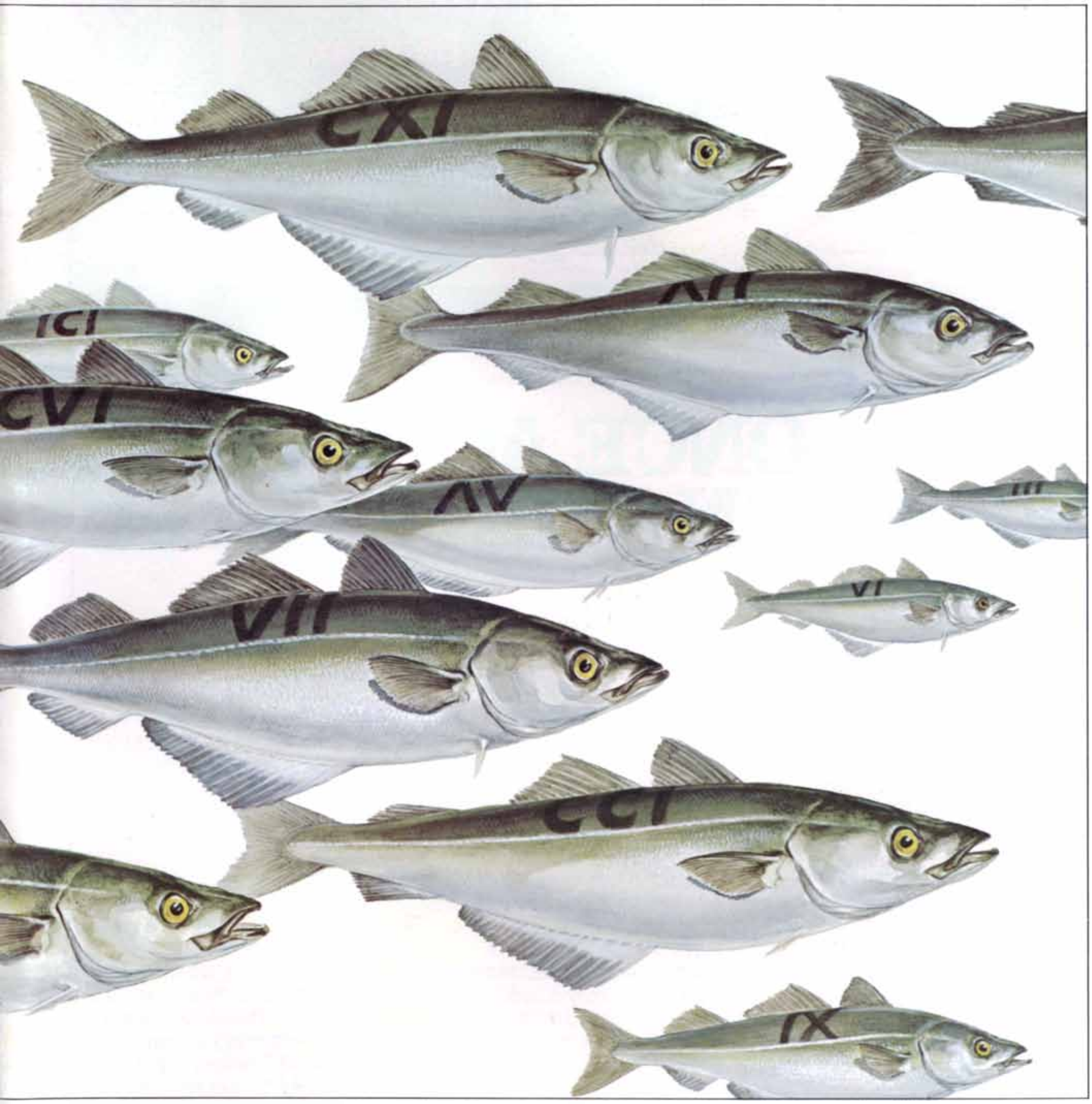
# SCIENTIFIC AMERICAN




SCHOOLING OF FISHES

$2.00

June 1982


# ONCE AGAIN, SONY BETWEEN THE AMATEUR

Introducing the Sony Betamax SL-2500. The most advanced home videocassette recorder ever conceived.

