wave the path of a "particle," or small volume, of water is circular (*1b*). Hence although the wave transports energy, no water is actually carried forward as the wave passes. As the waves enter the estuary they begin to interact with the sloping bottom and to move as shallow-water waves. The interaction "compresses" the waves and decreases the wavelength. Moreover, since the crests are deeper than the troughs, they move faster and overtake them (*2a*). The waves become asymmetric, with the advancing side steeper than the following side. In addition the path of the water particles within the wave is flattened into an ellipse (*2b*). When the height of the wave becomes greater than about half the depth, the crest overtakes the trough and breaks into it (*3a*). The path of the water particle has become flatter (*3b*). When the wave breaks, it forms a bore: a turbulent wall of water (*4a*). The particle's path has become linear toward the shore (*4b*). Thus the bore carries water as well as energy upriver.

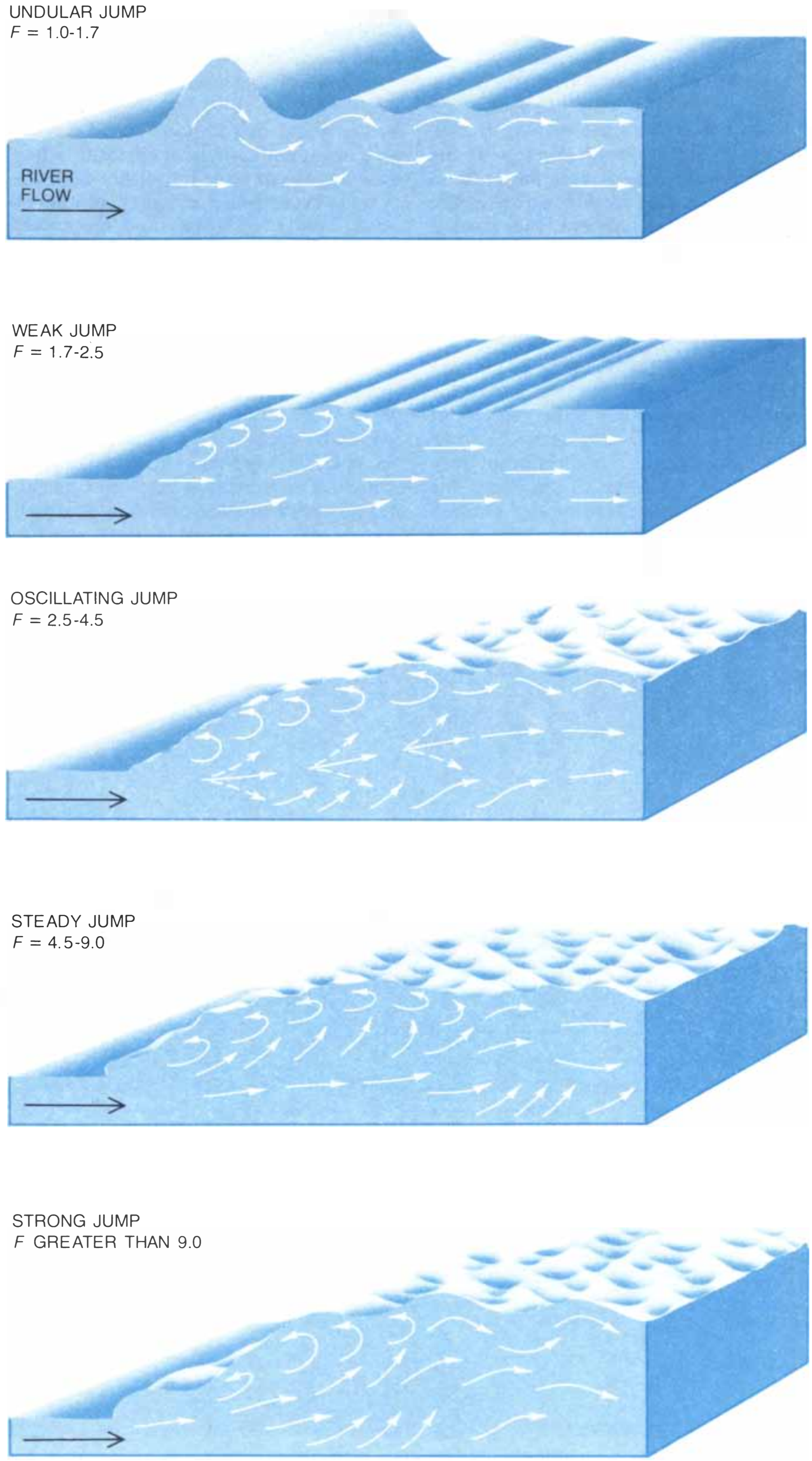
stantly changing as the faster components overtake the slower ones, unite with them and then leave them behind.

Gravity waves with a wavelength much longer than the depth of the water are called shallow-water waves. The speed of a shallow-water wave depends only on the depth of the water; it is proportional to the square root of the depth. Therefore in water of a particular depth all the waves move at the same speed. As a result they do not disperse but travel together with their waveforms relatively undisturbed. It should be kept in mind that deep-water and shallow-water are terms based on the relation between the wavelength and the depth and not on the absolute depth of the water or the wavelength. For example, a wave with a wavelength of three feet would travel across the oceans as a deep-water wave but would move across a bathtub as a shallow-water wave.

The difference between deep-water and shallow-water waves provides much of the basis for understanding the formation of tidal bores. The incoming tide arrives in a tidal basin in the form of rapidly moving waves with a long wavelength. As the waves enter the basin they are confined at the sides and the bottom by the narrowing estuary. As a result of this funneling action the height of the waves increases. The transition of the tide into a bore, which follows the increase in height, is much like the development of breakers from waves moving onto a shelving beach.

The transition from deep-water waves to shallow-water waves is associated with refraction. As the tide comes in and begins to interact with the bottom, refraction slows the waves. Since the offshore waves are in deeper water, they move faster than the waves closer to the shore. As a result, farther inshore the distance between successive crests is reduced. Furthermore, since the crest of each wave is in deeper water than the adjacent troughs are, it begins to overtake the trough in front. The wave, which was previously symmetrical, becomes asymmetric: the leading side gets steeper and the trailing side gets flatter. If the difference in the depth is great enough, the crest will begin to break into the trough and form a bore.

Another significant process develops as the wave moves toward the shore. In deep water the waves transport energy but do not transport the water itself. Each "particle," or small volume of water, at or near the surface describes a circular course and returns to its original position as the wave passes. When tidal waves advance toward the shore, however, refraction changes the circular path. The refraction is due to the interaction of the wave with the bottom. (It is refraction that slows down the waves in shallow water in proportion to



FROUDE NUMBER gives much information about the shape and force of a hydraulic jump such as a tidal bore. Abbreviated *F*, the number is named for William Froude, a 19th-century English civil engineer who worked on hydraulic jumps. *F* is equal to the ratio of the speed of the bore to the speed of the shallow-water waves on the undisturbed river in front of the bore. In each schematic diagram the observer is assumed to be moving at the same speed and in the same direction as the bore. In this frame of reference the bore's motion upstream disappears and the overall flow is from left to right, with the velocity decreasing as the flow passes through the bore. If *F* is from 1.0 to 1.7, the bore is undular. Above 1.7 all bores break. From 1.7 to 2.5 a weak jump is formed. From 2.5 to 4.5 a violent oscillating jump is formed; such a jump can send large waves downstream. From 4.5 to 9.0 a steady jump is formed; it is the best-defined type of bore. If *F* is greater than 9.0, large jets of water often extend ahead of the advancing bore.

their wavelength and therefore causes all shallow-water waves to move at the same speed.)

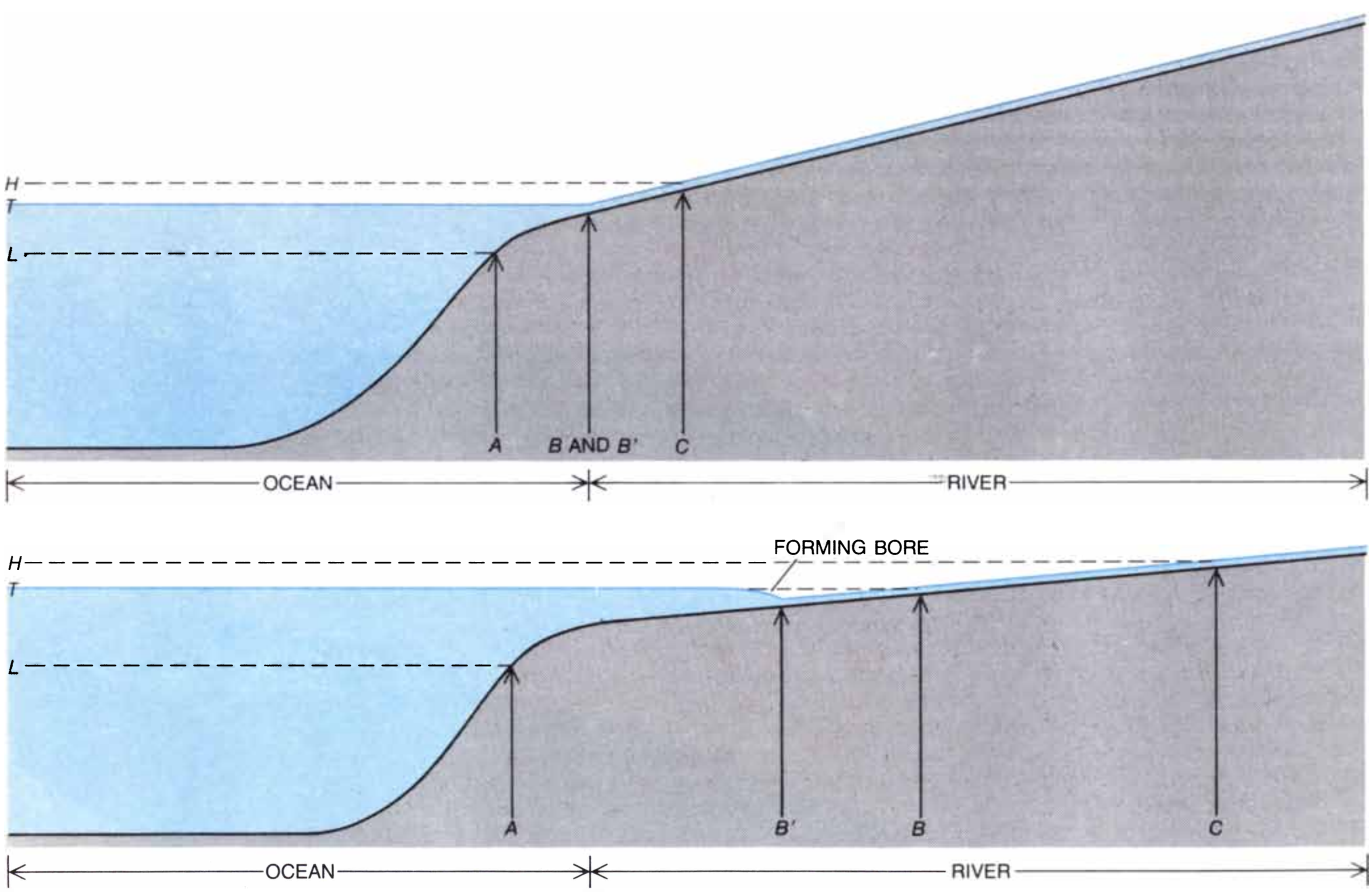
The path of the water particles is flattened into an ellipse by refraction. As the wave moves up the estuary the ellipses of particles near the top of the wave become longer and flatter than those of particles farther down in the wave. When this happens, the wave is said to be unstable; the paths of particles within it tend to become linear in the direction the wave is moving. At a certain point water begins to be carried toward the land as well as energy.

There are thus two velocities that are significant in analyzing the bore: the velocity of the wave and the velocity of the water particle. The relations between these velocities are an important part of the hydrodynamics of the bore. The study of the hydrodynamics of solitary waves has a long history. John Scott Russell, a hydraulic engineer, made the first report of a solitary wave on the Glasgow and Ardrossan Canal in Scotland in 1840. His account stimulated the interest of many workers; among them was the 19th-century British physicist Lord Rayleigh (John William Strutt), who made numerous contributions to the theory of solitary waves. Solitary waves were long thought to be merely an elegant mathematical curiosity. They now appear to have important applications in crystal-lattice theory, nonlinear optics, particle physics and molecular biology, among other disciplines.

As we have seen, the velocity of the water particle and the velocity of the wave can be related in several different ways. When the flow speed of the particle is less than the shallow-water wave velocity of the wave, the flow is said to be subcritical; when the speed of the particle is greater, the flow is supercritical. The transition between subcritical flow and supercritical flow is called a hydraulic jump; a tidal bore is a hydraulic jump moving upriver. As the tide enters the estuary the water particles begin to move forward. At the point where the particles assume the shallow-water wave velocity the bore forms.

Hydraulic jumps are common, although they rarely assume a form as dramatic as that of a tidal bore. For example, when the column of water from a faucet meets the surface of a large sink, a shallow layer of water spreads out to a certain radius, where it rises up in a circular wall of water. Outside the wall the water is deeper than it is inside.

The circular wall of water is a hydraulic jump. It results from the fact that when the column of water meets the surface of the sink, it spreads out with decreasing speed. Since the depth of the water flowing out from the base of the column is small, the shallow-water waves generated by random motions of the water in the column move slowly; indeed, the shallow-water wave velocity is much less than that of the flow. The flow is therefore supercritical and the waves are carried outward with the flowing water. Farther away from the column the water slows down and the flow becomes subcritical. The transition between the two regions is the hydraulic



CROSS-SECTIONAL VIEWS of the topography of the tidal basin and river mouth in two types of estuary show why bores form in some rivers but not in others. For a bore to form there must be a substantial tide and a gently sloping estuary. The diagrams show why that is so. In both the upper panel and the lower panel *H* indicates the high-tide level, *L* the low-tide level and *T* the current tide level. Because of the slope of the river bed the low tide intersects the river at *A;* the high tide intersects the river at *C.* The current tide would intersect the river at *B* if the tide level in the basin were extrapolated inshore. *B′* is the point to which the current tide actually extends. The speed of the intersection point *B* inshore depends on the slope of the estuary and the speed with which the tide rises. The actual extent of the tide inshore (*B′*) is limited by the velocity of the shallow-water waves that carry the tide. In most rivers (*top*) the tide rises slowly and the estuary is steep; the intersection point moves slowly and the tide can keep up. Therefore *B* and *B′* are at the same point. If the tide rises rapidly, however, and the estuary is gently sloping, as it is in some rivers (*bottom*), the intersection point will move inshore faster than the shallow-water waves of the tide can move. Therefore *B′* falls behind *B* and a difference in water level at the river mouth results. This is the bore.

jump. The elevated region beyond the jump is the accumulation of waves that propagate inward with a velocity equal to that of the flowing water and so are not swept away.

As we have seen, the formation of the hydraulic jump that constitutes a tidal bore requires a high tide and an estuary of the right shape. In addition the tide and the estuary must have the proper relation to each other. This relation can be explained as follows. As the tide rises it flows inshore to cover the previously exposed ground. The level of the tide can be extended horizontally inshore to the point where it intersects the surface of the river. The intersection point moves inshore as the tide rises. In most estuaries the intersection point moves inshore relatively slowly and the waves of the tide can propagate as fast as the intersection point moves inshore. In such estuaries no bore forms. In some instances, however, the estuary has a gradual slope and the intersection point moves inshore faster than the tidal waves can propagate. When the velocity of the shallow-water waves of the tide falls below that of the intersection point, the tide wells up higher than the river level and forms a bore.

As the tidal bore moves upstream it must move faster than the current of the river; otherwise it would be swept downstream. To an observer on the bank the bore is seen to move with a speed equal to the difference between the propagation speed of the hydraulic jump and the speed of the river.

The shape and the dynamics of the tidal bore, however, are not affected by the speed of the river. Indeed, the bore is most conveniently studied in a coordinate system that moves at the same speed as the bore, so that the bore appears to be stationary. In such a coordinate system the bore can be described by means of the Froude number (F), named in honor of the 19th-century English civil engineer William Froude. The Froude number is the square root of the ratio of the kinetic energy of the moving water to the gravitational energy in the waves. If F is smaller than 1, the flow is subcritical; if it is greater than 1, the flow is supercritical.

It turns out that the Froude number is equal to the ratio of the speed of the waves behind the bore to the speed of the waves on the undisturbed river in front of the bore. This fact yields a simple method of calculating F when the height of the bore is much less than the depth of the river. Since the speed of the waves on each side of the bore is proportional to the square root of the depth of the water, F is equal to the square root of the ratio of the depth behind the bore to the depth in front of it.

The Froude number is a compact description of the appearance and shape of a tidal bore. When F is between 1.0 and 1.7, an undular bore results; the advancing tide moves gently upriver. Only in the shallows near the river bank, where F is higher, does the bore break. When F is greater than 1.7, all bores are breaking bores. Between 1.7 and 2.5 the bore is called a weak jump; small rollers slide down the leading edge and the water behind the bore is smooth. Between 2.5 and 4.5 the jump is called oscillating; a vertical jet oscillates between the river bottom and the surface and generates surface waves with a large amplitude. Between 4.5 and 9.0 the jump is called steady; the leading edge is turbulent but the surface of the trailing edge can be smooth. Above 9.0 the jump is called strong; its energy is dissipated rapidly and the bore moves upstream in a thrashing, jerky way, with high-speed jets extending ahead of the main jump.

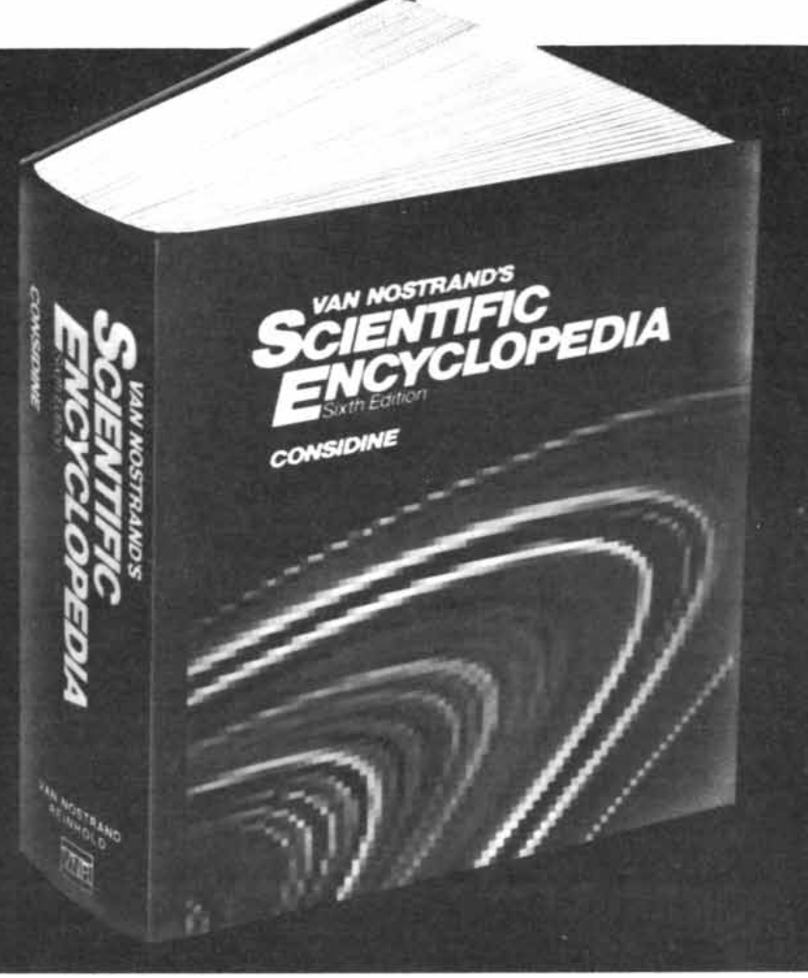
A common property of tidal bores is that their front is neither straight nor uniform. Some parts of the wave front lag behind the main wave and some propagate ahead. The cause of the variation is the change in the depth of the river from bank to bank. The parts of the wave in deeper water propagate faster and have a smaller Froude number. As a result bores are often led by a low undular wave; where there is a local shallow spot near the shore the bore will be higher and can be breaking.

If bores are dispersionless waves, how can they end? Two mechanisms drain the energy of the bore as it moves upstream. One is friction between the water and the bottom; such friction converts the kinetic energy of the moving water into heat. The other is viscosity, which is much like friction between adjacent molecules. Viscosity in the breaking wave transforms the propagation energy into turbulence and ultimately into heat. In addition, in rivers that become deeper or wider inland the bore can disappear because the funneling action that raised it in the estuary is reversed. When enough energy is dissipated, the speed of the bore becomes less than that of the river and the bore is carried downstream.

Bores, like many other phenomena on the earth, are vulnerable to the activities of man. A powerful bore known as the burro, which was as much as 15 feet high, once regularly went up the Colorado River. Over the past 15 years, however, agricultural water control and silting have transformed the mouth of the river. As a result the bore has been greatly reduced and is usually little more than a ripple lost in the river's other waves. The burro is still occasionally reported; other bores have been lost altogether. Still, the wonderful sight of a tidal bore surging upstream can be seen in many places by those willing to follow the new or full moon to where the river meets the sea.

Your fastest way to find sure answers in every area of science...

Examine without risk the new, fully updated Sixth Edition of the world's most popular science encyclopedia.



The equivalent of a multi-volume science library—at a fraction of the cost.

Here, in one convenient volume, are the answers to virtually all questions on every aspect of today's science, engineering, and technology.

This authoritative reference volume contains thousands of definitions, explanations, overviews, and specialized discussions. Reflecting the contributions of more than 200 distinguished authors and advisors, it provides a complete range of scientific and technical knowledge in clear language.

Newly revised from cover to cover.

Long hailed as the most comprehensive and practical reference volume of its kind, *Van Nostrand's Scientific Encyclopedia* has been considerably expanded and fully updated. More than 40% of the text and illustrations in the Sixth Edition is brand new, keeping pace with the explosive growth of technical knowledge.

This new edition contains over 3,000 pages, almost 3,000,000 words, and over 2,500 diagrams, graphs, and photographs. References within entries provide bridges to further information, and there are bibliographies for all major entries.

Examine this 'library' without risk.

This reference classic, in large format, is strikingly handsome. Its authority and clarity make it an ideal gift and a fine addition to any library.

Mail the handy coupon now, to see for yourself how this convenient single volume is like a multi-volume science library at a fraction of the cost. Examine the new Sixth Edition of *Van Nostrand's Scientific Encyclopedia* for 15 days without risk. Then you, and you alone, decide how important it can be to you.

Authoritatively covers ■ animal life ■ biosciences ■ chemistry ■ earth and atmospheric sciences ■ energy sources and power technology ■ mathematics and information sciences ■ materials and engineering sciences ■ medicine and physiology ■ physics ■ plant sciences ■ space and planetary sciences

New Sixth Edition gives you ■ over 3,000 pages ■ over 7,300 entries ■ over 2,500 diagrams, graphs, and photographs ■ almost 3,000,000 words ■ nearly 600 tables ■ over 9,500 cross references ■ metric and English units

On the previous edition...

"Highly informative, authoritative, accurate." —*American Association for the Advancement of Science*

"The best work of its kind." —*Industry Week*

VAN NOSTRAND REINHOLD Mail Order Service, 7625 Empire Drive, Florence, KY 41042

Act now and save.

This new edition is being offered at the special introductory price of $99.50 through December 31, 1982. On January 1, 1983 it will cost $107.50. You could not choose a better time to order this one-volume science 'library' than right now.

VNR **MAIL COUPON NOW FOR NO-RISK EXAMINATION**

VAN NOSTRAND REINHOLD
Mail Order Service
7625 Empire Drive, Florence, KY 41042

Please rush me the new Sixth Edition of *Van Nostrand's Scientific Encyclopedia*. If not completely satisfied, I may return it in 15 days for full refund.

☐ **Budget Plan:** Enclose $15.50 deposit and remit balance in three monthly payments of $28.00, for a total of $99.50 plus postage, handling, tax. Sorry, no credit cards on this plan.

☐ **Single Payment:** Save postage, handling. Enclose check or money order for $99.50 plus local tax or charge to credit card.

☐ **MasterCard** ☐ **Visa**

No. ________ Exp. ________

Sig. ________

Name ________
(Please print)

Address ________

City ________

State ________ Zip ________

Offer good in U.S.A. only. Subject to credit department approval. Payment must accompany order with P.O. box address. For introductory price savings, order must be postmarked no later than 12/31/82.

25161-0 S 7394

How Honeybees Find a Home

In most of the Temperate Zone a new colony of honeybees must locate a snug shelter in order to survive the winter. The search is carried out by the older "scout" bees with remarkable rigor

by Thomas D. Seeley

The honeybees arose in the Tropics, and two of the four living species (*Apis florea* and *A. dorsata*) still live only there. *A. cerana* and *A. mellifera,* however, have gradually extended their ranges far north and south of the Equator and therefore have confronted the problem of surviving through the winter, when there is no nectar or pollen to be gathered and when for days or weeks on end the temperature is below the level at which honeybees are able to fly (10 degrees Celsius, or about 50 degrees Fahrenheit). The remarkable method honeybee colonies have evolved for getting through the winter depends critically on the type of enclosure a colony occupies; in a suitable enclosure the colony may be able to carry on for several years, whereas in a poor one the bees may perish in the first winter. The adequacy of the enclosure in turn depends critically on the activity of the small group of older workers that chose it: the scout bees.