Its professional features are so sophisticated they are utterly simple to use. As a totality, those features combine to form a machine whose capabilities narrow the gap between the professional and the amateur video enthusiast.

Notice the height. It's equal to just three videocassettes. Or, 3⅛" high. Notice the width. It's equal to the width of standard audio components as Sony moves to integrate video and audio into one entertainment system.

Notice the advanced front loading and location of the feather-touch controls. All human engineered for ease of operation.

**NOW RECORDING, EDITING, AND REVIEWING ARE AS EASY AS THUMBING THROUGH A BOOK.**

To Sony, the ultimate goal of sophisticated technology is simplicity, not complexity.

Nowhere is this more evident than in the SL-2500.

We saw the process of recording, editing, and reviewing a video home movie as similar to that of thumbing through, reading, and referring to a book.

If you think of a Table of Contents, you'll understand the new Electronic Tab Marker index. You'll be able to program up to nine points, or chapters, on your recording. And be able to go directly to each of them at the push of a button.

If you think of skimming through a book, you'll understand BetaScan. You're able to quickly review material to find the spot you're looking for.

If you think of flipping back and forth between two pages to locate the place you need to find, you've got the idea behind Swing Search. In five different special effects speeds you can review with crystal clarity the frames where the action has occurred. So you get the exact frame where the edit should occur.

Sometimes, when you read, you wish to study a page for a long time. The same holds true when reviewing a scene. For that, you need Freeze-Frame, and the SL-2500's double Azimuth Recording Heads provide a crystal-clear one.

With the optional BetaStack tape changer, you can record or play back up to four separate pro-


# NARROWS THE GAP AND THE PROFESSIONAL.

grams, each on its own cassette. Each show becomes a single volume, as you match the cassette length to the program length. Making it easy to refer to and enjoy. Should you wish to compile anthologies, the BetaStack gives you a maximum of 20 hours of consecutive recording or playback.

### LINEAR TIME COUNTER. YOU KNOW EXACTLY WHERE YOU ARE IN HOURS, MINUTES, AND SECONDS.


Just as pages in a book are marked by a number, the Linear Time Counter electronically treats each second you record or play back as if it were a page. So instead of relying on a mechanical counter, you easily know where each scene and each program falls precisely on the tape.

And, as you can judge how much of a book is left to be read by the thickness of the remaining pages, the LED Tape-Remaining Indicator tells you how much tape is left. And even blinks as you approach the end of the tape.

### MICRO TIME PHASE CIRCUITRY FOR CLEAR, NOISELESS EDITS.

An editor is only as good as his tools. And in the past, those tools could not deliver edits that didn't call attention to themselves.

The Micro Time Phase Circuitry of the SL-2500 has made that problem ancient history.

When you push the pause or stop button, the tape rewinds 15 frames. When the pause button is pushed again, the tape advances 12 frames, which gives the motor a chance to get into the Record mode. The edit is clean, without a "glitch."

Beta

Linear Time Counter, Electronic Tab Marker and Remote Commander are trademarks of Sony Corporation.

### WIRELESS REMOTE COMMANDER. PROFESSIONAL CONTROL FROM YOUR EASY CHAIR.

Channel selection, BetaScan, Play, Record, and all the Swing Search functions can all be accomplished with our wireless remote. Sophisticated technology again bringing simplicity and ease to operation.


This is the Sony SL-2500.

Suddenly, there are no amateurs.

**SONY®**
**THE ONE AND ONLY**

See your Sony dealer for a full demonstration. Or write: Sony Video Information Center, CN 02950, Trenton, NJ 08650. Models shown: SL-2500 and RMT-312.


You belong where the Beefeater® is.
Imported From England
BEEFEATER
LONDON 1873 1911 1924
London Distilled Dry Gin
94 PROOF
Distilled and bottled by James Burrough Limited in London England
Sole U.S. Importers Kobrand Corporation N.Y.
100% grain neutral spirits.
JAMES BURROUGH LTD. DISTILLERS
BEEFEATER® GIN. The Crown Jewel of England.™

Established 1845 **SCIENTIFIC AMERICAN** June 1982 Volume 246 Number 6

ARTICLES



DEPARTMENTS





SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1982 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. 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

# Robert and Mathilde, fresh from riding in a Peugeot 505 S for 3,000 miles on one of the most horrendous roads in the world.

It is late afternoon. The rain has turned to a cold drizzle. In the gloom, a mud-spattered car pulls off the road and stops in front of a small, nondescript building.

Men in blue lab coats hurry over to the car and open the doors. The driver's seat is empty. There are no human beings in the car.

### Orders from under the road

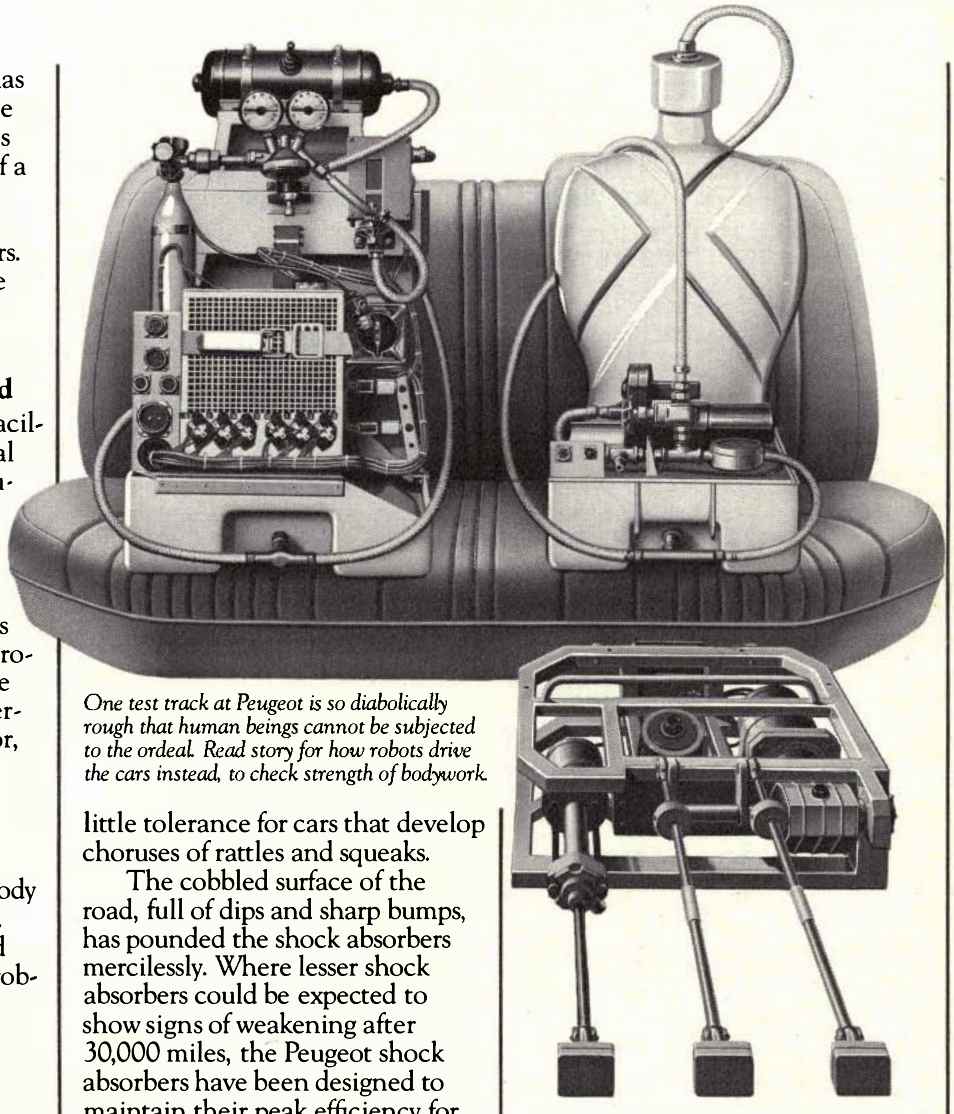
The place is a Peugeot testing facility in Europe. The road, so brutal that Peugeot will not subject human test drivers to the ordeal. In their place... robots.