In the Tropics *A. florea* and *A. dorsata* nest in the open, hanging their nest from the underside of a tree branch, whereas *A. cerana* and *A. mellifera* nest in a cavity. It is this habit that has enabled the latter two species to extend their range. In the temperate zones the bees in a colony of *A. cerana* or *A. mellifera* spend the nonwinter months storing up 10 kilograms or more of honey that will serve as food during the winter and thus indirectly as a heating fuel. When winter comes, the colony contracts into a tight cluster. The bees in it generate heat by small vibrations of their powerful flight muscles, maintaining the surface temperature of the cluster at about 10 degrees C. In a snug shelter that has an adequate store of honey the bees usually get through the winter quite well.

Clearly, then, an important first step by a new colony is the selection of a snug shelter. In the wild the shelter is usually a cavity in a tree. Over the past five years I have been studying the house-hunting behavior of *A. mellifera* honeybees living in central New York State. These studies have revealed that the process of nest-site selection is remarkably thorough and have yielded findings that are of practical value to beekeepers.

House hunting is a part of swarming, which is the honeybees' method of establishing a new colony. Colonies swarm in the late spring; for bees in New York State the time is from late May to the end of June. Swarming follows a period of intense rearing of brood, which considerably enlarges the population of the colony and creates congestion in the nest cavity. Studies by James Simpson of the Rothamsted Experimental Station in England indicate that overcrowding stimulates the colony's workers to begin rearing a batch of daughter queens. The strongest queen (determined by stinging duels among the young queens) inherits the established nest. Once the rearing of new queens is well under way, and even before the first daughter queen emerges, the mother queen moves out with about half of the colony's 30,000 or so workers to start a new colony in another location.

The old queen and her retinue of workers depart in a mad whirl, pouring out of the nest and beginning a short flight. The swarm travels only a few tens of meters and then settles in a beardlike cluster on some object, usually a branch of a tree or bush. Soon after the swarm has settled the scout bees fly off in all directions to begin their reconnaissance.

Their search extends out to 10 kilometers or more from the old nest. The scouts are the oldest bees in the swarm, the ones that have already foraged for the colony and so are familiar with the territory around the old nest. They number a few hundred, or about 5 percent of the swarm's population.

Once the scouts have selected the future home site they make zigzag runs (punctuated by bursts of wing buzzing) through the swarm, thereby signaling the other bees to break up the cluster. Bernd Heinrich of the University of Vermont has shown that before lift-off a swarm warms itself throughout to about 36 degrees C., the temperature at which flight muscle functions best. In the final minutes before lift-off a swarm teems with scouts scrambling over the cluster, vibrating their wings and boring through the interlocked nets of hanging bees. A loud humming noise emanates from the cluster, a mixture of deep wing buzzes and shrill piping sounds. The sound reaches a climax as the once solid surface of the swarm appears to melt because the chains of hanging bees begin to disintegrate. Within another minute the entire swarm is airborne, filling the air with the penetrating buzz of thousands of bees circling tightly overhead.

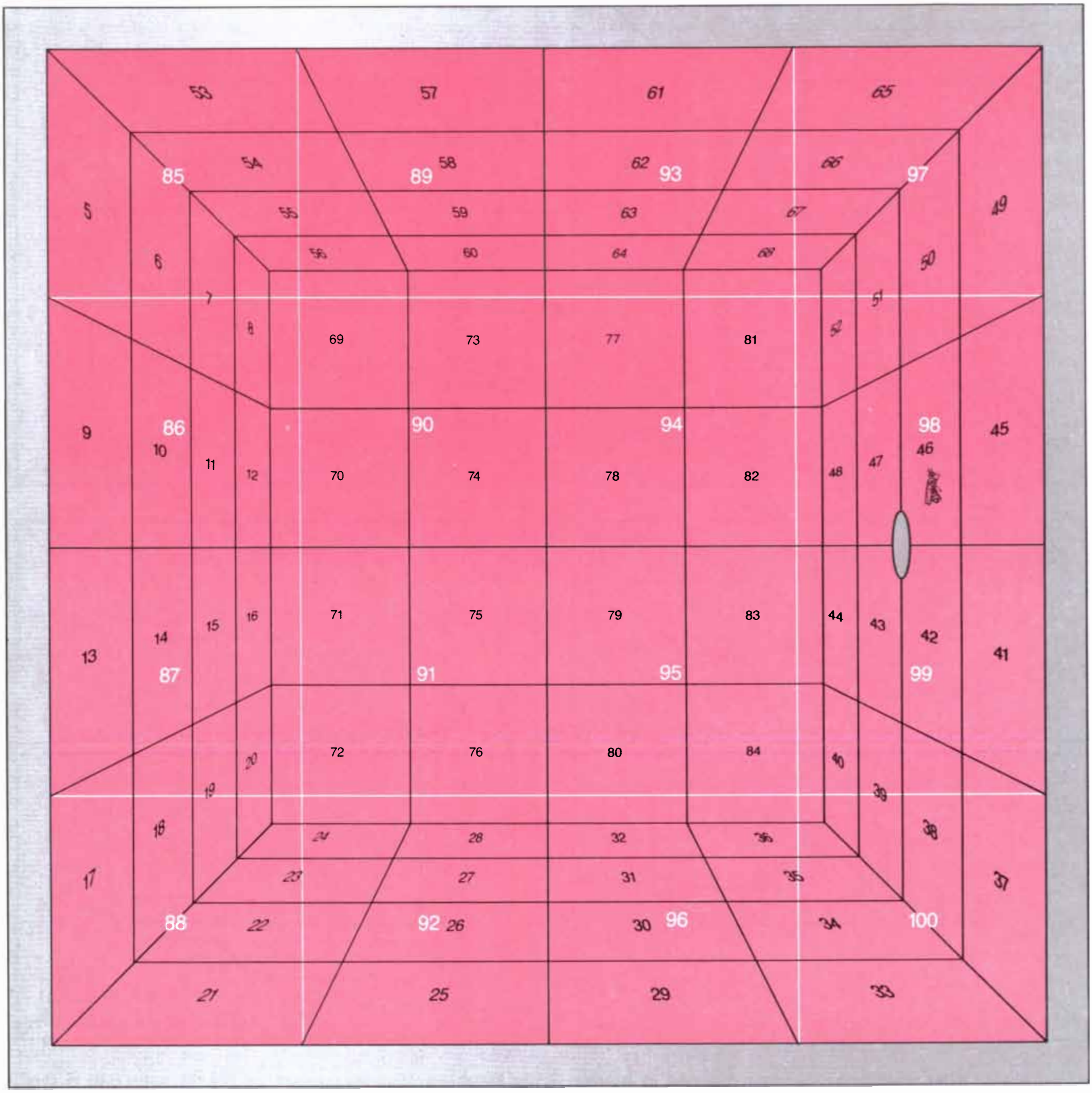
The airborne swarm forms a cloud about 10 meters in diameter. To pilot their sisters to the new home site the scouts streak through the swarm in the direction of the site. The swarm moves slowly at first, traversing the first 30 meters at a speed of less than one kilometer per hour, but within 200 meters it accelerates to 10 kilometers per hour or more, flying a few meters above the vegetation.

When the swarm reaches the nest site, the scouts somehow signal it to stop. Then they drop down from the cloud formed by the swarm, alight at the entrance to the nest (a knothole, a gap among the roots of a tree or a narrow crack in a limb) and release assembly pheromone from the Nassanoff gland in the tip of the abdomen. The pheromone, a chemical signal that the bees sense as an odor, pinpoints the entrance to the nest. Soon the other bees are streaming into the nest cavity in a formation that resembles a whirlpool. Within 30 minutes of lift-off nearly all the bees are safely inside their new home. Within a few hours they are cleaning out debris, constructing combs and flying off to forage for nectar and pollen. A new colony has been established.

Martin Lindauer of the University of Würzburg studied honeybee swarms in Munich just after World War II while he was working at the Munich Zoological Institute. He analyzed the process whereby scout bees decide which of the various sites discovered will be the swarm's future home. His first observation was that soon after a swarm assembles in a cluster near the old nest bees appear on the surface of the cluster doing "dances" of the type that Karl von Frisch, who was then also at the Munich Zoological Institute, had found to be the means by which foragers direct their nestmates to new sources of food.

Lindauer also observed that the dancers on a swarm never brought back nectar or pollen; therefore they were apparently not foragers. To make sure he marked the dancers with spots of paint and read from their dances the locations of the targets indicated. Further investigation called for a compass to measure the orientation of the dance with respect to the position of the sun and a stopwatch to time the cycles of the dance. The orientation indicates the direction to the site and the tempo of the dance indicates the distance: the faster the dance, the closer the target. Lindauer marked the indicated spots on a topographical map and set out to find precisely where the bees were going. In a few instances he found the bees he had marked. They were not foraging but were busily inspecting holes in the ground, hollows in trees or cracks in old walls. These bees were house hunters. Their dance announced prospective dwelling places, not patches of flowers.

By following the dances of the scouts continuously from the time a swarm settles until it lifts off Lindauer was able to determine how the scouts reach accord on a site. At first they search independently, and each one that finds a prospective site announces it independently by dancing on the swarm. Some sites are



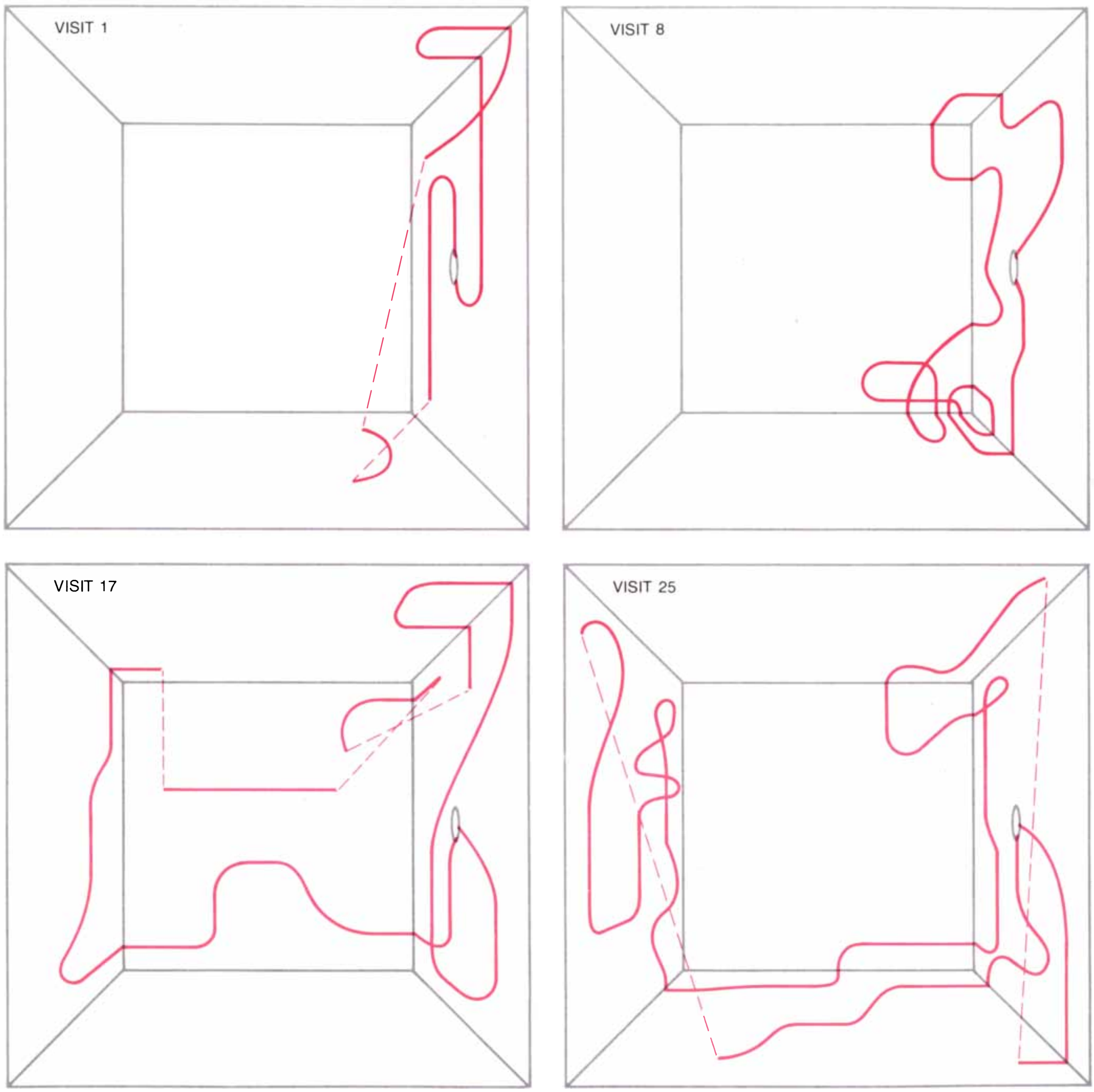
EXPERIMENTAL NEST BOX devised by the author enabled him to observe a scout bee as she explored a potential nest site. The box was mounted outside a red window on the side of a hut. Because bees do not see red light the experimenter could look through the window without disturbing the scout and could record her movements by means of numbered squares. Squares 85 through 100 are on window.

represented by a lively dance that is repeated for a long time, others by a dance a human observer can only characterize as unenthusiastic. The liveliness of a scout's dance reflects the quality of the site she has found. A dance for an inferior dwelling place is sluggish.

When a scout engaged in a sluggish dance encounters one that is dancing vigorously, she senses the lively dance and flies off to inspect the site. If her inspection reveals that it is indeed superior, she begins advertising it in her dance at the swarm. In this way the scouts gradually reach a consensus about the best dwelling place. Often it takes them several days to come to an accord. It is when they are all advertising the same goal in unison that they give the signal that causes the swarm to break up the cluster and fly to the new nest.

Lindauer's work revealed what happens at the clustered swarm while the scouts are house hunting but gave few clues about what happens at the prospective sites. What do the scout bees seek as an ideal site? How do they inspect a potential site and ascertain its properties? After repeating Lindauer's observations on the way the scouts reach and convey a consensus at the swarm I decided to explore these questions as the research on which I would base my doctoral thesis at Harvard University.

A logical starting point for studying what honeybees look for in a site for a



SCOUT'S METHOD of exploring a cavity is indicated by tracings of what a single scout bee did on four out of 25 visits during her initial inspection of a potental nest site. Where the line is solid the bee was walking; where it is broken she was flying. At first she stayed near the entrance, but eventually she explored the entire cavity. After the initial inspection she returned sporadically, apparently to check the site as external conditions changed. When she was away, she was either advertising the site to other members of the colony by means of a stereotyped "dance" on the surface of the swarm (the beardlike cluster the bees had formed) or visiting sites found by other scouts.

home is to study the nests of honeybee colonies living in nature. Each colony occupies a site its scouts have selected, and so it seemed reasonable to expect that patterns in the properties of such sites would yield clues about what the bees prefer. Working with Roger A. Morse of Cornell University, I found 21 honeybee colonies living in hollow trees in the forests around Ithaca, N.Y.

In order to fully describe the architecture of these natural nests we cut down each tree, transported the section with the nest to the laboratory and split open the log to expose the nest. We measured all the properties we could think of as likely to be significant to the bees: the height and size of the entrance and the shape and volume of the cavity (measured by removing the combs and filling the cavity with sand). Numerous consistencies emerged. A typical nest occupies a vertically elongate tree hollow with a volume of about 45 liters.

Once I knew what the natural homes of honeybees are like I sought to test whether the patterns we had found reflect the preferences of scout bees as they pursue their search. The idea for the experimental design of the test came from what I had read about beekeeping in East Africa. There the beekeepers acquire bees by hanging hives (hollowed logs with the ends stoppered except for an entrance hole) in trees and waiting for swarms to occupy them.

I tried the same thing around Ithaca except that I put up nest boxes in sets of two or three. In each set the boxes were identical except for one property, such as cavity volume or the height of the entrance above the ground. I hoped that scouts from wild swarms would discover my nest boxes and reveal their nest-site preferences by choosing among the boxes in a set.

The plan worked quite well. Over four summers approximately half of the nest-box sets attracted a swarm each summer. Moreover, the patterns of occupation were sufficiently consistent to show the nest-site preferences of the bees. Of course, in order to get statistically meaningful data I needed several occupations of the nest-box set for each variable in nest-site properties. To test for preferences in 11 variables I built 276 nest boxes, using up enough plywood (more than 70 sheets) for a small house.

The bees indicated preferences in the following nest-site variables: the volume of the cavity, the size of the entrance, the height of the entrance above the ground, the height of the entrance above the floor of the cavity, the direction the entrance faced and the presence of combs in the cavity. Honeybees avoid cavities with less than 10 liters of volume and more than 100. A small cavity cannot hold the store of honey the colo-

HONEYBEE SWARM hangs from a branch of a tree. A swarm consists of the queen from an established colony and about half of the colony's 30,000 or so workers. Soon after leaving the old nest the group settles in this manner and the scouts (a few hundred of the older workers) fly off to begin looking for a new site. They may go as far as 10 kilometers from the old nest.

SWARM IN FLIGHT is on its way to a new nest, on which the scouts have reached a consensus after a few hours or days. The swarm is moving from left to right. Most of the bees, which appear as dots or short streaks, are flying slowly. The scouts fly through the swarm rapidly, showing the direction of the new site. The scouts appear in the picture as relatively long streaks.

ny needs in order to survive the winter, and a large one may be difficult to heat in the winter.

Bees prefer an entrance that is less than 50 square centimeters in size, is at least two meters above the ground, faces south and opens into the bottom of the nest cavity. A small entrance is easily defended and helps to isolate the nest from the outside environment. High entrances are inaccessible to predators that cannot fly or climb trees and are harder to find than entrances near the ground. An entrance that faces south provides a warm, sunny perch from which foragers can take off and on which they can land. (Beekeepers face their manmade hives to the south to help their bees fly out in cool weather; the orientation is particularly important in the winter months, when bees fly out on sunny days to eliminate accumulated body wastes.) An entrance at the bottom of the nest cavity rather than at the top may help to minimize the loss of heat from the nest by convection currents. The preference for a site already filled with combs (built by a preceding colony that did not get through the winter) doubtless reflects the tremendous saving of energy that would otherwise have to go into the building of combs.

NATURAL HONEYBEE NEST was a cavity in a tree. Here the section containing the nest has been split open, revealing the combs containing brood and honey. The entrance hole is on the left side, about two-thirds of the way up the cavity. The author and Roger A. Morse of Cornell University examined a number of natural nests to determine what scout bees look for.

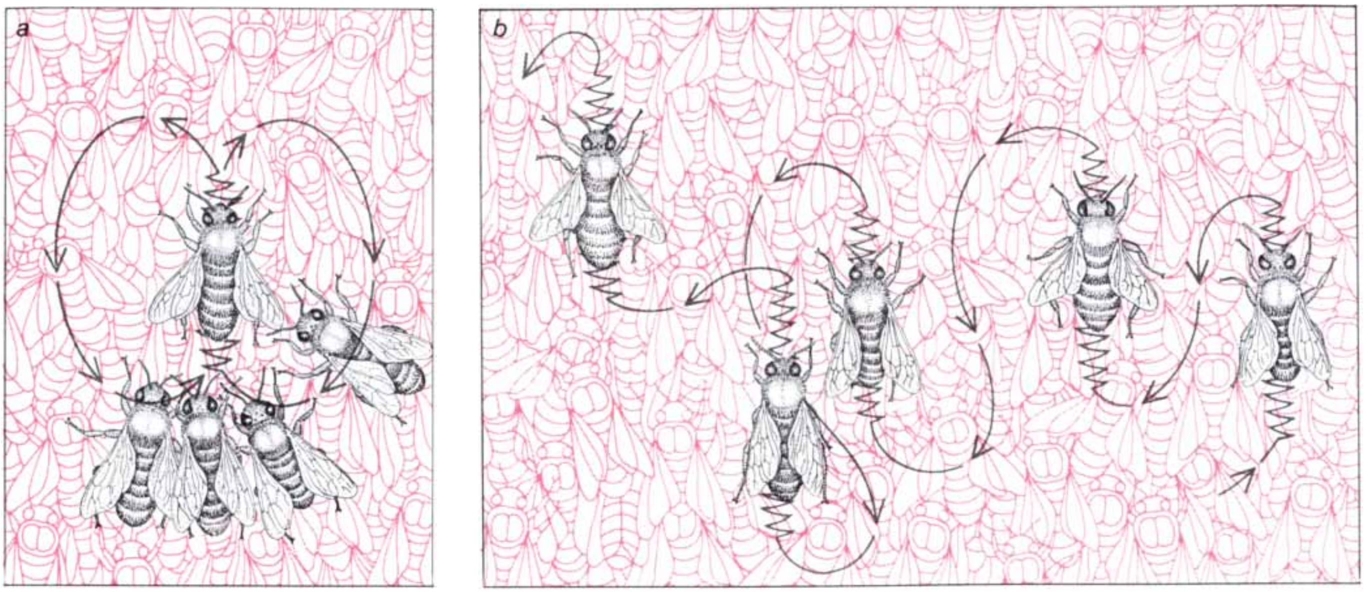
The nest-site properties for which I detected no preference were the shape of the entrance, the cavity's shape, its draftiness, its dryness and the distance of the site from the old nest. Honeybees probably prefer draft-free and dry nests, but because the colony can plug with tree resins any cracks that let in drafts and water the scouts apparently do not give much weight to these properties. In contrast, bees cannot modify the volume of a cavity or the height of the entrance and the direction in which it faces, and so they must make a choice in these matters before they move in.

The ability of bees to remedy a drafty or damp site was demonstrated by colonies that occupied our experimental nest boxes. I had made some of the boxes drafty by drilling into the sides a grid of six-millimeter holes spaced 10 centimeters apart; the bees plugged up all the holes with tree resins within a few days after moving in. They also quickly dried out damp nest boxes by dumping outside the soggy sawdust I had put on the floors and sealing the leaky wall and ceiling joints with resin.

Although I did these studies only to satisfy my curiosity about how honeybees choose a home when they are left to their own devices, the research yielded information of practical value to beekeepers. Applying the knowledge of what honeybees seek in a nest site, Morse and I designed what we call a bait hive, that is, a hive designed to catch wild swarms of honeybees. Physically the hive is basically just a box nailed to a tree, but its size, the characteristics of the entrance and the height above the ground meet the preferences of the bees and so make it an ideal nest site. (We do not fill bait hives with combs for fear of transmitting certain honeybee diseases to the incoming swarms.)

Between 1975 and 1980 we collected 251 bait-hive-years of data on our design; in the process we captured 124 swarms for an occupation rate of 49 percent per year. In the past beekeepers wanting to collect wild swarms had to rely on being notified when a swarm had settled somewhere; then they had to hurry to collect it in a hive before the bees could finish choosing a nest site and moving into it. With bait hives beekeepers can collect swarms automatically.

Up to this point in my studies I had treated the evaluation process of the nest site like a black box. I knew what scout bees seek in a prospective nest but I had little information on how they inspect one. To look into their inspection behavior I had to leave the countryside around Ithaca and shift my investigation to Appledore Island in Maine. This 39-hectare island lies 10 kilometers offshore and is the site of the

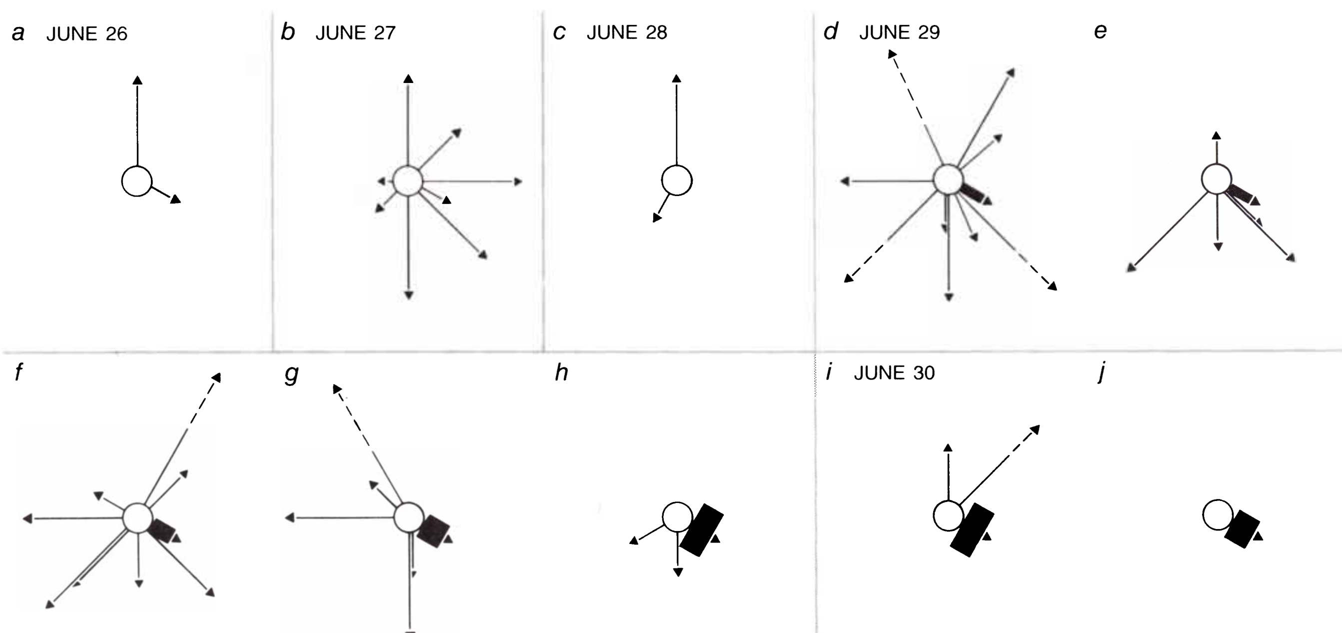


SCOUT'S DANCE advertising a nest site she has found is made on the surface of the swarm. Often the dancer moves in a pattern resembling a figure eight (*a*), waggling her abdomen during the central part. The direction of the waggling run conveys the direction of the site in relation to the sun; here the run is upward and means the site is toward the sun. The rhythm of the dance indicates the distance to the site; the farther the site, the longer the waggling middle part of the dance. Four other scouts are shown observing the dance. Sometimes a scout will dance in one place and then will walk a short distance along the surface of the swarm and perform the dance again (*b*).

Shoals Marine Laboratory of Cornell and the University of New Hampshire. Its particular importance for my work is that it has no large, hollow trees. Hence when I took honeybee swarms out to the island, I forced them to concentrate their house hunting on the experimental nest sites I provided. In this way I could easily observe and analyze the process of nest-site evaluation.

My first goal on Appledore Island was to record the behavior of scout bees inspecting nest sites. I hoped the observations would suggest how scouts evaluate the critical nest-site properties. It was particularly important to be able to watch scouts inside a prospective nest in order to understand how they judge such things as the volume of the cavity and the height of the entrance.

To this end I built a hut with a cube-shaped nest box mounted on one of its walls. The box was positioned outside a red window (bees cannot see red light) so that I could look in without disturbing the scouts. The inner surfaces of the box bore a grid-coordinate system that enabled me to record where a scout went while she was in the cavity. After setting up the hut on one



SUMMARY OF SCOUTING is given from the time a new colony left the old nest (at 1:35 P.M. on June 26) until it moved into the new nest site (at 9:40 A.M. on June 30). The observations were made by Martin Lindauer of the University of Würzburg. In each case the circle represents the location of the swarm, the arrows indicate the direction in which scouts found potential nest sites and the thickening of the arrows shows that increasing numbers of scouts were visiting a particular site. The longer an arrow is, the farther away the site was; distances beyond 1,000 meters are represented by broken lines. The most distant site (*f*) was 4,500 meters away. The scouts discovered 21 sites, but gradually they reached a consensus on the one that was 350 meters to the southeast, a site found by one of the first two scouts to advertise a find (*a*). On June 28 scouts were inactive because of rain. Multiple entries for June 29 and 30 are several periods of observation.

Physician, did you miss any of these significant developments in medical science?

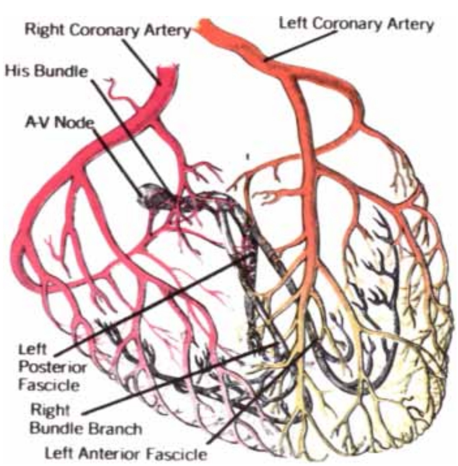
• ***Campylobacter fetus*** **subsp. *jejuni* is associated with a colitis that can clinically and sigmoidoscopically resemble acute idiopathic ulcerative colitis. Stool cultures are in order for *C. fetus* before beginning nonspecific anti-inflammatory therapy.**

• **Coumarin derivatives cross the placenta. A recent study shows that the consequences for the fetus can be severe. These include embryopathy, stillbirth, and premature delivery.**

• **Nonsteroidal anti-inflammatory drugs may produce a marked reduction in glomerular filtration rate; with termination of the drug, GFR returns to normal.**

• **The first documented incident of indigenous transmission of dengue in the continental United States since 1945 has been reported in Brownsville, Texas.**

IF THESE ITEMS are familiar you must be a prodigiously energetic or prodigiously lucky reader. With 2,000 or more journals published each year, information that significantly affects patient management all too easily slips by. Textbooks are out-of-date before they are published.



Branches of the right and left coronary arteries supply blood to the A-V node and intraventricular conduction system.

SCIENTIFIC AMERICAN *Medicine* is lucidly illustrated with drawings and photographs. Some examples are seen here and on the facing page.

SCIENTIFIC AMERICAN *Medicine* is the busy clinician's answer to this problem.

Because its authors update SCIENTIFIC AMERICAN *Medicine* every month, it is always current. Because the new information appears in a single source, it is there when you need it.

This 2,500-page, innovative union of publishing and electronic technology is the work of leading scholar-practitioners from Harvard and Stanford. The editors are Edward Rubenstein, M.D., F.A.C.P., and Daniel Federman, M.D., F.A.C.P.

Each month as authors update their contributions, revisions are entered on the magnetic tape on which the text and index are stored. The tape drives high-speed phototypesetting equipment so that subscribers receive about eight new chapters and a new index every four weeks; a bulletin highlights new developments.

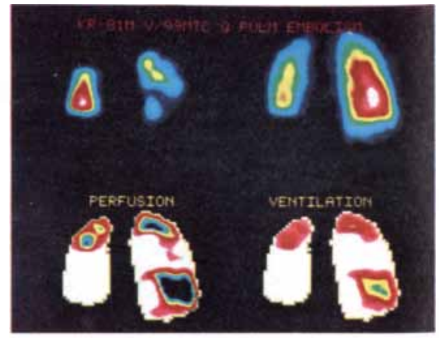
New material replaces old material in the living text, so that the information is there — up-to-date, at your fingertips.

A CME program of eight patient management problems offered over a 12-month period is available at no extra cost. As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions. This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians. This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

Trial Offer

We invite you to try SCIENTIFIC AMERICAN *Medicine* — for two months at no cost. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, we will bill you for $220 for the full 12 months (renewal is currently $170); otherwise return the two volumes.

Please mail the coupon today and let us take the hassle out of keeping up.



Computerized scintigraphy reveals pulmonary thromboembolism.

Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.

THE DISTINGUISHED AUTHORS AND THE FIFTEEN SECTIONS OF SCIENTIFIC AMERICAN ***MEDICINE***

1. Cardiovascular Medicine

Edgar Haber, M.D., Harvard Medical School and Massachusetts General Hospital

E. William Hancock, M.D., F.A.C.P., Stanford University School of Medicine

Roman W. DeSanctis, M.D., Harvard Medical School and Massachusetts General Hospital

Adolph M. Hutter, Jr., M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Eve Elizabeth Slater, M.D., Harvard Medical School and Massachusetts General Hospital

2. Dermatology

Eugene M. Farber, M.D., Stanford University School of Medicine

Elizabeth A. Abel, M.D., Stanford University School of Medicine

3. Endocrinology

Daniel D. Federman, M.D., F.A.C.P., Harvard Medical School

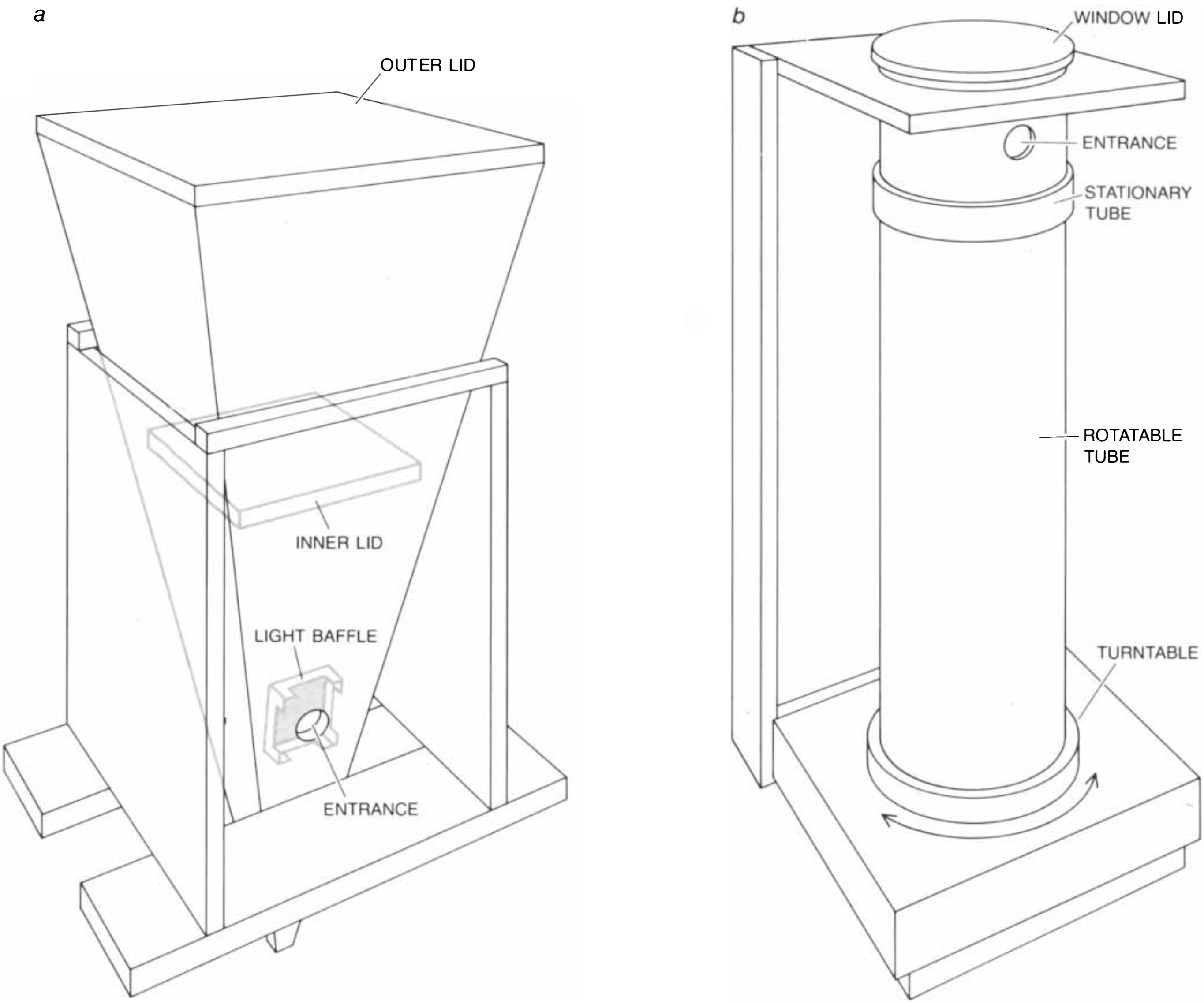
side of the island I would position a small swarm (about 2,000 bees) on the opposite shore, putting a dot of paint on each bee according to a color code that made them individually identifiable. Then I would retire to the hut to wait for the scout bees.

A scout bee needs about 40 minutes for inspecting a nest site. The complete inspection is a summation of numerous excursions inside the cavity, each one lasting for less than a minute and alternating with equally brief periods outside. I call this initial phase, in which a scout is primarily at the nest site, popping in and out of the cavity, the discovery phase.

Following the discovery phase a scout continues to visit a good site, but the visits become sporadic, each visit perhaps an hour after the preceding one and lasting for less than a minute. Evidently a scout conducts a detailed inspection of a prospective site during the discovery phase and thereafter spends most of her time elsewhere, either back at the swarm advertising the site or off inspecting other sites. The sporadic returns probably enable scouts to quickly check a site under different conditions, such as later in the day when the angle of the sun has changed or after a rainstorm that may have flooded the cavity.

When a scout is inside a cavity, she devotes most of her time (about 75 percent) to rapid walking about on the inner surfaces. This quick pacing is interspersed with brief flights, each usually less than a second in duration, in which she moves from one point to another. A geometric pattern in these inspections is that early in the discovery phase a scout walks about primarily near the entrance, whereas later she penetrates to the deepest recesses of the cavity. Three-dimensional reconstructions of the walking paths of individual scouts reveal that when the inspection is finished, a scout has walked 50 meters or more around the inside of the cavity and has covered all its inner surfaces.

The volume of a cavity is the nest-site property that is perhaps most critical to a colony's long-term survival. Since the colony needs at least 10 kilograms of honey to get through the win-



EXPERIMENTAL APPARATUS was devised by the author to test how a scout bee measures the volume of a cavity. The colony requires a site with a minimum volume of about 15 liters in order to store up the 10 kilograms of honey needed to get through the winter; anything larger than 100 liters is usually avoided, probably because the colony would have difficulty keeping warm inside it in the winter. With one apparatus (*a*) the experimenter could vary the volume between five liters (with the inner lid down) and 25 (inner lid up). By means of the light baffle he could change the amount of light coming in through the entrance hole to see whether scout bees estimating the size of the cavity relied mainly on walking or on vision. The chief sensory input turned out to be walking. In the other apparatus (*b*), which had a volume of 14 liters, the experimenter could rotate the wall in order to increase or decrease the amount of walking a scout bee had to do in order to measure the volume. The number of other scouts she recruited to the box was the measure of how she perceived its size.

Over a million people use Microsoft® software on their microcomputers. There's a reason.

Powerful simplicity. That's the concept behind every microcomputer software program we write. Powerful programs that allow you to do more with less effort. Programs that let you spend more time thinking about your problem... and less time thinking about the computer. It's a good reason why over a million people depend on Microsoft software in their everyday computing... and why the number is growing. Rapidly.

The value of experience. Microsoft made a giant step toward putting the power of computers in the hands of non-technical people when we wrote the first BASIC programming language for the very first microcomputer. It's an English-like programming language simple enough to teach in grade schools. Yet, it's so powerful that more application programs have been written in Microsoft BASIC than any other programming language.

Tools that help you think. Microsoft is an old-timer in one of the world's youngest industries. Microsoft BASIC was just the beginning. Today, we offer a broad range of proven tools for microcomputers. A complete line of programming languages. From the original BASIC to Pascal. Operating systems like XENIX™ and MS™-DOS that allow your computer to utilize a wide variety of programs. Products like the Microsoft SoftCard™ system which dramatically extends the capabilities of Apple® II computers. Management software such as Multiplan,™ which allows you to explore more alternatives to reach more intelligent business decisions. Better tools. Because they're powerfully simple.

Building a library. The great majority of microcomputers come equipped with Microsoft software. But you'll undoubtedly want more. Specialized programs that solve your particular problems. Programs for your industry, profession, or home. Programs that do more and ask less. Look to Microsoft, or to programs written in Microsoft languages. Software that's proven on millions of computers. Worldwide.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT®

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark of Microsoft Corporation
Multiplan, SoftCard, XENIX and MS are trademarks of Microsoft Corporation
Apple is a registered trademark of Apple Computer, Inc.

PROPERTY	PREFERENCE	FUNCTION
HEIGHT OF ENTRANCE	MORE THAN THREE METERS	DEFENSE OF COLONY
AREA OF ENTRANCE	LESS THAN 60 SQUARE CENTIMETERS	CONTROL OF MICROCLIMATE IN NEST AND DEFENSE OF COLONY
POSITION OF ENTRANCE	BOTTOM OF CAVITY	CONTROL OF MICROCLIMATE IN NEST
DIRECTION OF ENTRANCE	FACING SOUTHWARD	CONTROL OF MICROCLIMATE IN NEST
VOLUME OF CAVITY	BETWEEN 10 AND 100 LITERS	STORAGE SPACE FOR HONEY AND CONTROL OF MICROCLIMATE IN NEST
COMBS	COMBS IN CAVITY	ECONOMY IN CONSTRUCTION OF NEST

NEST-SITE PROPERTIES that honeybees apparently prefer are listed on the basis of nest-box selections made by bee swarms. A site with all the properties presumably would be ideal.

ter and that amount of honey requires about 15 liters of storage space, the volume must be at least 15 liters if the colony is to survive the winter.

How do scout bees measure the volume of a cavity? Their active walking in the course of an inspection may provide the basis for an estimate, but another hypothesis is that they simply go inside and look around. By means of experiments with nest boxes in which the traversable surface area and the interior illumination could be varied I found that in order to measure a volume scout bees need either inner surfaces that can be traversed or interior illumination of more than .5 lux (about the illumination provided by a full moon).

What are the conditions inside a cavity in a tree? Certainly the inner surface of bare wood is easily traversed by a scout bee. To measure the level of illumination in a similar cavity I built a model based on the measurements I had made of natural nests. It had a series of openings into which I inserted a light meter. I found the illumination to be less than .5 lux except near the entrance. Evidently in nature scout bees rely primarily on walking about in a prospective nest site to measure its volume.

To test this hypothesis more directly I tried modifying a scout's perception of volume by manipulating the amount of walking required to move from point to point inside a cavity. The device for this experiment was a cylindrical nest box mounted vertically on a turntable, enabling me to smoothly rotate the box while a scout bee was inside. By means of a window at the top I could look inside and see which way the bee was walking; then I could turn the walls according to whether I wanted to increase or decrease the amount of walking required for her to complete a horizontal circuit.

The volume of this experimental box was 14 liters, on the boundary between an unacceptably small cavity and a suitably large one. If walking contributes to the perception of volume, the first scout to discover the box should find it either more or less attractive than its true volume would suggest according to whether she had been made to walk more or less than she would in a normal 14-liter cavity. The test was the number of other scouts the first scout recruited to visit the box; she should recruit more scouts if she found the box suitably large than she would if the box had seemed unacceptably small. That is what I observed in four trials of this delicate experiment. It seems clear a scout's estimate of the volume of a cavity is proportional to the amount of walking she must do to circumnavigate it.

The founding of a honeybee colony is fraught with danger. To survive the first winter the colony must surmount the many hurdles of finding a good site, building a nest of energy-expensive beeswax combs, rearing offspring that can outlive the winter and gathering the necessary provisions for winter. Most colonies do not succeed. Long-term observations of the forest-dwelling colonies around Ithaca have revealed that only 24 percent of the newly founded colonies survive their first winter, whereas the survival rate for established colonies is 78 percent.

These observations have also revealed that if a colony does survive the critical first winter, it will endure on the average for another five years. In short, a colony has the potential to survive for a long time but faces great risks in moving from an old nest to a new one. Therefore a swarm cannot rely on trial-and-error methods in finding a suitable site. Each colony must make a single, careful decision with which it can live for many years.

THE CARE AND FEEDING OF A LIGHTWEIGHT.

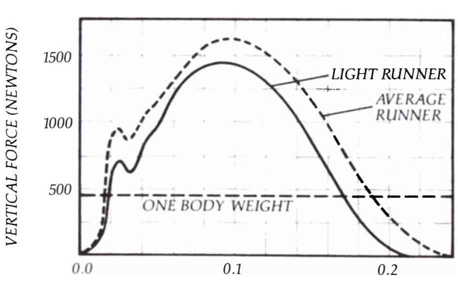
Believe it or not, one of the most envied group of runners—the lightweights—has a very unenviable problem.

While the rest of us are out there causing street lamps to sway, these folks are hitting the ground with considerably less force. The typical 6 footer who is 20 pounds underweight, is receiving a jolt about 14 percent less than normal.

This is a problem? You bet. Especially when you start cranking out the weekly miles in shoes designed for the average weight runner.

What lightweights should fear most is too much cushioning. An overdose here and the shoe can become dangerously inflexible.

If you're one of nature's more slimmed down versions, you won't be applying enough weight to penetrate the forefoot area as much. And what you lack in pounds, you'll have to make up for in effort. It takes more energy to flex a thicker sole, so you'll run less efficiently.



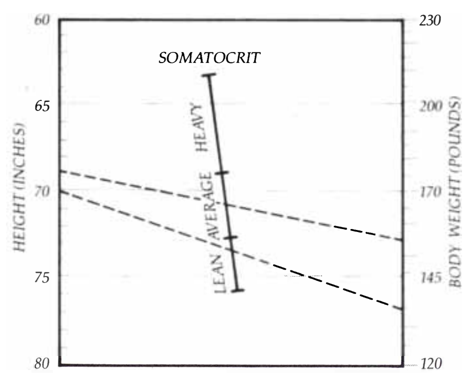
Vertical ground reaction forces for a 150 lb. runner and a 100 lb. runner at a 6:00 mile pace. Forces under the heel and forefoot are both proportionally smaller for the lighter runner.

You may also be flirting with shin splints and Achilles tendinitis, the most common results of too little flexibility.

Unfortunately, things really get sticky if your shoe size is less than a size 9. Because most midsoles aren't scaled the way the human body is.

Until we started investigating, the general assumption was that the forces under the foot were directly proportional to shoe size.

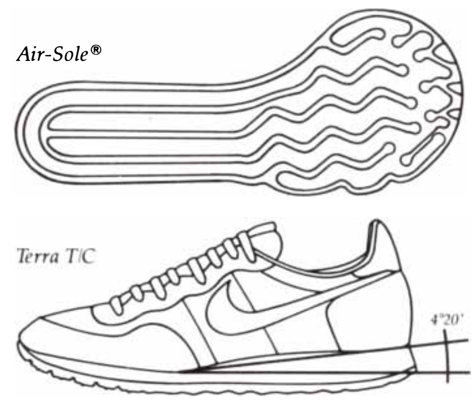
In other words, if a size 7 shoe is 9 percent smaller than a size 10, then a typical runner in a size 7 will experience 9 percent less vertical force.



Are you a lightweight? Draw a line between your height and weight and see where it intersects the somatocrit scale.

As it turns out, nature doesn't work that way. In actuality, that size 7 runner will hit the ground with 18 percent less force.

All of which means the smaller sizes of most running shoes have an over-generous amount of cushion. For even the average weight runner.



The Air-Sole® (top) in the Columbia, Aurora and Tailwind has different air pressure in the various sizes to give more appropriate cushioning. The Terra T/C (below) has a Phylon™ midsole also scaled for cushion as well as the same degree of heel lift in all sizes.

It was a startling discovery. And we did something about it.

First with our air shoes—the Columbia, Aurora, and the Tailwind. Not only are these shoes slip lasted for the flexibility lightweight runners need, but we also adjusted the air pressure in the various sizes for more appropriate cushioning.

Lightweights will also appreciate the Terra T/C and the Lady Terra T/C. They are slip lasted in the front to make them flexible and board lasted in the rear to make them stable. In addition, the midsole has been molded to give the proper cushion and heel lift for each size.

So if you're on the thin side, put your feet on a diet. When it comes to cushioning, don't ask for second helpings.

Or you'll wind up a glutton for punishment.

NIKE
Beaverton, Oregon

Historical Eclipses

Reliable records of solar and lunar eclipses go back as far as 750 B.C. They bear on such questions as whether the sun is shrinking or the earth is not spinning as fast as it once did

by F. Richard Stephenson

Observers on our planet—astronomers, historians and even poets—have been recording eclipses of the sun and the moon for more than 2,500 years. Whatever their motivation, their records, both ancient and more recent, can help to answer questions that challenge investigators today. One such question is: Why is the length of the day (or the earth's rate of rotation) changing? Another is: Is the sun shrinking? On the first question eclipse observations long before the rise of telescopic astronomy have provided information of much value. On the second only the records of total solar eclipses since the 18th century are pertinent.

We owe the term "eclipse" to the ancient Greeks: *ekleipsis* means "failure," in the sense of something gone wrong. The image is apt, and if the orbits of the earth and the moon were in exactly the same plane, these failures would be far more frequent: two per month. At each conjunction, or new moon, the sun would be eclipsed; at each opposition, or full moon, the moon would be eclipsed. Actually, since the orbital planes are tilted with respect to each other by some five degrees, eclipses can come only when the new or full moon happens to be near one of the two "nodes" of its orbit. These are the points, 180 degrees apart, where the plane of the moon's orbit intersects the plane of the earth's. This limitation means that any one year may have as few as two eclipses (the number in 1984) or as many as seven (the number this year).

Weather permitting, a man in his lifetime might expect to see some 50 lunar eclipses, more than half of them total, and perhaps 30 partial solar eclipses. A total eclipse of the sun, however, is a rare event at any one location. For example, the last total solar eclipse visible in the vicinity of New York City was in 1925 and the next will not be until 2079. What makes the spectacle so rare is that the sizes of the sun and the moon in the sky are almost identical, and as a result the conical shadow cast by the moon barely reaches the surface of the earth. The path of totality may be some 15,000 kilometers long, sweeping across as much as 140 degrees of longitude, but the umbra, or region of dark shadow, is seldom more than 250 kilometers wide.