For 3,000 torturous miles, the robots have been under the command of a computer. Orders were sent to the car via an electromagnetic cable buried under the road. The "driver" robot has operated the steering, the accelerator, the clutch, the brakes.

### A merciless pounding

The car is a Peugeot 505 S. Its body is welded rigidly in 4,032 places. The robot testers have punished the body along this vile road, probing for weaknesses. Peugeot has little tolerance for cars that develop choruses of rattles and squeaks.

*Sections of Belgian pavé that pound cars mercilessly are replicated on robot test track.*



*One test track at Peugeot is so diabolically rough that human beings cannot be subjected to the ordeal. Read story for how robots drive the cars instead, to check strength of bodywork.*

The cobbled surface of the road, full of dips and sharp bumps, has pounded the shock absorbers mercilessly. Where lesser shock absorbers could be expected to show signs of weakening after 30,000 miles, the Peugeot shock absorbers have been designed to maintain their peak efficiency for 60,000 miles of normal use.

### Inside a tough car... soft touches

Lest all this talk of Peugeot's toughness and durability mislead the reader into thinking of the car as some crude, range-riding vehicle, now is a good time to step inside the car.

The interior of the Peugeot 505 S teems with soft touches. The seats welcome you like an old friend. They have no jouncy springs; in their place is polymerized foam, molded to the shape of the human anatomy under the guidance of physiologists.

A pleasant surprise; this is a *true* 5-seater, with room in the back for adult-length legs and adult-height heads.

The heating and ventilating system can be directed to play cool air across your face while simultaneously bathing the rest of you in warm air—a great help in keeping a driver fresh and alert on long trips in the winter.

The radio is a marvel of electronic wizardry. It will automatically search the dial for stations, pause a few seconds at each one so that you can sample the program, then move on to the next station unless you order it to stop.


*A Peugeot winning the car-breaking Bandama Rally in Africa, 1978. The almost eerily smooth suspension helps make Peugeot the best-selling car on the continent, which has some of the roughest roads in the world.*

### 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 505S, this is the "S" (for Special) equipment that comes with the car:

- Air conditioning, factory-installed
- Electric sunroof

### 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).

- 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 transmission ($370) in place of the 5-speed overdrive manual gearbox, and by specifying metallic paint ($295).

### Crotchety about quality

Only rashness bordering on insanity would lead any carmaker to promise you a car free of even a single fault. Human beings make cars, and human beings make mistakes. Peugeot, however, is extremely crotchety about quality.

The company maintains a special team of 65 test drivers. Their job is to put *every single car* through an unusual test on a special track after it rolls off the assembly line.

The driver will personally 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.

Only then is the car permitted to leave Europe for the U.S.

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**

67-000SA-0600

*The Peugeot 505S, from the second-largest carmaker in Europe. A car that does not nibble you to death with extra-cost options.*

PEUGEOT

## THE COVER

The painting on the cover shows part of a school of pollock (*Pollachius virens*), a saltwater fish that is much like a cod. The pollock is known in Europe as the saithe. It reaches about 35 centimeters in length. Schools of pollock were employed in experiments that have provided the most extensive information yet obtained on the structure of the fish school (see "The Structure and Function of Fish Schools," by Brian L. Partridge, page 114). The number on each pollock is a temporary brand made by applying a piece of cold metal to the side of the fish. The numbers helped to identify individual fish in schools of from 20 to 30 pollock swimming around the circumference of a large circular tank. The school was followed around the tank by a gantry that held an observer and a videotape camera. The observer gave a tape-recorded commentary on the position of the fish. When the videotape was analyzed with the aid of the commentary, it yielded abundant information about the minute changes of speed and direction that are made continuously by each member of the school.