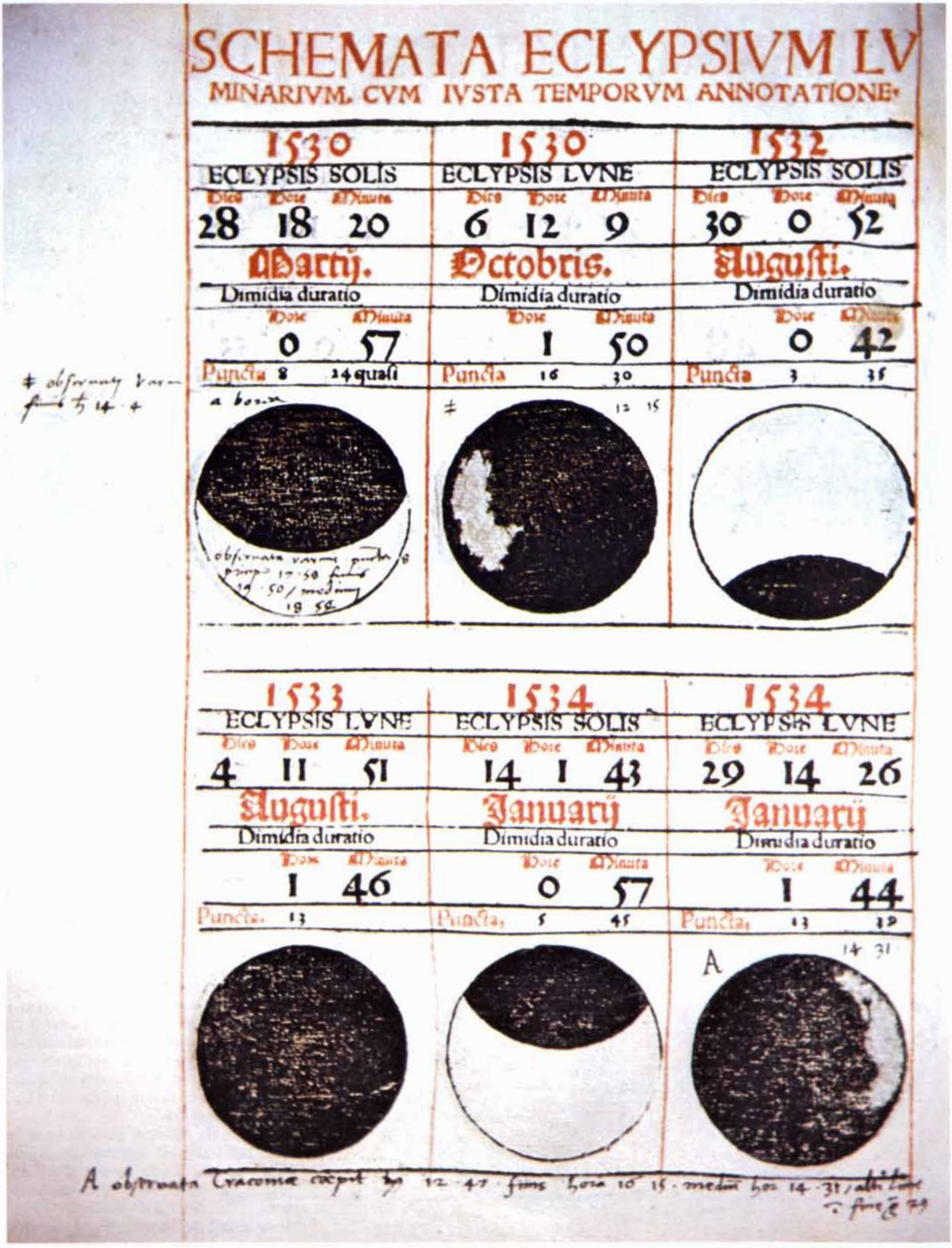
Whether a given solar eclipse is total or annular (with a bright ring of sun surrounding the disk of the moon) depends on the position of the moon in its elliptical orbit. When the moon is near perigee, its point of closest approach to the earth, it is able to cover the sun completely for as long as 7.5 minutes. When the moon is near apogee, its point of most distant recession, its shadow falls considerably short of the earth and the eclipse is annular. The duration of an annular eclipse can be as long as 12.5 minutes. Astronomers call total, annular and near-total eclipses of the sun large events, as opposed to the smaller events of partial eclipses.

Is the sun really shrinking? Two large bodies of data bearing on the subject are the daily observations, weather permitting, of the transit of the sun (in order to measure its right ascension) at the Royal Observatory in Britain since 1750 and at the U.S. Naval Observatory in Washington since 1846. The observations include the time required for the sun's image to pass a wire fixed in the eyepiece of a telescope; the interval, of course, also measures the size of the image. In 1979, after analyzing substantial amounts of the data from both observatories, John A. Eddy of the High Altitude Observatory of the National Center for Atmospheric Research and Aram A. Boornazian, a mathematician at S. Ross and Company in Boston, announced their conclusion that the sun is indeed shrinking, and at a considerable rate. They estimated the shrinkage of the sun's horizontal diameter to be about two seconds of arc, or about .1 percent, per century.

The sun's disk is brilliant, however, and the transparency of the earth's atmosphere is both imperfect and variable; these factors make accurate determination of the solar diameter by this kind of observation difficult. It is therefore not surprising that individual transit observations are less than precise and are also subject to considerable observer bias. As a result at least part of the trend detected by Eddy and Boornazian could actually be an artifact of observational error. An alternative and independent source of data regarding solar shrinkage is the duration of total solar eclipses. (A second independent check is provided by the time the planet Mercury takes to cross the solar disk.) Such observations allow a monitoring of changes in the sun's diameter at fairly regular intervals back to about the start of the 18th century, or roughly twice the span of time covered by the British and American transit data analyzed by Eddy and Boornazian.

Few total solar eclipses were visible in Europe in the 17th century. I know of only two observations: one made in southern Italy in 1605 and one made in northern Ireland in 1652. Indeed, the first known telescopic observation of a total solar eclipse was made in France in 1706. The first total solar eclipse to be carefully timed by a substantial number of observers in Europe was the eclipse of 1715, which was on May 3 (according to the Gregorian calendar). Most of the credit for this effort goes to Edmund Halley. Well before the event he "caused a small map of *England*, describing the Track and Bounds [of the eclipse], to be dispersed all over the Kingdom, with a Request to the Curious to observe what they could about it, but more especially to note the time of total Darkness, as requiring no other instrument than a *Pendulum Clock* with which most Persons are furnish'd, and as being determinable with the utmost Exactness, by reason of the momentaneous [sic] Occultation and Emersion of the luminous Edge of the Sun, whose least part makes Day."

Halley's motive in promoting the timing experiment was to improve the accu-



SCHEMATA ECLYPSIVM LV
MINARIVM. CVM IVSTA TEMPORVM ANNOTATIONE.

1530 ECLYPSIS SOLIS	1530 ECLYPSIS LVNE	1532 ECLYPSIS SOLIS
Dies 28 Hore 18 Minuta 20	Dies 6 Hore 12 Minuta 9	Dies 30 Hore 0 Minuta 52
Martij.	Octobris.	Augusti.
Dimidia duratio Hore 0 Minuta 57	Dimidia duratio Hore 1 Minuta 50	Dimidia duratio Hore 0 Minuta 42
Puncta 8 24 quasi	Puncta 16 30	Puncta 3 35

1533 ECLYPSIS LVNE	1534 ECLYPSIS SOLIS	1534 ECLYPSIS LVNE
Dies 4 Hore 11 Minuta 51	Dies 14 Hore 1 Minuta 43	Dies 29 Hore 14 Minuta 26
Augusti.	Januarij	Januarij
Dimidia duratio Hore 1 Minuta 46	Dimidia duratio Hore 0 Minuta 57	Dimidia duratio Hore 1 Minuta 44
Puncta. 13	Puncta, 5 45	Puncta, 13 39

THREE PARTIAL SOLAR ECLIPSES (in 1530, 1532 and 1534) and three total lunar eclipses are predicted on this page of *Calendarium Romanum Magnum,* published in 1518. The author, Johann Stöffler of Tübingen, was a prominent Renaissance astronomer-mathematician. The handwriting in this copy of Stöffler's work, now in the library of the University of Uppsala, is that of Nicolaus Copernicus.

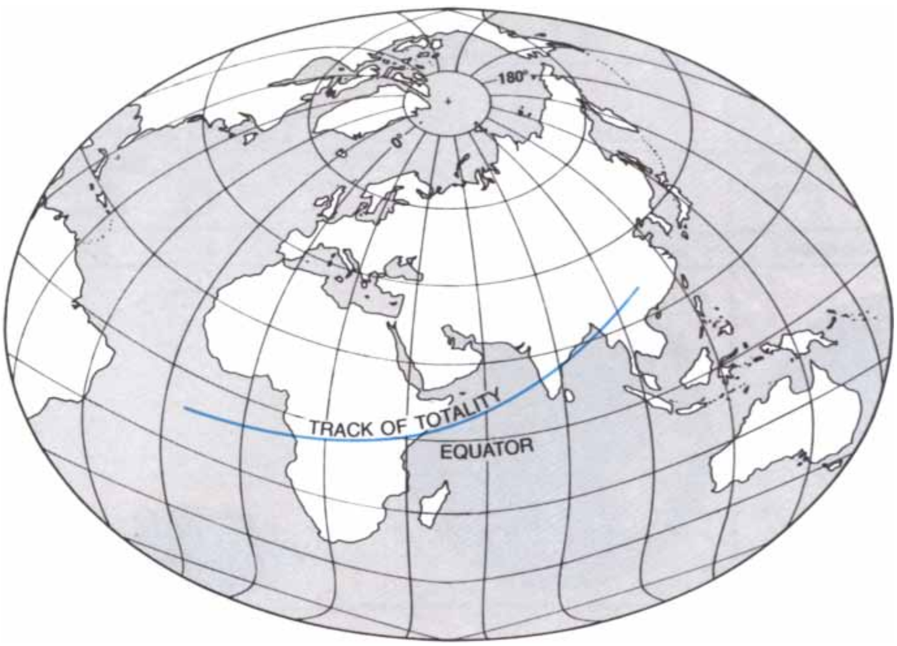
racy of future eclipse predictions. He harvested nine ostensibly accurate measurements from various parts of Britain. The duration of totality (about 3.3 minutes) was expressed to the nearest second of time; all but one of the nine measurements showed a small scatter, not much greater than the rounding error. Even the one discordant observation deviated by only seven seconds from the 200-second mean of the other eight, hardly a bad result. Converting the average of the eight into a measurement of the sun's diameter as it is now standardized, the sun in 1715 was smaller than it is today by two-tenths of a second of arc (with a probable error of four-tenths of a second).

The next total solar eclipse after the one of 1715 that was carefully timed came more than 125 years later. From 1842 until early in this century, however, most total eclipses were accurately timed, and astronomers often traveled to remote places to observe them. After 1925 the focus of eclipse observation shifted, to the neglect of totality timing, to such topics as the identification of lines in the solar spectrum and the study of the structure of the solar corona. Only lately has the topic of timing, with the specific objective of measuring the diameter of the sun, regained its place. Much of the credit must go to Eddy and Boornazian.

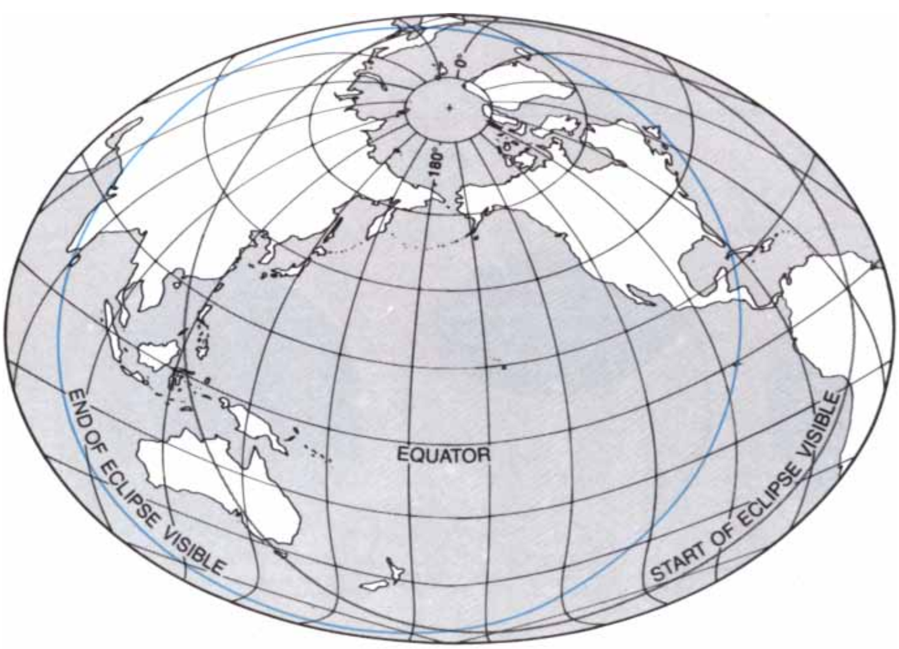
When the times of total solar eclipses over the past three centuries or so are analyzed in conjunction with the data on the transits of Mercury, they suggest a conclusion contrary to the one based on solar right-ascension observations. In brief, the data indicate only a negligible change in the sun's diameter. John H. Parkinson of University College London, Leslie V. Morrison of the Royal Greenwich Observatory and I have calculated that the percentage of decrease is $.008 \pm .007$ per century. Surprisingly, however, the data provide fairly strong evidence that the diameter of the sun oscillates. The period of the oscillation is some 80 years and its amplitude is about .025 percent. After investigating a wide variety of data Ronald L. Gilliland of the High Altitude Observatory supports this interpretation. In any event measurement of the sun's diameter on a regular basis seems worth pursuing in the future.

When it comes to the variations in the earth's rate of rotation, only eclipse data from ancient and medieval times are of value. From about 1620 onward telescopic observations of the occultation of stars by the moon have replaced observations of solar and lunar eclipses as a source of information on the length of the day. The reason is that the occultation of a star is almost instantaneous. Among eclipses only total solar obscurations can be timed with comparable precision, and they are much rarer than occultations.

The study of star occultations has revealed irregular fluctuations in day length over periods on the order of a decade, hence the term "decade fluctuations." The fluctuations themselves are quite small, no more than two or three milliseconds above or below the average, but their cumulative effect is fairly large. For example, if the length of the day remained continuously about three milliseconds above the average for a full



TOTAL ECLIPSE OF THE SUN in 1980 was visible on February 16 along a track no more than 150 kilometers wide that ran from west of the prime meridian to western China, a distance of 13,500 kilometers. It was visible mainly at sea and totality lasted for only about four minutes.



TOTAL ECLIPSE OF THE MOON at the end of 1982 will also be visible mainly at sea, but if weather permits, it may be seen from start (*black ellipse*) to finish (*colored ellipse*) in Iceland, Greenland, all of Canada except eastern Newfoundland, all 50 states of the U.S., all of Mexico, all of Japan and Korea, much of the eastern and arctic U.S.S.R., most of the Philippines, parts of Indonesia and Australia and all of New Zealand during the night of December 30–31.

decade, the accumulated total of what is called clock error (the difference between a clock keeping earth time and an "ideal" clock) would be about 10 seconds. Morrison and I have used the occultation data to map the decade fluctuations in detail from the 17th century to the present.

How can these fluctuations be explained? One hypothesis suggests that fluid motions in the core of the earth, which are responsible for the earth's magnetic field, are coupled, electromagnetically or possibly topographically, with the surrounding mantle and disturb the rotation of the mantle. Such disturbances would then be communicated to the earth's surface. Another hypothesis points out that small changes in the global sea level, produced by the melting or freezing of polar ice, would alter the planet's moment of inertia and so affect its rate of rotation.

For all their relative imprecision ancient and medieval eclipse data are nonetheless valuable for a second kind of study: the long-term changes in the earth's rate of rotation. These changes are masked in recent centuries by the short-term changes described above. For example, it has long been realized that the tides of the earth's oceans, the product of lunar (and to a lesser extent solar) gravitational influences, have a braking effect that is responsible for a gradual increase in the length of the day. The moon's tidal influences can be readily estimated today because of modern studies of the moon's motions, such as lunar laser ranging. These studies show that the orbit of the moon is slowly getting larger, so that our satellite is receding from its planet at a rate of about four centimeters per year.

The two phenomena are interrelated. As the earth's rotation slows, the planet loses angular momentum. The angular momentum of the earth-moon system, however, is conserved, and so what the earth loses the moon gains. An orbiting satellite that gains angular momentum actually loses speed and so recedes from its planet; this is what the moon is doing. Calculations indicate that the tidal gain in the moon's angular momentum adds about .04 second to the length of the month each century.

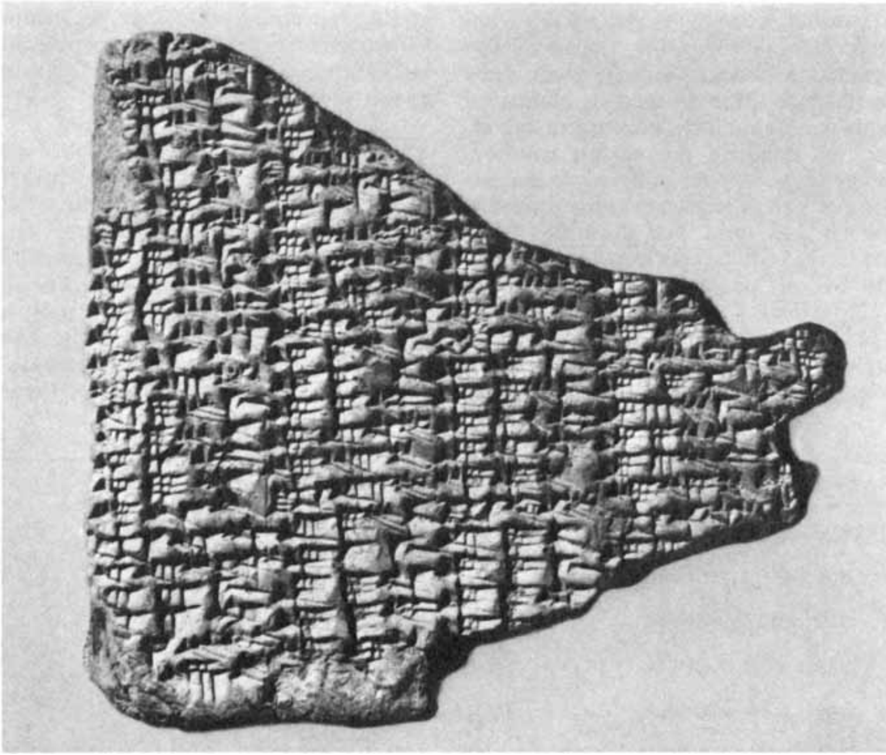
Like the decade fluctuation, such a small change seems a trifling matter. Its long-term effects are nonetheless important. For example, let it be assumed that tidal friction has been the same for many millions of years. Then 100 million years ago, in mid-Cretaceous times, the mean distance between the earth and the moon would have been 4,000 kilometers less than it is at present, and each month would have been shorter by somewhat more than 11 hours.

How can the record of ancient and medieval eclipses yield useful data on a gradual slowing of the earth? Consider clock error. Over a span of 100 years the clock keeping earth time would lose some 45 seconds compared with the ideal clock, because of the effect of tides on the earth's rotation. What is more, over still longer periods the clock error increases as the square of the elapsed time. For example, 1,000 years ago the expected accumulated clock error due to tides would have been an hour and a quarter, and 2,000 years ago it would have been five hours. Since at least some eclipse records are well over 2,000 years old, if certain facts about the observation can be established, the record can provide significant information on trends in the earth's rate of rotation.

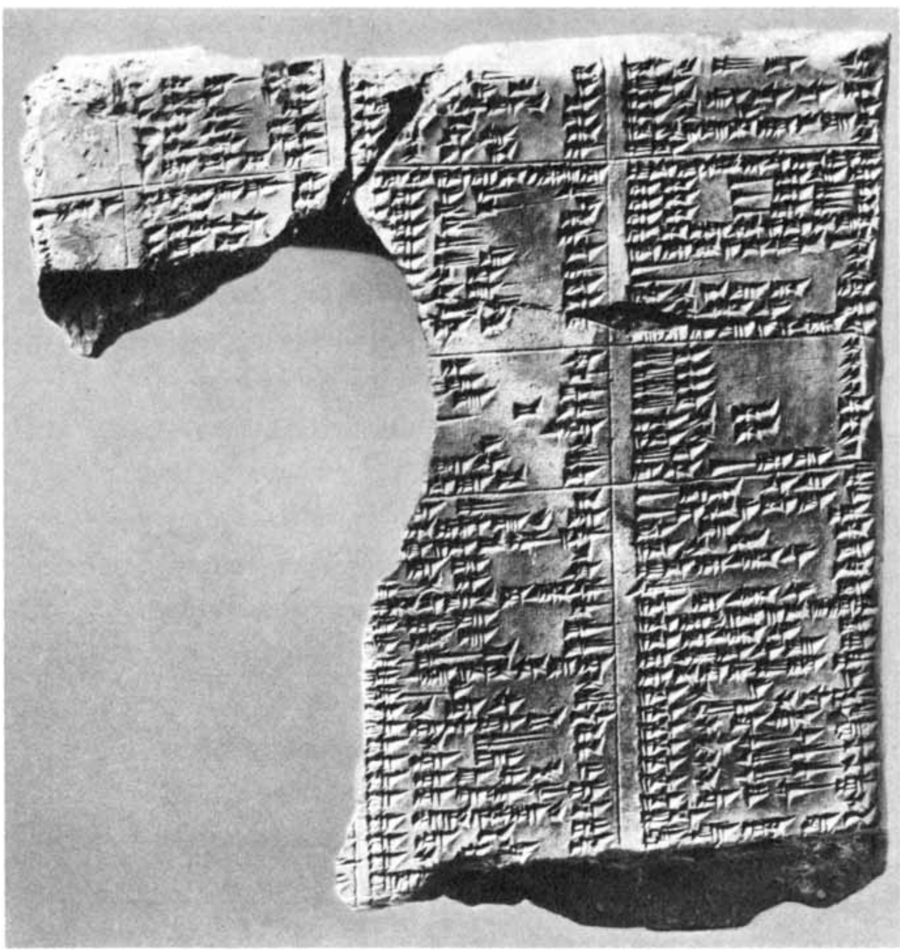
Before presenting the conclusions that can be drawn from such records it will be helpful to indicate what criteria the early observations must satisfy and to give an abbreviated account of the observations area by area. For an early observation to be of value it must fall into one of three categories. The first category consists of the most valuable observations: timed solar and lunar

YEAR	DAY	LOCATION	DESCRIPTION
B.C. 709	JULY 17	CHINA	TOTAL: NO OTHER DETAILS
601	SEPT. 12	CHINA	TOTAL: NO OTHER DETAILS
549	JUNE 12	CHINA	TOTAL: NO OTHER DETAILS
198	AUG. 7	CHINA	ANNULAR: NO OTHER DETAILS
181	MAR. 4	CHINA	TOTAL: NO OTHER DETAILS
136	APR. 15	MESOPOTAMIA	TOTAL: MANY STARS AND FOUR PLANETS SEEN
A.D. 2	NOV. 23	CHINA	TOTAL: NO OTHER DETAILS
65	DEC. 16	CHINA	TOTAL: NO OTHER DETAILS
516	APR. 18	CHINA	ANNULAR: NO OTHER DETAILS
522	JUNE 10	CHINA	TOTAL: NO OTHER DETAILS
840	MAY 5	ITALY	TOTAL: "SUN HIDDEN FROM WORLD"
873	JULY 28	PERSIA	ANNULAR: "MOON WITHIN BODY OF SUN"
912	JUNE 17	SPAIN	TOTAL: "DARKNESS JUST BEFORE SUNSET"
968	DEC. 22	TURKEY (2)	TOTAL: CORONA OBSERVED
975	AUG. 10	JAPAN	TOTAL: "INK-COLORED SUN"
1124	AUG. 11	RUSSIA	TOTAL: "SUN PERISHED COMPLETELY"
1133	AUG. 2	EUROPE (5)	TOTAL: "SUN BLACK AS PITCH"
1176	APR. 11	TURKEY	TOTAL: "NIGHT FELL, STARS APPEARED"
1185	MAY 1	RUSSIA	TOTAL: CHROMOSPHERE OBSERVED
1221	MAY 23	MONGOLIA	TOTAL: "STARS VISIBLE"
1239	JUNE 3	EUROPE (8)	TOTAL: "MANY STARS SEEN"
1241	OCT. 6	EUROPE (2) EGYPT	TOTAL: "WONDERFUL BLACKNESS"
1267	MAY 25	TURKEY	TOTAL: "MANY STARS APPEARED"
1275	JUNE 25	CHINA	TOTAL: "STARS SEEN"
1292	JAN. 21	CHINA	ANNULAR: "SUN A GOLDEN RING"
1406	JUNE 16	GERMANY	TOTAL: "GREAT DARKNESS"
1415	JUNE 7	EUROPE (2)	TOTAL: "ENTIRE SUN ECLIPSED"
1485	MAR. 16	AUSTRIA	TOTAL: NO OTHER DETAILS

TWENTY-EIGHT SOLAR ECLIPSES, either total or annular, are noted in records from the eighth century B.C. through the 15th century A.D. All but one of the 10 most ancient eclipses listed in this table were observed by Chinese astronomers. The later predominance of European eclipse observers is due to the rise of monastic communities. Multiple sightings are numbered.



EARLIEST SOLAR ECLIPSE for which the date is certain is recorded on this Assyrian tablet, a summary of past historical events. A large partial eclipse, it took place on June 15, 763 B.C.



LATE BABYLONIAN TABLET records new- and full-moon observations made between 323 and 319 B.C. Like the Assyrian tablet, it is among the ancient texts in the British Museum.

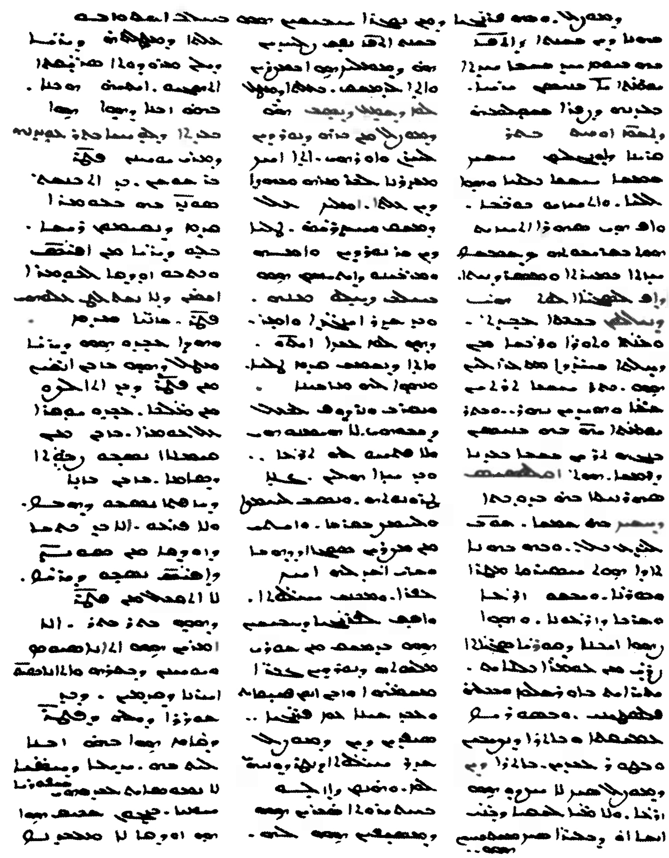
eclipses. The second consists of solar and lunar eclipses that are untimed but are reported as having come near sunrise or sunset. The third consists of solar eclipses that are untimed but are reported as large events: near-total or total.

As for the areas in which the observations were made, they begin with Babylonia. At least as early as 750 B.C. Babylonian astronomers began to take an active interest in the accurate observation of many celestial phenomena, including solar and lunar eclipses. We cannot say how much earlier observations may have been made in Babylonia or anywhere else simply because there is little by way of historical records from more ancient times. Systematic Babylonian observations continued from about 750 B.C. to at least 50 B.C. and possibly well into the first century A.D. This large corpus of astronomical data was well known to the ancient Greeks. The Greek astronomer Ptolemy of Alexandria, writing early in the second century A.D., claimed access to Babylonian eclipse observations going back to 747 B.C. It is a cause for regret that Ptolemy, in his *Mathematikē Syntaxis* (better known today by its Arabic name, *Almagest*), gives only brief extracts from the Babylonian record.

After about 300 B.C. Babylonia gradually decayed, and by A.D. 100 the city of Babylon was deserted. More than 1,700 years were to pass before the Babylonian astronomical tablets again came to light, accidentally unearthed by scavengers digging up the ancient baked clay bricks of the city to use in new construction. Most of the surviving astronomical tablets, many in fragmentary condition, are now in the British Museum. What was salvaged represents only some 5 percent of the original archive, and much of it is undatable. Only future excavation will determine whether more of the tablets remain to be found.

The traditional Babylonian practice was to measure the interval between the start of an eclipse and either sunrise or sunset with the aid of a water clock; the durations of the various phases were also usually measured. Time was expressed in units called *us,* the four-minute interval required for the celestial sphere to turn through one degree. It was also the Babylonian practice to estimate the degree of totality of an eclipse as a fraction of the solar or lunar diameter. Finally, if the sun or the moon rose or set while the eclipse was in progress, that fact was usually noted.

A typical eclipse entry, of 240 B.C., reads in translation as follows: "Month Eight, day 14, at three *uš* before sunrise [a lunar eclipse] began on the east side. [The moon] set eclipsed." The account, although brief, includes two vital pieces of information. First, the statement that



TWO HISTORICAL TEXTS that include references to eclipses of the sun are the official history of the Eastern Han dynasty (*left*) and the Chronicle of the Patriarch of Antioch, Michael the Syrian (*right*). The Chinese text describes solar eclipses of A.D. 118 and 120, the latter regarded as presaging the empress dowager's death in A.D. 122. The Syriac text describes in vivid detail a total eclipse in A.D. 1176.

the eclipse began only about 12 minutes before sunrise is highly accurate; it is in accord with the calculated clock error for 240 B.C. Furthermore, since the observation began so close to a well-defined moment, namely sunrise, there is little likelihood of timing error, always a serious problem with a primitive horological apparatus such as a water clock. Second, the statement that the moonset eclipsed counterbalances the possibility of a scribe's error in the time notation; some part of the eclipse must have been visible while part of the lunar disk was still above the horizon. Because local times of sunrise (and moonrise) and sunset (and moonset) can be computed with precision even for the distant historical past, being virtually independent of clock error, this second part of the observation gives a viable alternative to the first part if the first part proves to be defective. And because the text also gives the year of the observation according to the Seleucid calendar the Julian date can be readily deduced as November 3, 240 B.C. Modern calculation shows that an eclipse of the moon was indeed visible in Babylonia on that day. Such exactness in recording dates is characteristic of all Babylonian astronomical records.

The Babylonian records, impoverished as they now are, include some 40 useful observations of lunar and solar eclipses, both timed and untimed. Only a single total solar eclipse, however, was recorded. It is fortunate indeed that the British Museum collection contains two separate tablets describing the event. Its recorded date (the 29th day of the intercalary 12th month of the Seleucid year 175) corresponds to April 15, 136 B.C. One of the two tablets states that the eclipse was total and adds that it began 24 *uš* after sunrise, reaching totality after a further 18 *uš*.

The second tablet states that the duration of the entire eclipse was 35 *uš*, which is in close accord with the time to totality given by the first tablet. The scribe goes on to give additional descriptive details that make the text the most remarkable account of a total solar eclipse recorded before the 18th century. I owe the following translation to Abraham J. Sachs of Brown University: "Twenty-four *uš* after sunrise a solar eclipse [obscured] on the southwest side when it began.... Venus, Mercury and the normal stars [meaning those that were then above the horizon] were visible; Jupiter and Mars, which were in their period of disappearance [meaning the interval between setting with the sun and rising with the sun], were visible in that eclipse.... [The shadow] moved from southwest to northeast." Here, quite apart from the measured time of duration, the described visibility of planets and stars more than an hour and a half after sunrise is compelling evidence that the eclipse was indeed total, a very precise datum. The texts, in combination with the other Babylonian lunar and solar eclipse records, constitute by far the best collection of data bearing on changes in the rate of the earth's rotation from anywhere in the ancient world.

If Ptolemy's account of Babylonian observations is correct, the Near East has priority over the Far East in eclipse observations. For all its ancient historical tradition China is without a single reliable eclipse record before 720 B.C. Thereafter, over a period of 240 years, the astronomers of a single small state (Lu, the birthplace of Confucius) recorded 37 solar eclipses, of which three were said to have been total. This is a corpus that contains most of the known solar-eclipse observations from the entire world up to 480 B.C. The Chinese character used to identify totality, *chi*, was originally a pictogram of a man turning his head away from a plate of food, signifying that he was satiated;

Chinese astronomers continued to use it for another 2,000 years to describe total and annular eclipses.

The Lu records are particularly impressive for the accuracy of their dates; nearly all agree to the day with the dates arrived at by modern calculations. What is disappointing is that they omit information about the time of day and few of them give descriptive details; virtually all Chinese astronomical observations are deficient in these respects. Thus for the purpose of cross-checking long-term trends in the length of the day only the three total solar eclipses found in the Lu records are useful.

The tyrannical "burning of the books" by the founder of the short-lived Ch'in dynasty in 213 B.C. may be responsible for the scarcity of eclipse records from parts of China other than Lu. Immediately thereafter, however, with the beginning of the Han dynasty in 206 B.C., the situation changed dramatically. From that date until the fall of the Ch'ing (Manchu) dynasty in A.D. 1911 almost every kind of celestial phenomenon visible to the unaided eye in China was recorded.

The records make it clear that the main purpose behind all these observations was political astrology on behalf of the emperor and his family. For example, here is a description of a near-total eclipse on January 18, A.D. 120: "There was an eclipse of the sun. It was almost complete and on the earth it became like evening. [The sun] was 11 degrees in [the lunar mansion] (*Hsü-nu*). The woman ruler showed aversion to it. Two years and three months later, Teng, the Empress Dowager, died."

The imperial astronomical treatises give some 10 accounts of total and annular solar eclipses, the last two having occurred in A.D. 1275 (total) and A.D. 1292 (annular). The 1292 account, which states that "the sun looked like a golden ring," is one of the few clear descriptions of an annular eclipse from any part of the world before modern times.

To return to the West, at least to Alexandria, Ptolemy's *Almagest* records nine timed observations of lunar eclipses that were made by Greek astronomers in two brief periods: between 200 and 140 B.C. and between A.D. 125 and 136. With one exception the observers were at Alexandria. Their records were in general fairly crude, with the time of day given only to the nearest third of an hour or so, but taken as a set they constitute a useful supplement to the timed Babylonian and untimed Chinese observations. It is impossible to tell whether the nine observations were selected from a larger body of data or if they were all Ptolemy could find.

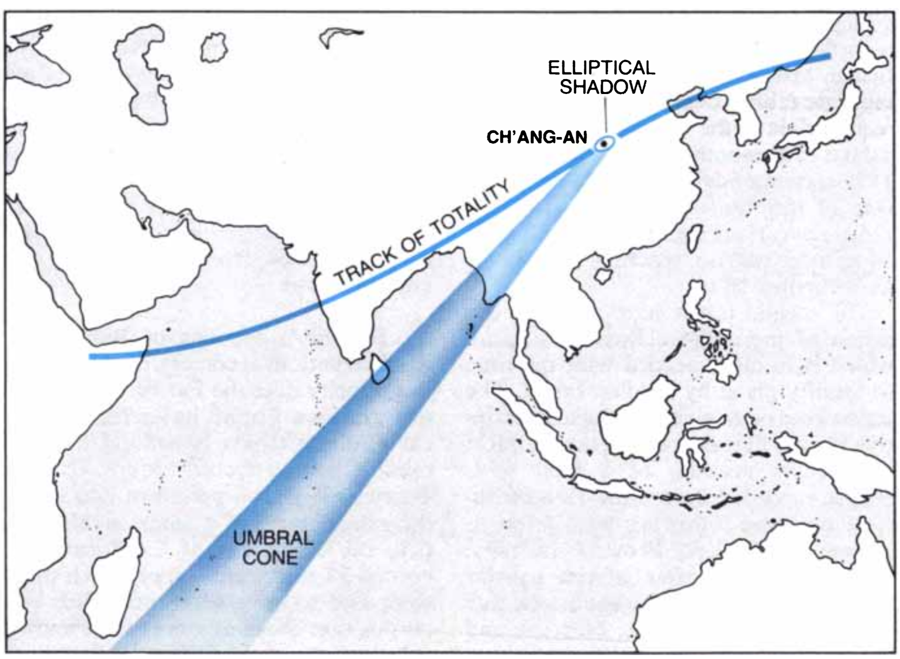
We have to come forward to medieval times to encounter further astronomically useful eclipse references, both from Europe and from the Arab dominions. In Europe an increasing number of monastic centers kept chronicles of major events, among them earthquakes, bright comets and eclipses of the sun and the moon. As a result more records of total eclipses of the sun are found in monastic chronicles than are found in all other sources before 1700 combined. As an example, the total solar eclipse of A.D. 1239 was noted by eight separate monastic communities ranging from Spain to Yugoslavia.

As with the near-total solar eclipse of A.D. 120 seen in China, the medieval record of a near-total eclipse can be of value. Here is the account of one such observation, made on August 2, 1133, at Prague: "An eclipse of the sun appeared in a wonderful manner; this defection gradually diminished so much that a crown like a crescent moon proceeded to the south part, which afterward turned round to the east, henceforth to the west. At length it was transformed to its original state." The account is somewhat garbled, but it seems clear that at no time was the sun completely covered.

Few monastic chronicles later than A.D. 1300 have been issued in compilations; the reason is they are too numerous for such publication to be practicable. As a result these later potential sources of European eclipse observations remain largely untapped.

In the Arab dominions scientific astronomy as we know it did not begin until about 150 years after the birth of Islam. The first leader to actively encourage astronomy was the caliph al-Mansur, who founded Baghdad in A.D. 762; indeed, the plans for the city were drawn up with the help of a Persian astronomer. Islamic interest in astronomy became high in Baghdad and Cairo and later at both extremes of the realm: Cordova in Spain and Samarkand in central Asia. Two different types of observations have come down to us from the Islamic world. On the one hand are carefully timed observations of solar and lunar eclipses made by skilled astronomers in well-equipped observatories. On the other are random sightings of near-total or total solar eclipses, recorded by various writers in much the nontechnical style of the European chronicles.

Numerous observatory sightings were made at Baghdad and Cairo between A.D. 800 and 1000. They were first studied by the American astronomer Simon Newcomb a century ago, and two of them will suffice to give some sense of the Arab astronomers' craft. In each case the altitude of the sun or the moon or a suitable bright star was measured to the nearest degree or half degree at the beginning and end of the eclipse. Thus for the partial solar eclipse of June 8, 978, observed at Cairo, the record reads: "Altitude of the sun when the eclipse began to be sensible to the view, 56 degrees approximately; altitude at the end, 26 degrees approximately." For a lunar eclipse of September 14, 927, observed at Baghdad, the record



TOTAL SOLAR ECLIPSE was noted by the court astronomers at Ch'ang-an (modern Sian), then the capital of China's Western Han dynasty, on March 4, 181 B.C. This observation is evidence that the earth's speed of rotation is gradually diminishing. Unless allowance is made for the resulting increase in the length of the day, a plot of this totality track would not pass over Ch'ang-an, as the Chinese historical records prove it did. Both the conical shape of the moon's shadow in space and the shadow's elliptical shape on the earth's surface are exaggerated.

A Special Profile by Andrew Hargrave

VEBA concentrating on energy.

VEBA is West Germany's leading industrial corporation in terms of sales. It plays a crucial role in the country's energy market. It is a major importer, processor, distributor and trader in oil and oil products and an important generator and supplier of electricity. It also includes major chemical as well as trading and transportation companies.

Rudolf v. Bennigsen-Foerder has been chief executive of VEBA for the past 11 years. In a recent interview he reflected on events in the last fiscal year and the corporation's future strategy.

According to Herr v. Bennigsen, the Group did well last year despite an environment of economic stagnation. External Group sales rose to almost DM 50 billion. The annual surplus amounted to DM 539 million. This performance permitted an unchanged dividend of 15 percent and a strengthening of reserves.

VEBA's relatively strong position, compared to various other German companies, is largely attributable to its electricity sector which has been allotted more than 50 percent of total investment funds in recent years. "In the most important sector of the Group, electricity, we are satisfied with business performance," says Herr v. Bennigsen. "This field is also the major contributor to the results of VEBA AG, the holding company."

The growing importance of nuclear energy

Herr v. Bennigsen has strong views on the desirability of increasing rapidly the nuclear element in the country's electricity generating capacity and has repeatedly suggested that public debate should be about the overall energy concept, not only its nuclear segment.

VEBA's utilities are grouped in two subsidiaries, the larger PREUSSENELEKTRA and VEBA

Rudolf v. Bennigsen-Foerder, Chairman

KRAFTWERKE RUHR (VKR). PREUSSENELEKTRA generated 43 percent of its electricity from a nuclear source. In the next couple of years there will be added three nuclear stations of 1,300 megawatt capacity each to the four already in service, and Herr v. Bennigsen is in no doubt about their contribution to the network. He insists that the quality and cost-effectiveness of power generation is as important as the availability of capacity itself: "If, for example, we are to use expensive German coal in electricity generation, then we need nuclear power to maintain electricity prices within reasonable limits." And while VEBA has proportionately more than 17 percent of nuclear capacity compared with 11.5 percent in the Federal Republic, it requires more for supplying the bulk of baseload with coal providing the middle load and a small input of gas at peak loads. Oil is being virtually eliminated as a fuel source.

In the case of coal, Herr v. Bennigsen envisages a step-by-step approach to improve the supplies for VEBA's power stations as well as for international trading operations. In June, a contract was concluded with the US coal company Westmoreland with 15 percent of its share capital going to VEBA. This stake can be increased to 20 percent.

More flexibility in the oil sector

Herr v. Bennigsen is less pleased with the situation in the oil sector. Results for 1981 and the first months of 1982 were unsatisfactory. Since May, however, an improvement has been discernible. Nevertheless, the role hitherto assumed by oil companies in securing the Federal Republic's long-term oil supply will have to be reviewed. At VEBA OEL this has led to a policy of increased flexibility. This means above all that contractual oil supplies will be largely replaced by more favorable purchases on the spot market. In addition to immediate steps to reduce losses, long range adjustments have been planned to improve results significantly, including further energy savings and increased flexibility in supply and refining.

One of these adjustments is an increase in the conversion rate within total refinery capacity to make operations more flexible by being able to switch to lighter distillates when necessary. This proportion is now 35 percent, well above the average for refineries in West Germany as well as in Europe as a whole; and it will be further increased after taking into operation a new hydrocracker being now under construction. Moreover, VEBA has, as a consequence of the BP deal in 1978, undergone structural changes, especially in refinery capacity and marketing, without which financial losses on these operations would have been much greater.

Restructuring in the chemicals sector

Decisive savings measures and restructuring are also being implemented in the chemicals sector.

VEBA AG had increased its equity stake to a majority holding in CHEMISCHE WERKE HÜLS some years ago, with the expressed objective of improving its product structure, switching from mass chemicals to more specialized and consequently valuable ones. In the case of mass produced plastic goods, which are particularly sensitive to economic cycles, this strategy has proved to be successful. The share of special items in the product range is 65 percent. However, earnings are unsatisfactory. This is also true for fertilizers. The other sectors are operating profitably. Sweeping measures to reduce costs were already introduced in 1981. In 1982, a supplementary rationalization program is being carried out. Herr v. Bennigsen is confident that these steps will lead to a positive development of results starting in 1983.

Trading satisfactory

The VEBA chief executive expressed satisfaction with developments in the Group's fourth major area of activity, trading and transportation: "Even if certain sectors experience difficulties in the second half of the year, we are reasonably confident that the projected results will be achieved and that they will make a stabilizing contribution to the overall results of the VEBA Group."

Confidence in future development

Herr v. Bennigsen firmly believes that VEBA is well-equipped to take advantage of opportunities provided by the international division of labor through its efficient trading set-up, its high level of technological knowhow and the cooperation and support it receives from the Federal Republic Government which is the Corporation's largest single shareholder.

VEBA - Energy is our Business.

VEBA was founded in 1929 as "Vereinigte Elektrizitäts- und Bergwerks-Aktiengesellschaft" in Berlin in order to combine the electricity and mining interests of the Prussian State. Until the mid-sixties it was a public-sector company. In 1965, VEBA was denationalized. The Federal Republic of Germany holds around 44% of the Company's equity, while the majority is in the hands of approximately 700,000 shareholders. The headquarters of the Company is in Düsseldorf.

The VEBA Group has around 83,000 employees and achieved a turnover of almost DM 50 billion in 1981. In terms of total sales, VEBA is therefore the largest company in West Germany.

The Company's main activities are electricity generation and supply, mineral oil, chemicals, as well as trading and transportation.

VEBA at a Glance 1981

Total external sales	million DM	49,428
Electricity generated	mill kWh	58,792
Natural gas production	mill kWh	2,539
Crude oil production	1,000 tons	2,052
Crude oil throughput	1,000 tons	11,890
Net income	million DM	539
Capital invested	million DM	2,277
Personnel		82,935

Electricity

Electricity is the main sector of the VEBA Group. PREUSSEN-ELEKTRA and its subsidiary NORDWESTDEUTSCHE KRAFTWERKE operate 36 power stations. These supply electricity for about 12 million people living between Flensburg and Frankfurt.

The second important VEBA energy company, VEBA KRAFTWERKE RUHR, operates 8 power stations, mainly supplying wholesale users with electricity. The pit coal in the Ruhr area provides 95% of the fuel for electricity generation.

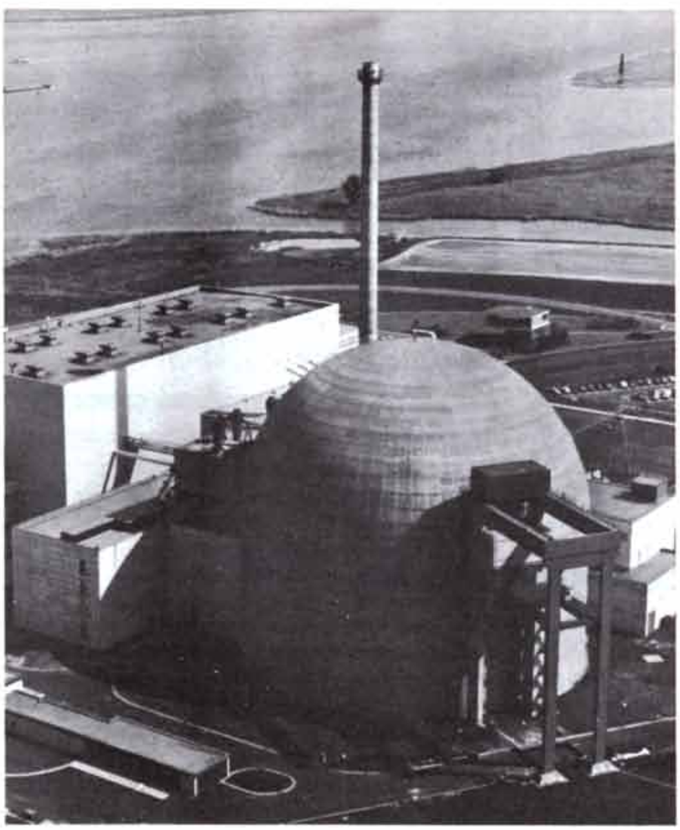
The two companies maintain a network of more than 100,000 kilometers for electricity distribution.

Petroleum

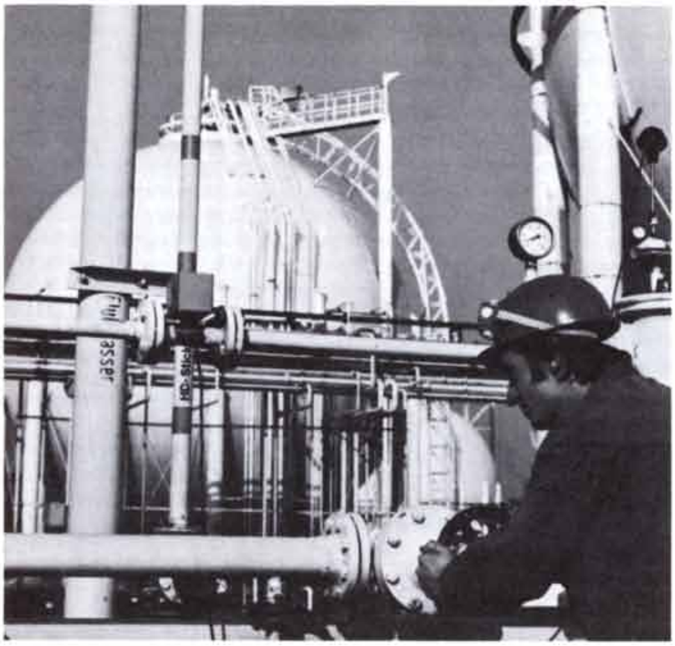
VEBA OEL is active in the total field of mineral oil: from exploration and production to transportation and processing for consumer supply.

Exploration and production of crude oil are realized by the company itself and by DEMINEX (VEBA share: 54%) in many countries.

Company-owned tankers carry the oil to the large European pipeline systems, in which VEBA has a stake. VEBA OEL has above average conversion capacity in refineries, and is very active in petrochemicals as well.



A GLOBAL PERSPECTIVE



The products of VEBA OEL are mainly sold through its own trading company RAAB KARCHER as well as ARAL (56% participation), Germany's foremost gasoline station network.

Chemicals

The VEBA chemical interests are concentrated at CHEMISCHE WERKE HÜLS.

This company produces more than 1,000 different products in its factories in West Germany. The program comprises chemical raw materials and inorganic chemicals, organic chemicals and detergent raw materials, lacquer raw materials, refined chemicals, plastics, latex and synthetic rubber as well as fertilizers and agrochemicals. In plastics and detergent raw materials, CHEMISCHE WERKE HÜLS ranks among the most important manufacturers in Europe.

Product distribution is achieved through its own worldwide sales organization.

Trading and Transportation

The most important trading and transportation company within the VEBA Group is STINNES. About 80% of this company's sales volume is in the trading sector, approximately 20% is in transportation.