## THE ILLUSTRATIONS

Cover painting by Enid Kotschnig

| Page | Source | Page | Source |
|---|---|---|---|
| 46–50 | Andrew Tomko | 108 | Andrew Tomko |
| 51 | W. R. Grace & Co. | 109 | Patricia J. Wynne |
| 52 | Palomar Observatory (*top*), Richard S. Simon (*bottom*) | 110 | Brigitte and Gilles Delluc (*top*); Arlette Leroi-Gourhan, Musée de l'Homme (*bottom*) |
| 54–57 | Allen Beechel | 111 | André Glory |
| 58 | Alan T. Moffet | 112 | Patricia J. Wynne |
| 59 | Palomar Observatory (*top*), Kevin J. Hussey (*middle*), Kitt Peak National Observatory (*bottom*) | 115 | Robert Holland, University of Miami |
| 60 | Kevin J. Hussey (*left*), Allen Beechel (*right*) | 116 | Brian L. Partridge |
| 61 | National Geographic Society–Palomar Observatory Sky Survey (*top*), Kevin J. Hussey (*bottom*) | 117 | Tom Prentiss |
| | | 118 | National Marine Fisheries Services (*top*); James A. Bohnsack, University of Miami (*bottom*) |
| 63–67 | Alan D. Iselin | 119–122 | Tom Prentiss |
| 68 | John R. Dedman, B. R. Brinkley and Anthony R. Means | 125 | Caroline F. Keating, Colgate University, and Gregory Keating, Upstate Medical Center at Syracuse, State University of New York |
| 69 | Jeffrey F. Harper and Alton L. Steiner (*top left and bottom*), John G. Wood (*top right*) | 126–132 | Patricia J. Wynne |
| 70 | Alan D. Iselin | 134 | Richard Jung, University of Freiburg |
| 83–100 | Jerome Kuhl | | |
| 105 | J. Oster, Musée de l'Homme | 137 | The Granger Collection |
| 106 | Denis Vialou, Muséum National d'Histoire Naturelle | 138 | The Birmingham Assay Office |
| | | 139–142 | Courtesy of the National Portrait Gallery, London |
| 107 | Andrew Tomko (*top*); Pamela B. Vandiver, Massachusetts Institute of Technology (*bottom*) | 144 | City of Birmingham, Public Libraries Department |
| | | 146–153 | Michael Goodman |

# LETTERS

Sirs:

In this age of hysteria with regard to any topic including the word "radiation" it was gratifying to read Arthur C. Upton's factual account ["The Biological Effects of Low-Level Ionizing Radiation," by Arthur C. Upton; SCIENTIFIC AMERICAN, February]. The statement "The evidence so far indicates that compared with other hazards [the low-level-radiation hazard] is slight" in the subtitle almost left me speechless.

I should nonetheless like to point out a serious shortcoming in the bar chart captioned "Perception of risk." The chart is presented with a logarithmic scale that overemphasizes the lower end. The 100 deaths attributed to low-level radiation from nuclear-power sources per year in the U.S. looks like more than a third the risk due to smoking, if a reader casually assumes the chart abscissa is linear, as most bar charts are. A better comparison would be to compare a stack of 100 sheets of paper from your magazine, representing the nuclear-power risk, with a stack of 300,000 sheets representing the risk of smoking, drinking and driving; the 100 sheets (almost the thickness of the magazine) are considerably less than a quarter of an inch high, whereas the top three risks represent a stack some 630 inches, or 52.5 feet, high. If one were to include all the risks greater than nuclear power, the stack would be some 61 feet high.

HELMUT WOLF

Raymond F. Giffels Distinguished
Professor of Mechanical Engineering
University of Arkansas
Fayetteville

Sirs:

Arthur C. Upton's article on the biological effects of low-level ionizing radiation graphically demonstrates that nuclear power is commonly perceived to be far riskier than it objectively is. This raises a perplexing problem for policymakers and regulatory agencies: should they take into account the perceived risk of nuclear energy in placing restrictions on the industry or should they instead impose only those controls warranted by the risks actually presented?

A compelling case can be made that perceived risks must be excluded entirely from the regulatory calculus in the establishment of health and safety standards and requirements. If regulators take into account unfounded fears of nuclear power, then far more stringent restrictions will be imposed than are warranted by the actual hazards. This will necessarily increase the cost of nuclear power relative to other energy technologies and will unduly predispose the marketplace toward riskier and less desirable alternatives.

Until recently a strong legal argument could also be mounted against taking perceived risks into account. The Administrative Procedure Act, a Federal statute, provides that arbitrary and capricious agency action is invalid. The perception of nuclear power as being ultrarisky, as Dr. Upton indicates, is arbitrary. A regulatory requirement founded on arbitrary fears is itself capricious. Agencies may not impose capricious requirements based on arbitrary fears of nuclear power under the Administrative Procedure Act.

Perceived risk still has a place in policymaking. Any interest group that considers itself unduly afflicted by a particular technology, including nuclear power, may always petition Congress to take irrational fears into account. Congress may impose stringent requirements not otherwise warranted by the facts in the event that it decides to oblige. However, when Congress has left the question of standards and requirements to unelected officials in regulatory agencies (as it has done in the case of nuclear power), those agencies must act on the basis of facts and rational analysis, not on the basis of speculation, conjecture and misconceived fear.

The United States Court of Appeals for the District of Columbia has recently thrown a substantial cloud over this analysis. In *People Against Nuclear Energy (PANE)* v. *NRC* the court ordered the Nuclear Regulatory Commission (NRC) to assess the "psychological health of neighboring residents" in the licensing of a nuclear reactor. The court evidently based this decision on the National Environmental Policy Act (NEPA). That statute hitherto had been viewed as calling for consideration of environmental costs and benefits of Government action. It had not been read as requiring that factors such as mental anguish be included expressly in the decision making of a regulatory agency. The PANE case, should it be something other than an aberration, may prove to be a significant impediment to the development of nuclear power and potentially to technology in general. . . .

CHARLES H. MONTANGE

Washington, D.C.

Sirs:

Does "Retrospective Instrumentation for Analysis of Physical Traces of UFOs" (*Journal of UFO Studies,* Vol. 2, pages 37–46; 1980) read like a *National Enquirer* headline? Does "ESP with Unbalanced Decks: A Study of the Process in an Exceptional Subject" (*The Journal of Parapsychology,* Vol. 37, pages 278–297; 1973) read like a *National Enquirer* headline? These are the titles of articles in two research journals, one dealing with UFO's, the other with ESP. To the majority of *Scientific American* readers the existence of these journals may come as a surprise. Their premises and individual articles may be wrong; certainly they are no less controversial than Darwin's *Origin of Species* and no less obscure than was Mendel's research before it was rediscovered. Research is published to be criticized. If Douglas R. Hofstadter wants to write about ESP or UFO's, he should have the respect for his readers to at least demonstrate some knowledge of research in these fields. To deal with these controversial topics by speculating about *National Enquirer* headlines or by wringing hands over credulous undergraduates reduces *Scientific American* to the level of a *National Enquirer* of science.