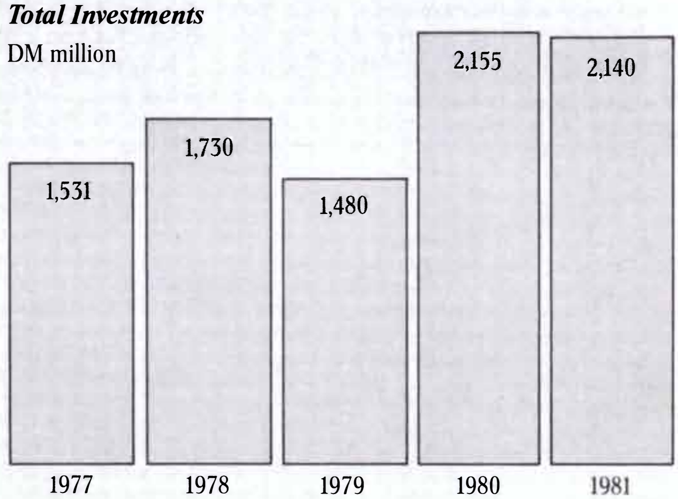
The main trading activities are international fuel business, sales of building materials, chemical products, ores and minerals, steel products, technical equipment and consumer goods.

An additional important trading company of the VEBA Group is RAAB KARCHER, which mainly concentrates on the distribution of VEBA OEL products but also sells products of other companies.

The broad diversification of VEBA with its multiple activities, especially in the energy sector, gives the Company a solid base for further growth as well as greater internationalization of the Group.

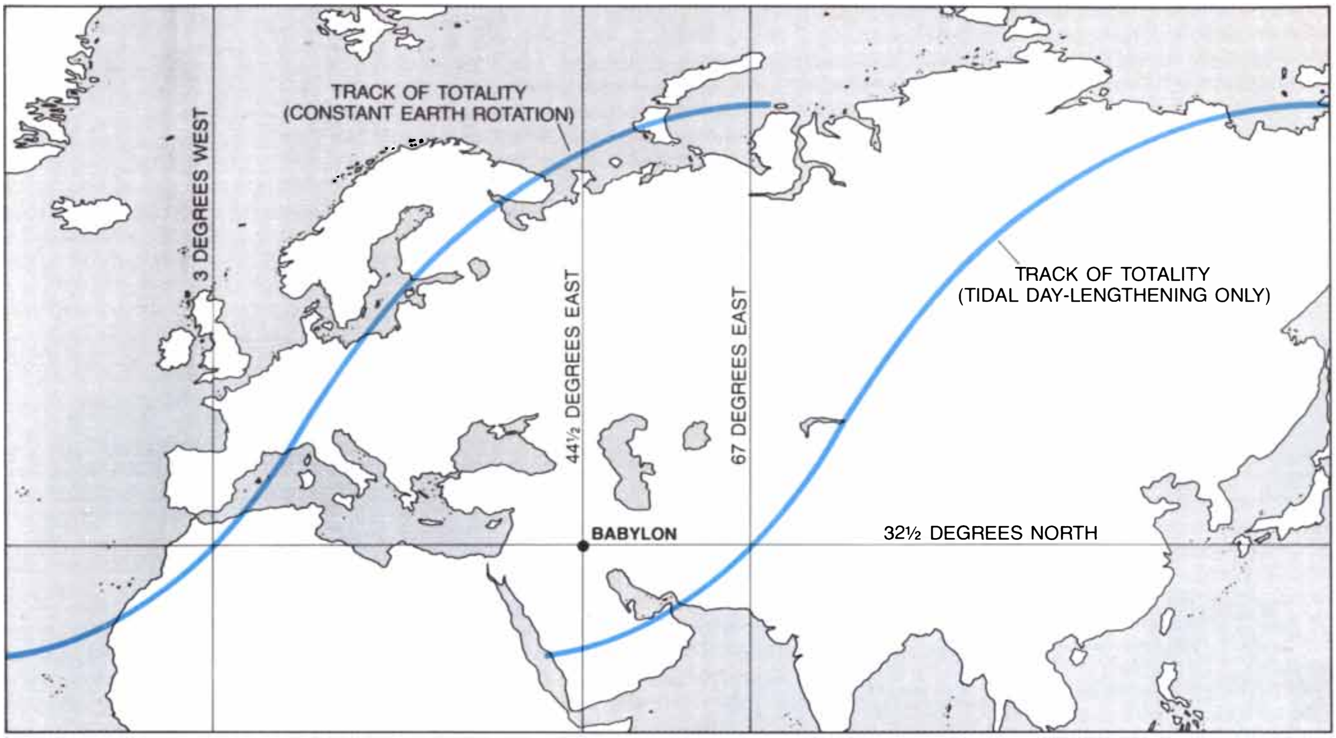
Total Investments

DM million



VEBA

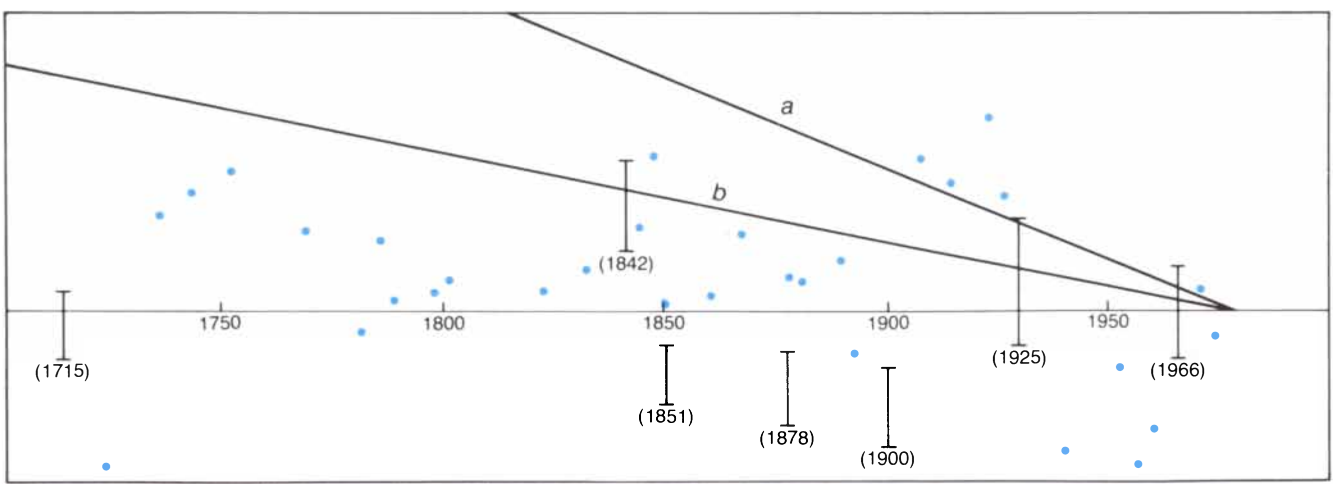
Karl-Arnold-Platz 3, 4000 Düsseldorf 30, West Germany



DUAL INFLUENCES on the net rate of the slowing rotation of the earth are presented in a plot of the total solar eclipse of April 15, 136 B.C., noted in Babylonian astronomical records. If the rate of the earth's rotation had been then what it is now, the totality track of the eclipse would have fallen far to the west of Babylon, missing the city by more than 47 degrees of longitude. If no factor other than tidal friction were involved in altering the rate of rotation, the track would have been displaced to the east of Babylon by more than 22 degrees. Because in fact the eclipse was total at Babylon some acceleration in the rate of rotation must be partly counterbalancing the tidal effect.

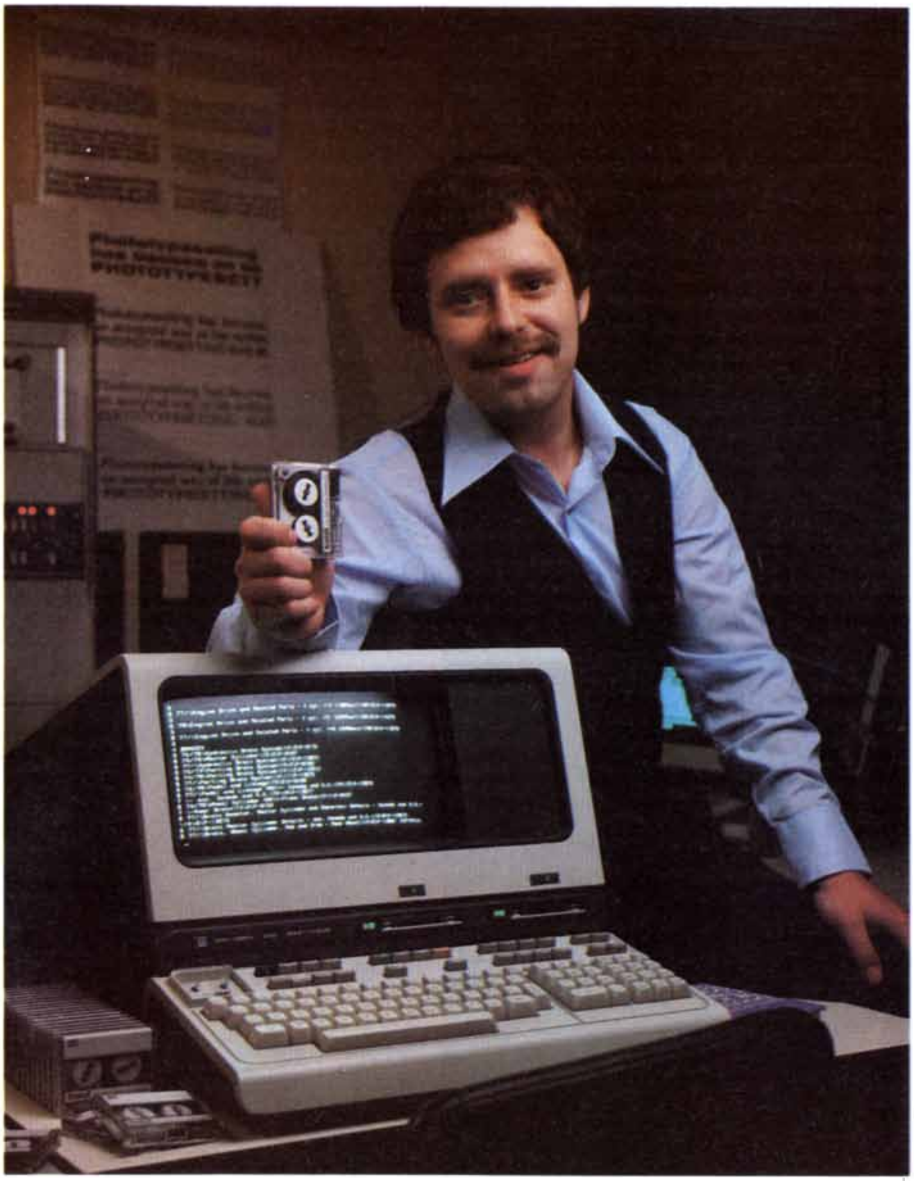
reads: "Altitude of Sirius at the beginning, 31 degrees in the east; revolution of the celestial sphere from sunset until the commencement of the eclipse, determined with the astrolabe, about 148 degrees." It is of interest that the determination of altitude rather than the direct measurement of time remained a European astronomical practice in observing eclipses and other events throughout most of the 17th century; the practice ended only when reliable clocks became available.

Known Arabian sightings of near-total and total eclipses of the sun are few. It is probably because the Arab chronicles, many of which exist only as manuscripts, have not yet been systematically searched. A recent study by Bernard R. Goldstein of the University of Pittsburgh has uncovered a brief but precise



HYPOTHESIS OF A SHRINKING SUN, proposed in 1979 following analysis of thousands of right-ascension measurements, is contradicted by measurements based on the observed duration of 30 transits of Mercury (*colored dots*) and six total solar eclipses from A.D. 1715 to 1925. The two slopes indicate where past measurements of the sun's diameter should have fallen if indeed it has been shrinking at a rate of one second of arc per century (*a*) or at half that rate (*b*). Typically each of the eclipses from 1715 to 1925 has been timed by some 10 to 20 observers and the transits of Mercury often by more than 100 observers. The means of the eclipse measurements are entered for the appropriate year; the error bars represent a 95 percent level of confidence. Analysis of both series of observations supports the conclusion that the sun's apparent diameter is changing at a rate of minus .16 ± .14 second of arc per century, essentially a null result. Substantial variations are evident, however, implying that on a time scale of decades fluctuations in the diameter of the sun do take place.

"Our reputation rests on digits, decimal points, and details. We wouldn't trust them to anything less than Scotch® Brand Data Cartridges."



Bill Birkett, Vice President, Trade Graphics, Inc., Livonia, Michigan

The unique design of a data cartridge provides great reliability, high storage capacity and long tape life. And where could you possibly get better data cartridges than Scotch Brand, made by 3M, the people who invented the data cartridge system itself?

3M controls every step in manufacturing. Top quality magnetic tape and precision components are part of every Scotch Data Cartridge. Over twenty-five years of service to the computer industry assure you of the utmost reliability.

Scotch Data Cartridges are available in miniature DC 100A, the standard-size DC 300A and now, an extra-length DC 300XL with 50% more storage capacity. They are compatible with most cartridge systems including Hewlett-Packard, IBM, NCR, Tektronix and TI.

To find out where you can find Scotch Data Cartridges or virtually any other data recording medium, call toll-free: 800-328-1300. (In Minnesota, call collect: 612-736-9625.) Ask for the Data Recording Products Division.

If it's worth remembering, it's worth Scotch Data Recording Products.

3M

INNER PART of the sun's corona, normally invisible, rings the circumference of the moon in this photograph of the eclipse of February 16, 1980. Flecks of color are solar prominences.

TWO COMPONENTS of the earth's shadow appear in this photograph of the total lunar eclipse of January 9, 1982, as totality was advancing. What appears to be a crescent missing from the moon's edge is the umbral sector; the lighter band beyond is the deep penumbral sector.

account of an annular eclipse. The account was mentioned in one of the treatises of the Arab astronomer al-Biruni and relates to the eclipse of July 28, 873, as seen at Nishapur in Persia. Al-Biruni's entry states: "[The astronomer al-Iranshahri] recorded that the body of the moon was in the middle of the body of the sun such that the light of the remaining portion of the sun surrounded it uneclipsed; thus it is clear that the apparent diameter of the sun exceeds that of the moon." The last clause, an interpolation by al-Biruni, is of course wrong, but Tycho Brahe later arrived at the same mistaken conclusion quite independently.

As these various examples, drawn from records over the past 2,500 years, make clear, there is a rich store of both solar and lunar eclipse observations that can be utilized in the study of the earth's rotational history. Nevertheless, these data are far from ideal. For instance, there is a gap between A.D. 100 and 800 and little usable material between A.D. 1300 and 1600. Hence it is not at present possible to trace a continuous history of the variations in the rate of the earth's rotation. How much can be done?

To analyze an eclipse observation it is necessary to compute the celestial coordinates of both the sun and the moon with high precision. With the sun the computation presents little difficulty. Toward the end of the 19th century Newcomb produced a detailed theory of the motions of the earth and the other inner planets around the sun; his theory is still used in computing the data in *Astronomical Almanac* and similar ephemerides. When the apparent positions of the sun as seen from the earth, computed from Newcomb's theory, are checked against the positions computed from a dynamical, integrated ephemeris, they show no significant differences throughout the historical period.

With computing the celestial coordinates of the moon, however, there are problems. Modern lunar theory, largely the work of Ernest W. Brown of Yale University at the turn of the century, would be able to provide a highly accurate lunar ephemeris for any day in the historical past if only the moon's rate of recession from the earth could be determined precisely.

This may eventually be possible, but astronomers today are unable to compute the coordinates of the moon 2,000 years ago to closer than a tenth of a degree. That rather large uncertainty upsets calculations bearing on the changing rate of the earth's rotation in the past. It does not, however, prevent ancient eclipse data from yielding useful information.

Consider the single total solar eclipse noted in the Babylonian records: the eclipse of 136 B.C. That is possibly the most reliable of all ancient eclipse observations. Neglecting the moon uncertainty, how fast would an ideal clock, started in 136 B.C., now be compared with an earth clock? The clock error would be in the range of 3.13 to 3.38 hours; any value outside that range would make the 136 B.C. eclipse only a partial one at Babylon, which would contradict the historical record.

Taking all the ancient data as a group (mostly Babylonian observations but including a much smaller number of Chinese ones), the average rate of day lengthening since ancient times is $1.78 \pm .11$ milliseconds per century. (The allowance for observational error is as small as $\pm .02$ millisecond per century.) When the medieval data are similarly grouped, the average value since A.D. 1000 is significantly smaller: $1.45 \pm .15$ milliseconds per century. These figures compare with the tidal result of 2.51 milliseconds per century.

As tidal friction is slowing down the earth's spin other processes act to speed it up. Sea level has an influence. Some scholars suggest that the planet is contracting; others suggest that the core is expanding, because of continued heating by the radioactive isotopes in it. Either process would diminish the planet's moment of inertia, accelerating its rate of rotation and thereby decreasing the length of the day. The average rate of day lengthening since ancient times would have been even greater except for this accelerating influence. Furthermore, the smaller average rate of lengthening indicated by the medieval eclipse data is due to an approximate doubling of the acceleration component in more recent centuries. Thus it would seem that over a period of some thousands of years the acceleration of the earth's spin has been rather variable.

In default of any other adequate explanation, the nontidal tendency toward a shorter day length may be attributable to a gradual decrease in the earth's moment of inertia, whether it is due to sea-level changes, to the planet's shrinkage or to the redistribution of material in its interior. To judge from the diminishing millisecond-per-century values for day length, the rate of diminution in the moment of inertia in recent centuries is about twice what it was in ancient times. Now, as simple a change as a 1.2-meter fall in the average sea level over the past 2,500 years could have produced the earlier, lower rate of diminution. Better historical sea-level data than those now available may eventually help to determine the actual importance of the other proposed causes for the rate change. In any event the data contained in eclipse observations of the past may well make significant contributions to a solution of these classic problems of geophysics.

THE AMATEUR SCIENTIST

Delights of the "wobbler," a coin or a cylinder that precesses as it spins

by Jearl Walker

Set a coin on its edge on a smooth surface and flick it with your finger. It will spin vertically but soon will start to wobble and tilt until, with a clatter that increases in frequency, it ends up flat on the surface. A cylindrical object such as a bottle can also be made to wobble on its base, but unless the wobble is so strong that it makes the object fall over it remains upright. A bit of experimenting reveals a lot about how such objects wobble and what determines whether a wobbler finishes flat or upright. My analysis is based on a study by Lorne A. Whitehead and Frank L. Curzon of the University of British Columbia that will appear in *American Journal of Physics*. Whitehead designed an apparatus for studying the phenomenon and initiated the mathematical analysis. I have also drawn on earlier work by Martin G. Olsson of the University of Wisconsin at Madison.

Wobbling has three noteworthy features. First, it is periodic. Second, the object rolls around on its supporting surface without slipping. Third, the center of mass of the object gradually moves downward but does not shift much horizontally unless the object has been started into a clumsy spin.

Wobble lasts for as long as it does because two losses of energy are minimized. The lack of slippage diminishes the loss of energy to friction with the surface on which the object is spinning. The lack of vibration of the center of mass and of the support point eliminates an additional loss of energy.

The top illustration on page 188 shows the main physical characteristics of a wobbling object. The object is in contact with the supporting surface at a single point. An axis of symmetry extending along the length of the object is usually at an angle to the vertical. As the wobbler spins around this axis the axis turns around the vertical. Since the center of mass is stationary as the object rotates, the bottom of the wobbler rolls in a circle around the vertical axis passing through the center of mass. This is precession, the movement usually seen when a toy top has been set spinning.

The force that generates precession is developed at the point where the wobbler touches the surface on which it is turning. Although each part of the wobbler is pulled downward by gravity, the combined weight can be assumed to act through the center of mass. If the wobbler were not also spinning around its axis of symmetry, the pull on the center of mass would cause the object to fall over.

A force usually makes an object accelerate in the direction of the force, but a force acting on something that is spinning creates a torque that makes the object turn. The torque on a wobbler causes precession. If the rate of rotation is too low or too high for the object to roll without slipping, friction quickly increases the rate of spin until the slipping stops.

What the torque is actually doing is redirecting the angular momentum of the wobbler. This quantity is the product of the object's distribution of mass (its moment of inertia) and its rate of spin. The angular momentum can also be described as a vector that lies in an imaginary vertical plane containing the axis of symmetry around which the wobbler is spinning. Rotating about the vertical axis that passes through the center of mass, the vector maps out an imaginary cone. Hence the axis of symmetry also rotates about the vertical. Lacking gravity and the torque it generates, an object could be made to spin without precessing. A wobbler then would not wobble.

A wobbler's rate of precession depends in part on the angle between the axis of symmetry and the vertical. When Whitehead and Curzon analyzed this relation, a surprise emerged. In general a wobbler will precess steadily at any angle of inclination between approximately zero (with the axis of symmetry nearly vertical) and 90 degrees (with the axis nearly horizontal). The surprise was that the wobbler will not precess steadily in a certain zone that is determined by its shape. The forbidden zone is easy to determine if the object rolls around on a circular rim, as coins and bottles do, but is hard to determine if the shape is more irregular.

Whitehead and Curzon worked with a cylinder in their study of the requirements for steady precession. The range of inclination angles at which a cylindrical object cannot precess steadily depends on the shape of the cylinder. In particular the ratio of half of the cylinder's length to its radius is crucial. A coin is a cylinder in which the ratio is small; with beer cans and most other cylinders the ratio is large.

The results of the analysis are shown in the graph at the bottom of page 188. The ratios of half length to radius are on the horizontal axis and the inclination angles are on the vertical one. For example, the inclination angle of a coin is small when the coin is nearly flat and large when it is on or near its edge.

The cylinder precesses steadily when its axis of symmetry is outside the forbidden zone. That zone is marked by the angles θ_1 and θ_2. The angle is θ_1 when the cylinder's center of mass is directly above the support point, in which case gravity provides no torque for the precession. At θ_2 the angular-momentum vector is vertical and torque again does not cause precession.

Ordinarily these angles differ. They are the same only for a cylinder with a half-length-to-radius ratio equal to $\sqrt{3}/2$ (half of the square root of 3). A cylinder with a smaller ratio precesses more like a coin and would be recorded on the left side of the graph. Long cylinders such as cans precess in a way that is reflected on the right side.

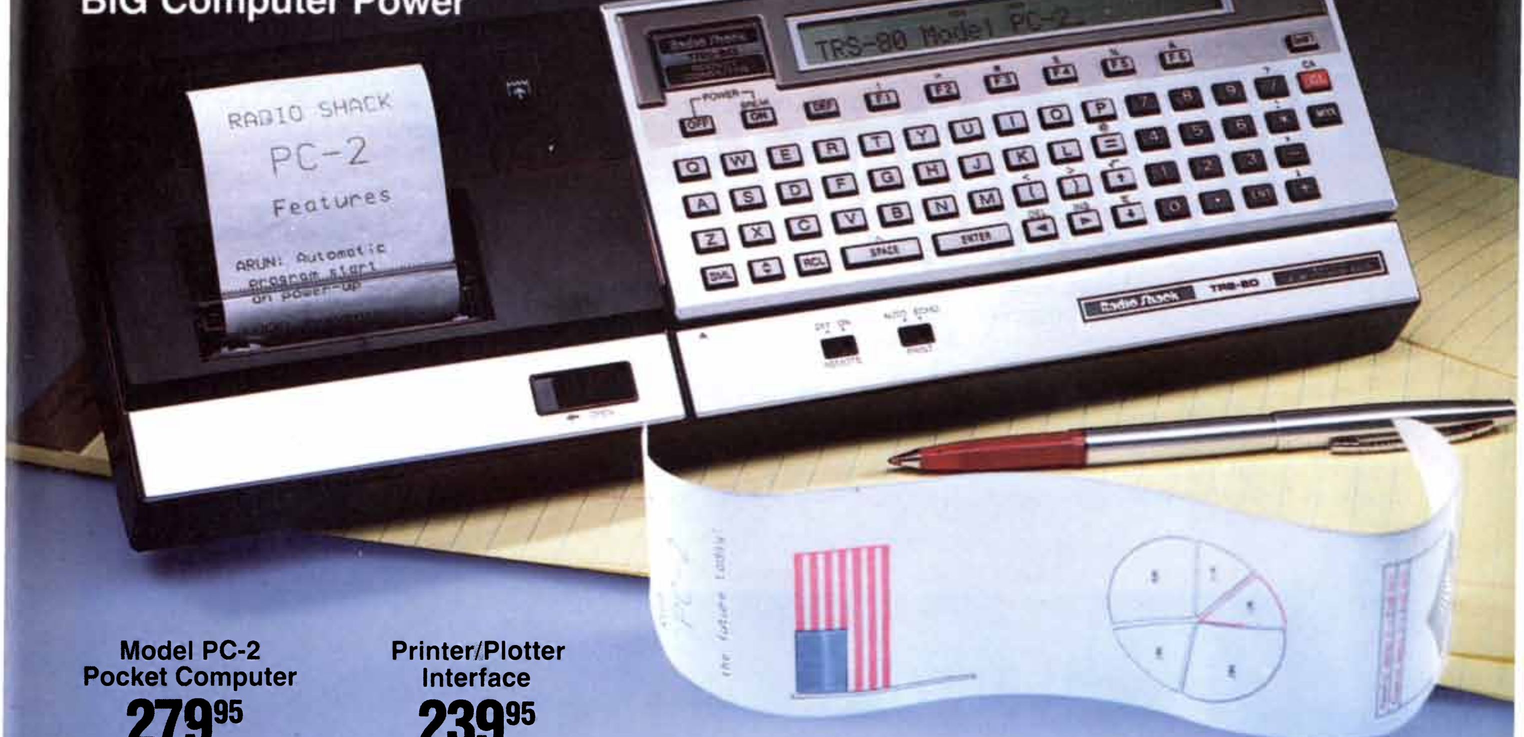
A coin spun on a surface begins to wobble in a steady precession when the conditions are as indicated in section *a* of the graph. As it gradually loses energy to friction its angle of inclination decreases. Initially the rate of precession decreases too, but when the angle of inclination approaches zero, the precession rate begins to rise. You can detect the change with your ears as well as your eyes. As the angle of inclination decreases, the clatter of the coin first drops in frequency (reflecting the initial decrease in the rate of precession) and then rises.

If you are looking down on the coin, you can also monitor the spin rate. At first it is too high for you to be able to see the face of the coin clearly. As the coin approaches its final horizontal position the spin rate diminishes and the face of the coin becomes clearer. The increasing clatter from the increasing precession rate contrasts sharply with the decreasing spin rate.

How can the precession rate of a wobbling coin increase while the coin is losing energy to friction? The coin has

Radio Shack's Astounding PC-2 is What Every Professional Needs!

The Hand-Held Computer With BIG Computer Power



Model PC-2 Pocket Computer
279^{95}

Printer/Plotter Interface
239^{95}

Out in the field, in meetings, on board airplanes—you need a real computer to get the job done. One with power and features like your micro back in the office. You need Radio Shack's Pocket Computer Model PC-2.

You Need Power. The PC-2 delivers it. RAM is 2640-bytes big—expandable to over 10K. Our phenomenal Extended Pocket BASIC programming language allows comprehensive handling of string variables up to 80 characters long, 2-dimension arrays, 42 statements, 34 functions, 6 commands and 10 digit accuracy.

You Need Speed. The PC-2's 8-bit CPU operates at a clock speed of 1.3MHz—as fast as many desktop computers. Its advanced design allows direct memory access, and maskable and timer interrupts.

You Need Features. PC-2 has a built-in real-time quartz clock, accessible from either BASIC or the keyboard. The 26-character Liquid Crystal Display can produce full upper and lower case alphanumerics, special engineering and scientific characters or 7x156 dot matrix graphics. There's even a programmable beeper.

You Need Peripherals. Add our unique 4-color Printer/Plotter/Dual Cassette Interface and expand your PC-2's usefulness even further. The Printer/Plotter gives you outstanding hardcopy—including full X/Y/Z axis graphics. Text can be printed in nine different character sizes. And you can add two cassette recorders for automatic overlays, chaining and data storage under program control.

Need more memory? It's easy. Just plug a 4K or 8K RAM Module into the special slot in the back.

You Need to See It to Appreciate It. You can try out the amazing TRS-80 PC-2 at a Radio Shack Computer Center, store or participating dealer near you. Stop by today.

Radio Shack
The biggest name in little computers™
A DIVISION OF TANDY CORPORATION

Send me your free TRS-80 Computer Catalog today!
Mail to: Radio Shack, Dept. 83-A-270
300 One Tandy Center, Fort Worth, Texas 76102

NAME ______________________________

ADDRESS ______________________________

CITY ______________ STATE ________ ZIP ________

Retail prices may vary at individual stores and dealers.

three types of energy. Two are rotational kinetic energies associated with the spin around the axis of symmetry and the precession around the vertical. The third type is gravitational potential energy. Every part of the coin except the point in contact with the table has potential energy, but the situation is easier to understand if the entire amount is associated with the center of mass.

The coin has to tilt increasingly as it wobbles because friction with the surface gradually removes the energy of the spin around the axis of symmetry. With a smaller rate of spin the axis of symmetry must move closer to the vertical. As a result the center of mass drops closer to the surface and the potential energy of the coin decreases. The loss of energy (both kinetic and potential) forces the coin to precess faster, somewhat like a rubber ball that bounces more frequently as it loses energy.

In theory you could also make a coin spin in the conditions represented by section *b* of the graph, but it would be difficult. The coin begins to spin while it is nearly vertical. As it loses energy to friction its angle of inclination increases to 90 degrees. In principle the coin would come to rest balanced on its edge. This position is unlikely in practice because the coin is too sensitive to small perturbations. It would probably fall over and lie on the surface.

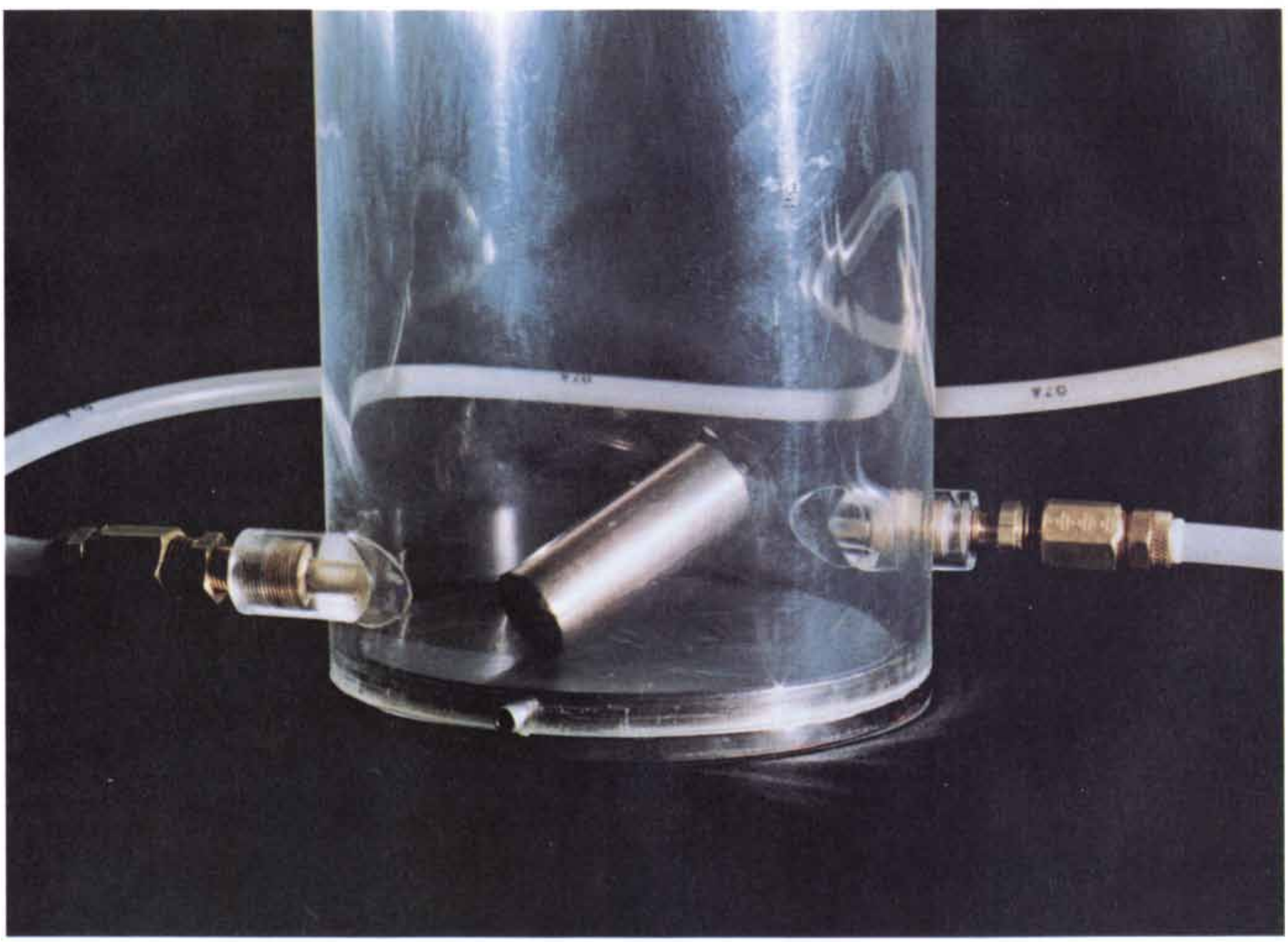
A cylinder with a ratio of half length to radius of $\sqrt{3}/2$ shows the greatest range of inclination angles at which it can precess smoothly. Longer cylinders such as bottles wobble in two different ways, which are represented by regions *c* and *d* on the graph. A cylinder set wobbling when it is nearly upright performs as shown in region *c*. You could start a bottle in this way by putting one hand on each side of it while it is inclined almost at the balancing angle of θ_1 and rapidly moving your hands in opposite directions.

The bottle rolls on its rim. It gradually speeds up, and the angle of inclination decreases as the bottle approaches the vertical. Since the bottle is tall, it can be affected by nonuniformities in the rim and the table. The wobble is usually unstable.

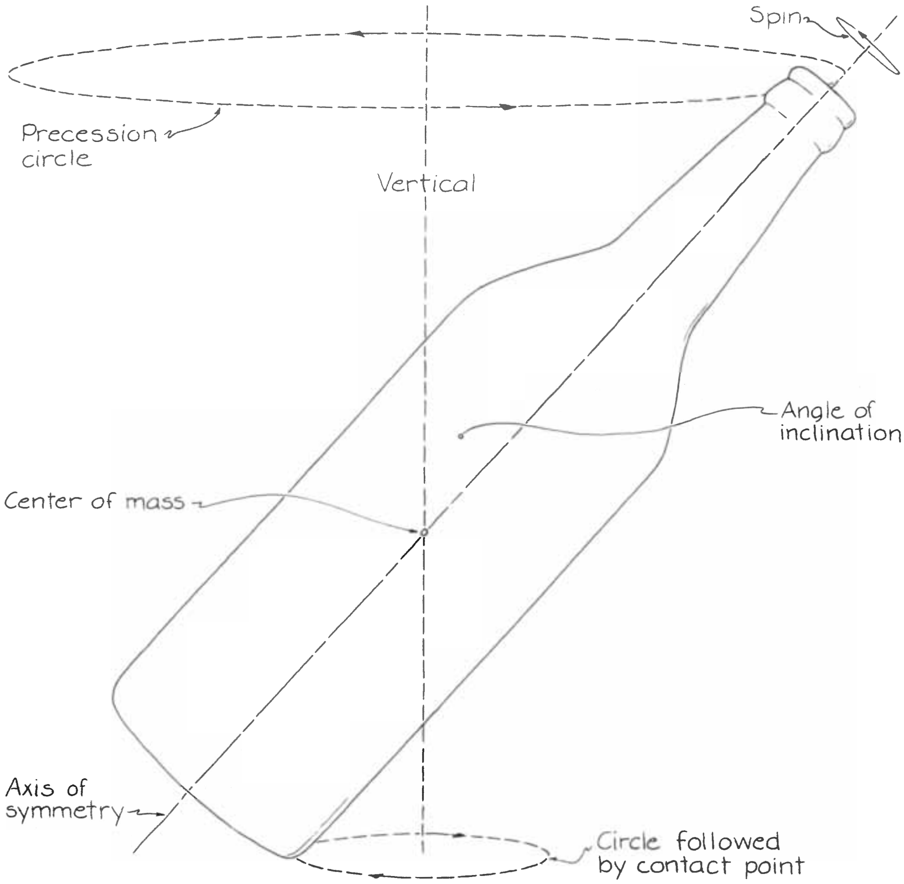
A cylinder wobbling under the conditions represented in region *d* begins with its axis of symmetry almost horizontal. The angle of inclination may be just slightly larger than θ_2. As with a coin the angle of inclination cannot be exactly θ_2 because the precession and spin would be infinitely fast. In losing energy the cylinder begins to lie down and its rate of precession decreases steadily. When it lies flat, it still has in principle some minimum rate of precession, but in fact it is quickly stopped by friction with the surface.

You can start a small cylinder in this type of motion with a snap of a finger. The cylinder precesses smoothly and gradually descends to the surface, where it rolls about more irregularly. Its clatter steadily decreases in frequency.

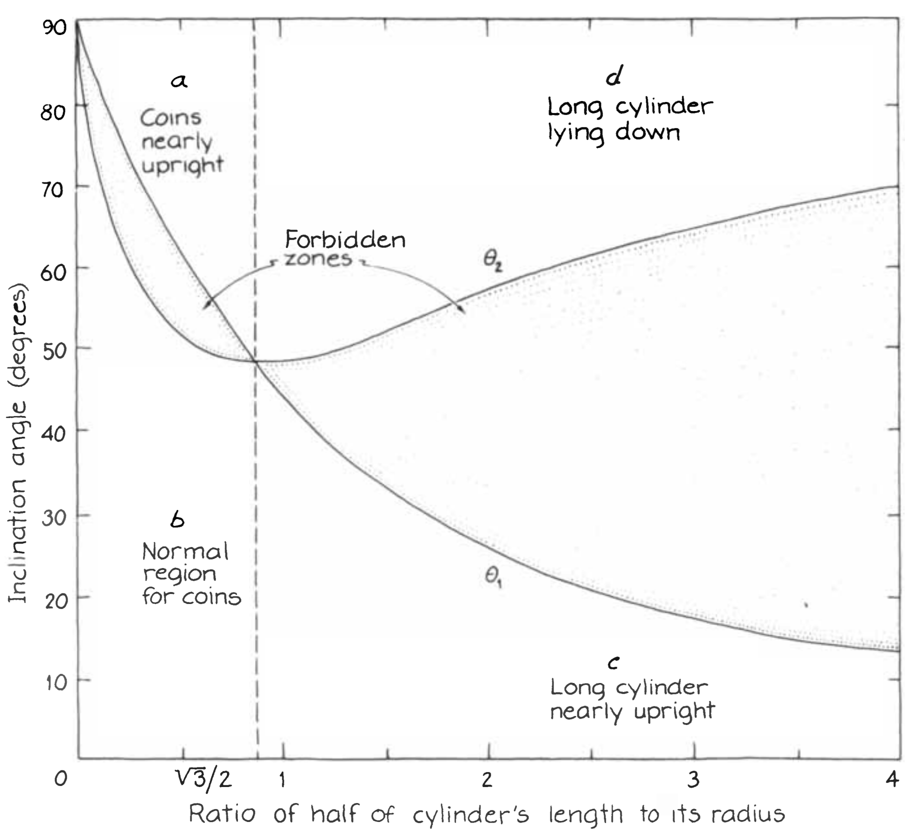
Whitehead realized early in his work that the wobble of a cylinder could be prolonged with jets of air. He rigged an apparatus in which jets of air were directed tangentially at a cylinder of aluminum eight centimeters long and three centimeters in diameter. Launched by hand at a large angle of inclination, the cylinder wobbled in a way represented by region *d* of the graph. Precession frequencies of up to 100 cycles per second



An aluminum cylinder wobbling in a chamber designed by Lorne A. Whitehead



The physical characteristics of a wobbling object



The conditions for steady precession

could be achieved by adjusting the jets. Although the motion was stable, the cylinder had a tendency to drift across the table. Whitehead therefore set out to build a better apparatus.

The result is a container with a Plexiglas base 15 centimeters in diameter. Whitehead machined the base so that it is concave; its radius of curvature is 50 centimeters. The top surface is polished to eliminate rough features that would interfere with a stable wobble. A layer of rubber an inch thick is glued to the surface to provide better resistance to the air. The concave surface keeps the wobbler from straying from the center.

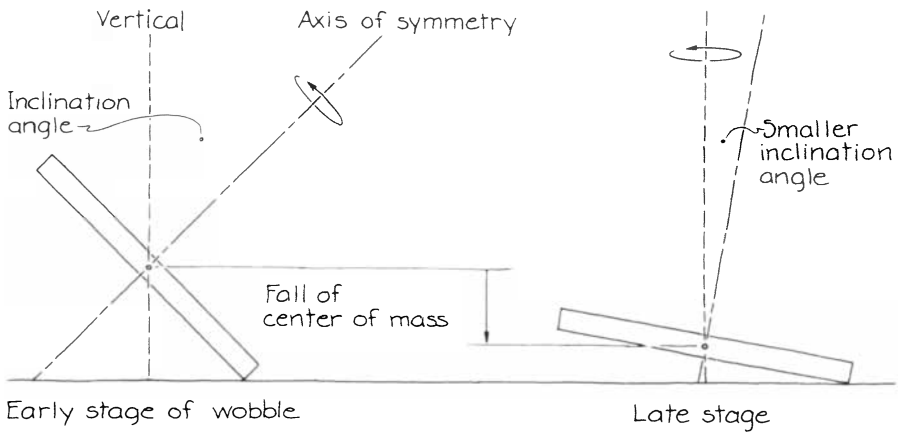
Air jets from a standard laboratory supply of air are directed into the container near the base. They maintain the energy of the wobbler, overcoming its losses to friction. The air goes out of the container through the top, which is a flat piece of Plexiglas with holes in it.

The main problem faced by Whitehead was how to freeze both the precession and the spin of the wobbler. Usually the two proceed at different rates. The solution was to deploy a wobbler in which the spin rate is an integral multiple of the precession rate. If a stroboscope is then set to flash at the precession rate, both the precession and the spin are frozen.

Whitehead chose as a wobbler a carefully machined cylinder of aluminum with a half-length-to-radius ratio of 2.71. When such a cylinder wobbles at an angle of inclination of 64 degrees, its spin rate should be twice as high as its precession rate. That angle is close to θ_2 for the cylinder. By adjusting the rate of flow in the air jets he could achieve fine control of the angle of inclination and thereby of the precession rate. The stroboscope was operated at a frequency of 50 hertz.

Whitehead had to remachine the cylinder a bit to achieve the desired ratio of 2 in its precession and spin rates. A precision of about .1 millimeter was necessary. When the cylinder was right, he set it in motion by hand to wobble under the conditions of region *d*. It quickly developed a stable wobble near the desired angle of inclination. Whitehead adjusted the flow of air to maintain the angle. Under these conditions the wobble could be maintained for three or four days. A cylinder of this kind could be kept wobbling longer if the concave base of the apparatus were made of a material harder than Plexiglas. To keep wobbling smooth on Plexiglas calls for an occasional repolishing of the surface.

Olsson has studied the wobble of a coin, analyzing the motion with methods that could also be applied to a child's top. The two objects are actually quite similar. When Olsson sets a coin spinning on its edge, the spin is at first stable against perturbations from the ta-



Stages in the wobbling of a coin

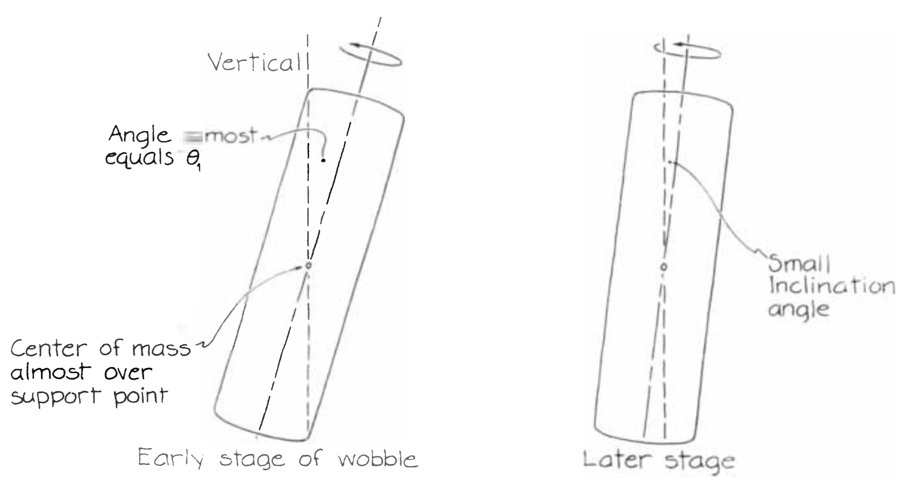
ble and from gravity. Friction gradually slows the spin, which soon drops below a critical value determined by the coin's mass, radius and moment of inertia. Then the coin begins to wobble.

Olsson also demonstrates wobbling in a large-scale way, spinning an aluminum disk that is an inch thick and about as big as a manhole cover. It wobbles with quite a noise, and toward the end, when the disk is almost flat, the racket is impressive. In a classroom where the seats are on a sloping floor the students can both watch the spin rate and hear the precession rate. The contrast between the decrease of the one and the increase of the other is mesmerizing.

In October, 1979, I described the rattleback, a toy that resists spinning in one direction but spins easily in the other. Many readers have asked where such a toy can be bought. It is now being sold by Toltoy, Inc. (5439 Schultz Drive, Sylvania, Ohio 43560), under the name of Space Pet.

Last December I examined the phenomenon of bubbles in beer and other carbonated drinks. Many people have written to me amplifying my explanation of why bubbles starting at the bottom of a glass of beer grow larger as they rise. I attributed the expansion to reduced hydrostatic pressure. My critics point out that the bubbles become larger than they would if reduced hydrostatic pressure were the only thing affecting them and that the bubbles must also be collecting dissolved carbon dioxide gas as they rise.

I erred too in stating that gas bubbles originate only on the sides of the glass where pockets of air or carbon dioxide gas already adhere to crevices. Such pockets do serve as nuclei for the formation of bubbles, but another important source is pointed out by Thomas D. Kunkle of the Los Alamos National Laboratory. The liquid contains microscopically small regions of gas left over from the cavitation that resulted when the beer was poured into the glass. I would have thought that the molecules of gas in such regions would quickly diffuse into the liquid, where they would dissolve. According to Kunkle, however, the regions are stabilized against diffusion by surface-active molecules that bond together to form a "skin" one molecule thick. These regions of gas then act as nucleating agents for the formation of bubbles as more gas comes out of solution.



Stages in the wobbling of a long cylinder

Rinehart S. Potts of Glassboro, N.J., asked why one should avoid cleaning beer glasses with a detergent. It matters only if you want lots of bubbles in the beer after the cavitation from pouring the drink has subsided. A detergent is likely to coat the surface of the glass, and so the crevices there are less able to serve as nucleating agents.

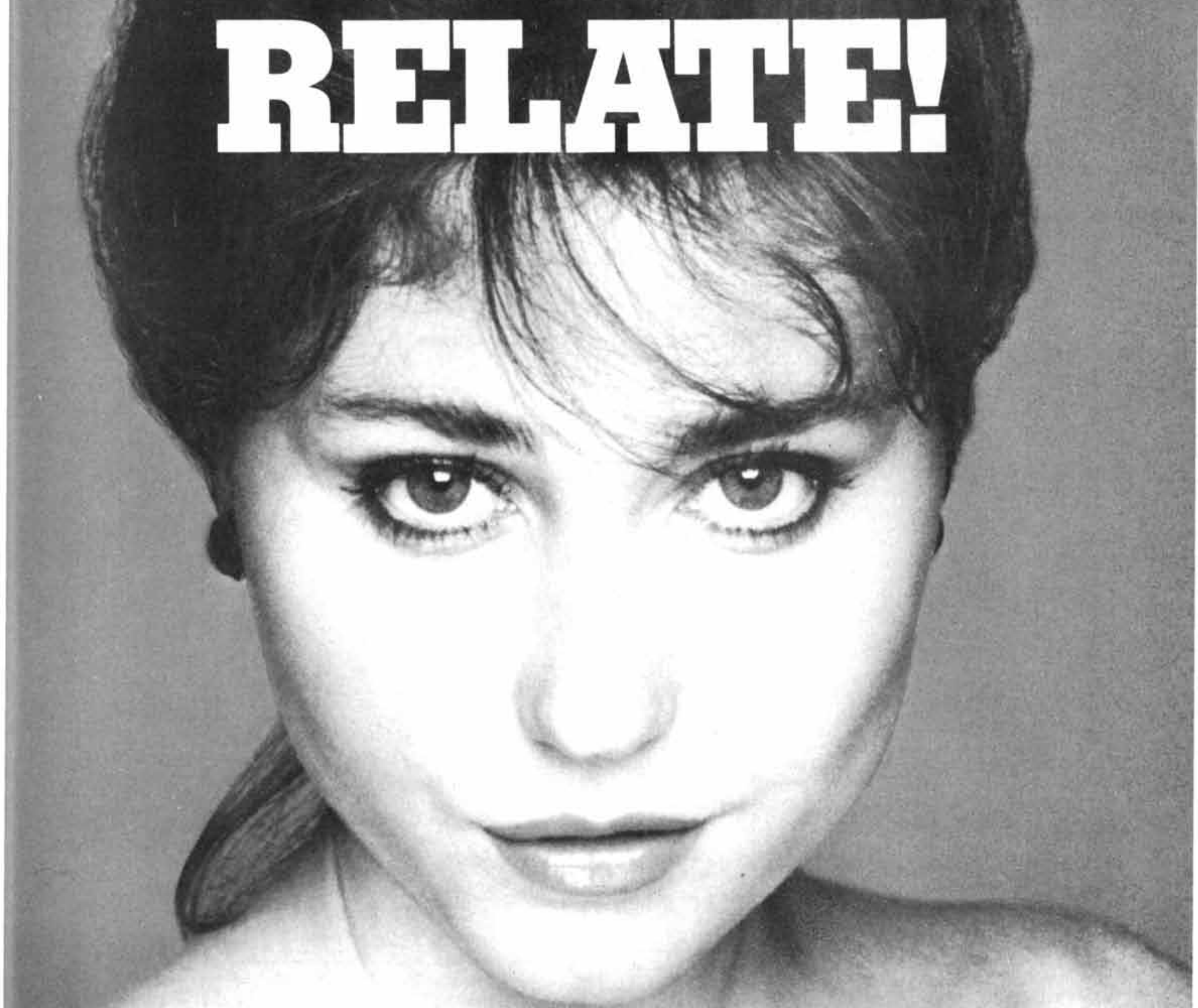
Paul C. Condit of San Anselmo, Calif., and Richard W. Hill of East Lansing, Mich., both told me about the practice of salting beer that has been poured into a glass. Some people salt beer because they like the taste, others because the salt generates more bubbles. The adsorbed pockets of air on each grain of salt serve as nucleating sites.