D. C. DONDERI

Department of Psychology
McGill University
Montreal


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

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

**Offprint correspondence** and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 75 cents.

**Subscription correspondence** should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name ______________________

New Address ______________________

______________________

______________________

Old Address ______________________

______________________

______________________

# 50 AND 100 YEARS AGO

## SCIENTIFIC AMERICAN

JUNE, 1932: "In 1915 William D. Harkins, professor of physical chemistry at the University of Chicago, proposed that atoms are built up in steps. In 1921 he began to photograph in the cloud chamber the rare collisions of helium nuclei with the nuclei of nitrogen and other atoms. The photographs showed conclusively that in some collisions the helium nucleus and the nitrogen nucleus unite and form a heavier nucleus, that of fluorine. This nucleus is very unstable and almost simultaneously shoots off a hydrogen nucleus and is thus converted into oxygen. In all the work done since then in Professor Harkins' laboratory only four atoms of oxygen have been synthesized, and to show this it has been necessary to take more than 100,000 photographs. The importance of atom building lies not only in the transformation of the material but also in the enormous energy changes involved. The best example is given in the well-known calculation of Wilson and Professor Harkins based on the special relativity theory of Einstein. It was shown by Wilson and Harkins that the union of one pound of hydrogen to give helium or any heavier element would yield as much energy as the combustion of 10,000 tons of coal. It is not improbable, however, that this synthesis is extremely slow, and while it is probably an important (possibly the most important) source of heat for the sun and other stars, where high temperatures and enormous amounts of material are involved, there is no indication that it will ever be a source of energy for industrial purposes under the conditions existent on the earth."

"The Rolls-Royce 'R' engine fitted to the Supermarine S.6.B seaplane is by far the most efficient aircraft engine in the world. In the S.6.B, Flight Lieutenant G. H. Stainforth has set a new world's record over a three-kilometer course, with a speed of 415.2 miles per hour. At present there seems to be no limit in sight for the maximum speed that can be reached. Already there are airplanes flying in which arrangements have been made to retract the undercarriage wheels into the body, and seaplane floats may be expected to follow. The saving in resistance, if floats and struts are omitted, would amount to as much as half the total for the entire aircraft. Given the necessary incentive and the funds, there appears to be no reason speeds should not continuously increase until they approach the velocity of sound at 740 miles per hour. Then, owing to the formation of pressure waves and the rapidly increasing resistance, it is difficult to predict what will happen."

"Before a recent meeting of the Optical Society of America, Dr. C. E. Kenneth Mees of the Eastman Kodak Company described how it is possible by means of a new dye to make plates sensitive to invisible heat rays. Xenocyanine is the name of the dye. Hitherto dicyanine and more recently neocyanine have been used to sensitize plates that have been particularly useful in photographing from an airplane through thin mists. Infra-red and red rays are not scattered by mist as are ultra-violet rays. With the aid of the new plates Drs. W. F. Meggers, C. C. Kiess and C. J. Humphreys of the National Bureau of Standards have found unsuspected lines in the spectra of 36 chemical elements. It is probable that in the stars elements will be discovered whose existence is not even suspected. There is even the hope that very dark stars that have not yet been detected by telescopic cameras may be photographed."

"Our nation faces an emergency unparalleled in all its history. While in the throes of the world's greatest depression, it piled up a Federal deficit of $903,000,000 in 1931; another of more than $2,000,000,000 is promised for the fiscal year 1932. The Federal budget must be balanced! But how? The first and most important source of revenue that meets the eye is the liquor industry, which has cost the Government more than $300,000,000 for 'enforcement' and from which, if it were legalized, the Government could obtain up to $1,000,000,000 annually."