A "yard" container for beer or ale is a distinctively shaped structure that is about a yard long, with a round bottom and a long, narrow neck. The drinker must know how to drink out of it or he will be drenched by a sudden gush of fluid. I said the trick is to tap the neck lightly so that bubbles will flow up the tilted container to replace liquid that is gently flowing down to the drinker. Martin Reid of Surrey in England describes a better technique, which works even if the fluid is uncarbonated (as some ales are in England). He rotates the neck of the yard glass as he drinks. In this way fluid travels along the sides of the neck as well as the bottom, and the drinker achieves more control over the flow.

H. D. Westphal of Puchheim in West Germany wrote to me about a way to remove the excess head of foam that can build up in a poured beer. Weissbier (or weizenbier), which is brewed from wheat, has yeast added to it so that the final stage of brewing takes place in the bottle. The beer contains much more carbon dioxide than most other beers do, and so a big head develops when weissbier is poured. The traditional glass for the beer is tall, curved along the sides and flat on the inside of the bottom. It is rinsed in cold water, a small amount of which is left on the bottom. Then the bottle is put in the glass with the neck close to the bottom. As beer emerges and foam is generated the bottle is pulled upward at the same rate and the foam is sucked into it. The bottle is then set aside so that the beer that materializes as the bubbles burst can be drunk later.

Two letters reminded me of stories I have heard for years but have not been able to confirm. James L. Ealy, Jr., of Pottstown, Pa., said that when an unopened bottle or can of beer has been inadvertently shaken, the carbon dioxide that has been released in it can be

RELATE!



The eyes have it. Your eyes and the lens—the camera's eye—are the key to better pictures. That's why, when we developed the Soligor 80-200mm OneTouch Zoom+Macro, we pulled out all the stops to combine top quality with true economy. Its design, its features and its value make it one of the finest lenses we've ever offered, and Soligor offers more different zoom lenses than anyone else in the world.

Use the lower end of its range for clear, sharp, intimate portraits. Then, without your eye ever leaving the finder, this lens lets you simultaneously zoom and focus on distant action, or continue right into macro to learn what being close really means. Lightweight, perfectly balanced, beautifully precision engineered for Canon, Minolta, Nikon, Olympus, Pentax and other fine cameras, the Soligor 80-200mm is a lens used by professionals and by discriminating photographers everywhere.

You've seen the ordinary; now relate to the world through our eyes. At your nearest Soligor dealer, or write for full information.

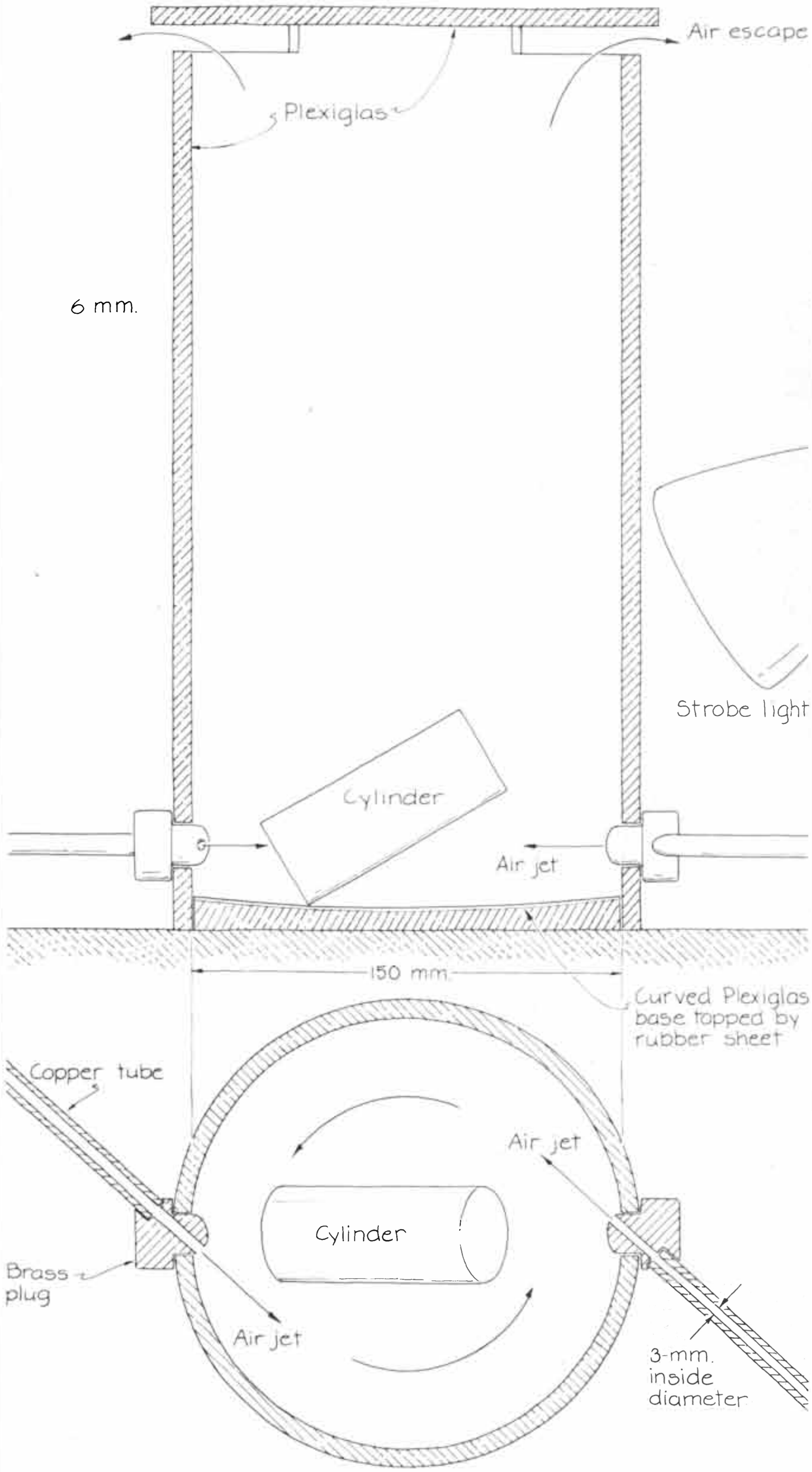
Soligor C/D 80-200mm OneTouch Zoom+Macro Lens
Aperture Range: f4.5—f22 *Construction:* 12 elements in 9 groups *Macro Ratio:* 1:4 *Filter Size:* 55mm *Length:* 4.8″
Weight: 16.9oz.

SOLIGOR®
MAKES THE DIFFERENCE

AIC **AIC Photo, Inc. Carle Place, N.Y. 11514**
In Canada: Kingsway Film Equipment., Ltd., Ontario

redissolved by tapping lightly on the container with a knife or some other object. Then one can open the container without releasing a gush of foam and liquid. I cannot explain how vibrations of the container and its contents could drive the gas back into solution.

Alvaro de Paiva Abreu of Rio de Janeiro reminded me that a carbonated drink in a container that has been opened can be kept from going flat for a while by putting the handle of a metal spoon in the neck. If this maneuver does in fact delay the loss of gas, how does it do so? I would welcome explanations of both of these tricks.



The apparatus designed by Whitehead for demonstrating wobble

NEW FROM McGRAW-HILL

For anyone thinking of a desk-top computer who doesn't know where to begin or what to actually expect when it gets going.

McGraw-Hill's At-Home Seminar in selecting the right microcomputer and software for your business or profession... and getting every benefit it has to offer quickly and trouble-free. On cassettes, with text, workbooks, and computer simulator.

Now, you can gain a clear understanding of whether your business will actually benefit from your having a microcomputer, and how much.

You can learn what a microcomputer will really do, how it does it, and what's required to make it happen within your organization.

You'll know the kind of computer that's best for you—in terms of your present needs and probable future requirements. Without paying too much for excess capacity, unnecessary functions, or the wrong peripherals; or, committing probably the greater mistake of paying too little for equipment inadequate to your needs. (There are more choices available than anyone can investigate on their own.)

You'll be able to evaluate the big name brands and the unfamiliar ones. Objectively and without having to depend on the "expertise" of a salesman.

Removes every mystery, helps you over every hurdle

In other words, with McGraw-Hill's at-home seminar, quickly and inexpensively, through cassettes by experts and vivid text, you'll be able to draw on the experiences of hundreds of others, become comfortable with the technical complexities (only a few are essential from the user's point of view), make knowledgeable, intelligent decisions about entering the computer age.

You'll know what programming is all about. Not to do it yourself but to evaluate those programs (software) that are available in terms of your own needs. As you'll see, the secret is not getting the right programs, difficult enough in itself, but programs that will work together so all your functions are compatible, particularly as your system develops and expands.

We take you through every step. Clearly and concisely. Without technical jargon. Of course, since this is one seminar that you control, repeating whatever's necessary is easy, while you can quickly skip through anything you already know or that's irrelevant to your purposes.

We show you exactly what to expect and what has to be done to perform with speed and efficiency the functions you have in mind already—inventory, general ledger, word processing, etc.

Then you'll see how your computer can be used to perform tasks no mechanical system can duplicate or human mind encompass. Controlling, coordinating, analyzing, forecasting... accessing information you never knew existed.

The Third Stage

Actually, this is the third stage of the computer revolution. From mainframes, which only the giant corporations could afford, to minicomputers costing $50 to $100,000 that brought computers to tens of thousands of companies, to the new generation of desk-top, micro/mini computers, ranging in price from $3-$4,000 to about $10 or $12,000 (with substantial tax benefits cutting the cost even more) that are feasible even for one to two-people firms.

What is truly extraordinary about this is that these new, little, inexpensive computers can do virtually everything those old giants could do—and in many ways more, since they're more flexible and can use preexisting programs for almost every function.

You don't have to delay or miss out on this for lack of understanding or fear you'll be getting in over your head. Conversely, plunging ahead without adequate preparation—depending on a salesman or a company representative to help you over the rough spots—could create a nightmare, disrupt your operation for months and cost you more money than a computer can save in years of operation.

If this seminar does no more than help you through the transition from mechanical to electronic operations, it will be one of the best investments of time and money you've ever made.

Free! Complete Information Without Risk or Obligation

Now, we'd like to send you complete information so you can make up your own mind about the program. You'll see that it covers every aspect, on nine cassettes, with accompanying text and a computer simulator that makes the flow of information through a computer system really easy to understand.

Of course, you can listen to the cassettes at home, in the office or while you're driving, turning to the text to brush up on fine points or as a ready reference. Equally important, you'll see the logical way the program develops, one section building a firm foundation for the next; so, grasping a function, for example, is easy since you already have command of the language.

It's quick, convenient, costs considerably less than any ordinary seminar and brings the computer to you from a business and professional point of view that's uniquely McGraw-Hill. For complete information call now. There's no risk or obligation.

Call Toll Free 1-800-323-1717
Ask for Operator 50 (In Illinois, phone 1-800-942-8881)
Or Mail Coupon

To: **McGraw-Hill Continuing Education Center**
Computer Science Division
3939 Wisconsin Avenue, NW
Washington, DC 20016

No Risk, No Obligation Reply Form
At-Home Seminar on Microcomputers with Cassettes

YES, I'm interested in finding out more about your microcomputer learning program. Please send me complete and detailed information. I understand this does not obligate me in any way and that no salesman will call.

Name (Please Print) Title

Organization

Street Address

City/State/Zip

3560-102

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

ONE BOOK/FIVE WAYS: THE PUBLISHING PROCEDURES OF FIVE UNIVERSITY PRESSES. Foreword by Joyce Kachergis and introduction by Chandler Grennis. William Kaufmann, Inc., 1978.

THE CONCEPT OF A META-FONT. Donald E. Knuth in *Visible Language,* Vol. 16, No. 1, pages 3–27; Winter, 1982.

THE VERIFICATION OF A COMPREHENSIVE NUCLEAR TEST BAN

TOWARD A COMPREHENSIVE NUCLEAR TEST BAN TREATY. Hearing before the Subcommittee on Arms Control, International Law and Organization of the Committee on Foreign Relations, U.S. Senate, May 15, 1972. U.S. Government Printing Office, 1972.

THE GREAT TEST-BAN DEBATE. Herbert F. York in *Scientific American,* Vol. 227, No. 5, pages 15–23; November, 1972.

NUCLEAR EXPLOSIONS AND EARTHQUAKES: THE PARTED VEIL. Bruce A. Bolt. W. H. Freeman and Company, 1976.

STUDY OF SEISMOLOGICAL EVASION, PART 1: GENERAL DISCUSSION OF VARIOUS EVASION SCHEMES. PART 2: EVALUATION OF EVASION POSSIBILITIES USING NORMAL MICROSEISMIC NOISE. PART 3: EVALUATION OF EROSION POSSIBILITIES USING CODAS OF LARGE EARTHQUAKES. Jack F. Evernden in *Bulletin of the Seismological Society of America,* Vol. 66, No. 1, pages 245–280 (Part 1), pages 281–324 (Part 2), February, 1976; No. 2, pages 549–592 (Part 3), April, 1976.

THE COMPREHENSIVE TEST BAN. Stockholm International Peace Research Institute in *World Armaments and Disarmament: SIPRI Yearbook 1978.* Taylor and Francis Ltd., 1978.

TRIPARTITE REPORT TO THE COMMITTEE ON DISARMAMENT (UNITED NATIONS). In *World Armaments and Disarmament: SIPRI Yearbook 1981.* Taylor and Francis Ltd., 1981.

CALCIUM IN SYNAPTIC TRANSMISSION

A STUDY OF SYNAPTIC TRANSMISSION IN THE ABSENCE OF NERVE IMPULSES. B. Katz and R. Miledi in *Journal of Physiology,* Vol. 192, No. 2, pages 407–436; September, 1967.

MEMBRANE ULTRASTRUCTURE OF THE GIANT SYNAPSE OF THE SQUID *LOLIGO PEALII.* D. W. Pumplin and T. S. Reese in *Neuroscience,* Vol. 3, No. 8, pages 685–696; August, 1978.

PRESYNAPTIC CALCIUM CURRENTS IN SQUID GIANT SYNAPSE. R. Llinás, I. Z. Steinberg and K. Walton in *Biophysical Journal,* Vol. 33, pages 289–321; March, 1981.

RELATIONSHIP BETWEEN PRESYNAPTIC CALCIUM CURRENT AND POSTSYNAPTIC POTENTIAL IN SQUID GIANT SYNAPSE. R. Llinás, I. Z. Steinberg and K. Walton in *Biophysical Journal,* Vol. 33, pages 323–351; March, 1981.

TRANSMISSION BY PRESYNAPTIC SPIKE-LIKE DEPOLARIZATION IN THE SQUID GIANT SYNAPSE. R. Llinás, M. Sugimori and S. M. Simon in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 79, No. 7, pages 2415–2419; April, 1982.

CHARGE-COUPLED DEVICES IN ASTRONOMY

CHARGE-COUPLED DEVICES AND SYSTEMS. Edited by M. J. Howes and D. V. Morgan. John Wiley & Sons, Inc., 1979.

SOLID STATE IMAGERS FOR ASTRONOMY. Edited by John C. Geary and David W. Latham. SPIE–The International Society for Optical Engineering, Bellingham, Wash., 1981.

ARTIFICIAL INTELLIGENCE

ARTIFICIAL INTELLIGENCE AND NATURAL MAN. Margaret Boden. Basic Books, 1977.

ARTIFICIAL INTELLIGENCE: AN MIT PERSPECTIVE. Patrick Henry Winston and Richard Henry Brown. The MIT Press, 1979.

MACHINES WHO THINK. Pamela McCorduck. W. H. Freeman and Company, 1979.

THE MIND'S I. Douglas R. Hofstadter and Daniel C. Dennett. Basic Books, 1981.

PERSPECTIVES IN COGNITIVE SCIENCE. Edited by Donald A. Norman. Ablex Publishing, 1981.

MAGNETIC FLUIDS

FERROHYDRODYNAMICS. Joseph L. Neuringer and Ronald E. Rosensweig in *The Physics of Fluids,* Vol. 7, No. 12, pages 1927–1937; December, 1964.

FLUID DYNAMICS AND SCIENCE OF MAGNETIC LIQUIDS. Ronald E. Rosensweig in *Advances in Electronics and Electron Physics,* Vol. 48, pages 103–199; 1979.

PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON MAGNETIC FLUIDS. In *IEEE Transactions on Magnetics,* Vol. Mag-16, No. 2, pages 171–415; March, 1980.

TIDAL BORES

A THEORY OF THE PROPAGATION OF BORES IN CHANNELS AND RIVERS. M. R. Abbott in *Proceedings of the Cambridge Philosophical Society,* Vol. 52, Part 2, pages 344–362; April, 1956.

THE SEVERN BORE. Fred Rowbotham. David & Charles, Newton Abbot, 1964.

BORES, BREAKERS, WAVES AND WAKES: AN INTRODUCTION TO THE STUDY OF WAVES ON WATER. R. A. R. Tricker. American Elsevier Publishing Co., 1965.

THE SEAS IN MOTION. F. G. Walton Smith. Thomas Y. Crowell Co., 1973.

HOW HONEYBEES FIND A HOME

THE SOCIAL BEHAVIOR OF BEES: A COMPARATIVE STUDY. Charles D. Michener. Harvard University Press, 1974.

THE NEST OF THE HONEY BEE. T. D. Seeley and R. A. Morse in *Insectes Sociaux,* Vol. 23, pages 455–512; 1976.

HISTORICAL ECLIPSES

GROWTH RHYTHMS AND THE HISTORY OF THE EARTH'S ROTATION. INTERDISCIPLINARY WINTER CONFERENCE ON BIOLOGICAL CLOCKS AND CHANGES IN THE EARTH'S ROTATION: GEOPHYSICAL AND ASTRONOMICAL CONSEQUENCES, NEWCASTLE-UPON-TYNE, 1974. Edited by G. D. Rosenberg and S. K. Runcorn. John Wiley & Sons, Inc., 1975.

APPLICATIONS OF EARLY ASTRONOMICAL RECORDS. F. Richard Stephenson and David H. Clark. Oxford University Press, 1978.

TIDAL FRICTION AND THE EARTH'S ROTATION. Edited by P. Brosche and J. Sündermann. Springer-Verlag, 1978.

THE EARTH'S VARIABLE ROTATION: GEOPHYSICAL CAUSES AND CONSEQUENCES. Kurt Lambeck. Cambridge University Press, 1980.

PROCEEDINGS OF A CONFERENCE ON THE ANCIENT SUN. Edited by R. O. Pepin, J. A. Eddy and R. B. Merrill. Pergamon Press, 1980.

SOLAR RADIUS VARIATIONS OVER THE PAST 265 YEARS. Ronald L. Gilliland in *The Astrophysical Journal,* Vol. 248, No. 3, Part 1, pages 1144–1155; September, 1981.

THE AMATEUR SCIENTIST

COIN SPINNING ON A TABLE. M. G. Olsson in *American Journal of Physics,* Vol. 40, pages 1543–1545; October, 1972.


dimension

"Bell put reduced operating costs at our fingertips."

If you've got a business, the Bell System has a way for you to reduce operating costs. By installing a Dimension® PBX from Bell, you gain control over many features of your communications system.

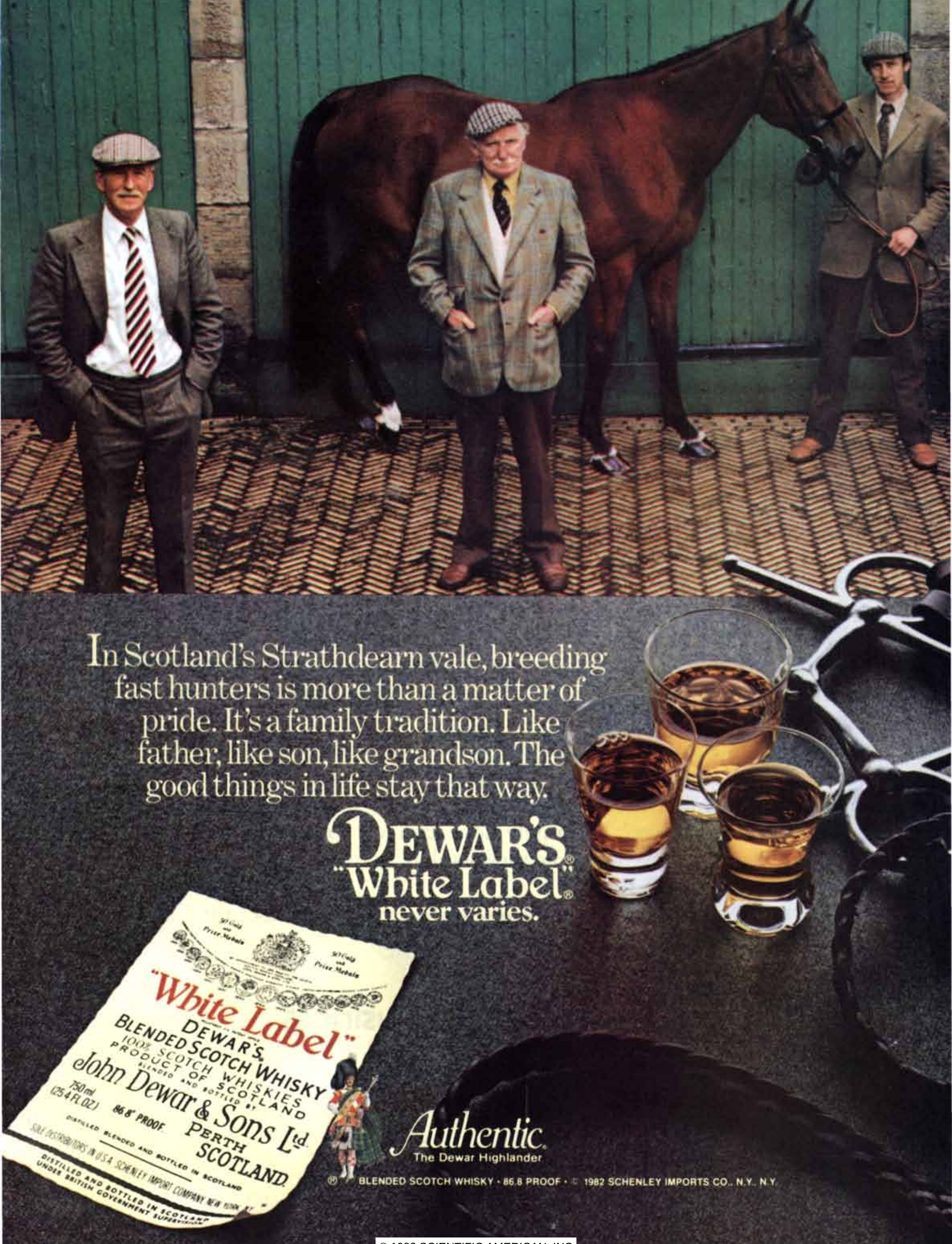
You will be able to assign your own extension numbers, which means that an employee can change offices and keep the same number. Any member of your staff can transfer a call without involving your operator. Any call can be turned into a three-way conference call, with callers speaking from three different telephone numbers. Any employee can instruct the system to direct his or her calls to another telephone. Special rings distinguish outside calls from inside calls.

You can assign special features to specific phones. You can limit which phones have access to long distance calling. You can print out your own record of calls for individual phones.

But you won't be entirely alone. Bell will continue to check your system to identify problems before they develop. In many cases, repairs will be made remotely.

Turn your telephone system into a productivity resource. Call your Bell System Account Executive.

The knowledge business



In Scotland's Strathdearn vale, breeding fast hunters is more than a matter of pride. It's a family tradition. Like father, like son, like grandson. The good things in life stay that way.
DEWAR'S® "White Label"® never varies.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
PRODUCT OF SCOTLAND
BLENDED AND BOTTLED BY
John Dewar & Sons Ltd
PERTH
SCOTLAND
750 ml
(25.4 FL.OZ.)
86.8° PROOF
DISTILLED BLENDED AND BOTTLED IN SCOTLAND
SOLE DISTRIBUTORS IN U.S.A. SCHENLEY IMPORT COMPANY NEW YORK
DISTILLED AND BOTTLED IN SCOTLAND UNDER BRITISH GOVERNMENT SUPERVISION
Authentic.®
The Dewar Highlander
® BLENDED SCOTCH WHISKY · 86.8 PROOF · © 1982 SCHENLEY IMPORTS CO., N.Y., N.Y.