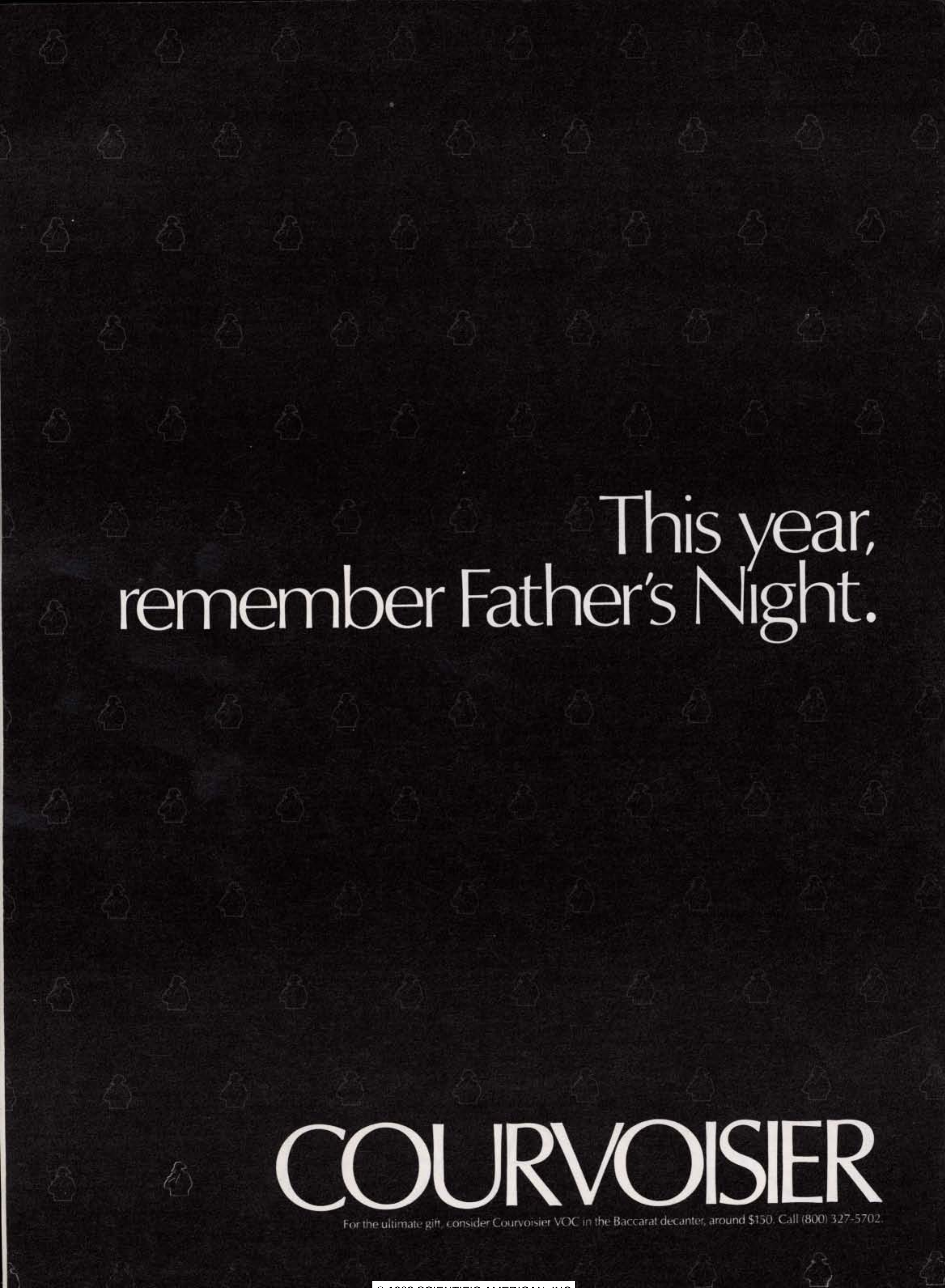
## SCIENTIFIC AMERICAN

JUNE, 1882: "Few objects in the heavens have been treated with such unmerited neglect as the Great Nebula in Andromeda. Notwithstanding its enormous extent, as followed by the Harvard 15-inch achromatic, of $2\frac{1}{4}$ degrees in length and one degree in breadth, and its conspicuous brightness, which makes it readily perceptible to the naked eye, it has resisted all inquiry. The telescope has failed. The spectroscope gives some kind of reply, but it is an indecisive one. It precludes at once the idea of a simple gaseous condition such as that of the Ring nebula, or the Dumb-bell or the wonder in Orion. What is that at which we gaze, overspreading field after field of the telescope with soft yet vivid light? If it is gaseous, the gas is unknown or in some hitherto unknown condition. If it is stellar, how are its components so concealed that they cannot be isolated? If stars are there, they must be numbered by the hundreds of thousands, yet they must be of a lesser magnitude than we associate with the idea of a star."

"The production of artificial indigo is making progress. Artificial indigo is not itself manufactured, the coloring matter being deposited by the printer himself in the fiber of the cloth from propiolic acid. For this purpose a mixture of propiolic acid and xanthate of soda is printed on the cloth. The color makes its appearance on exposure to the air. The first necessity in the technical preparation of a propiolic acid is a cheap cinnamic acid. W. H. Perkin was the first to make this acid by a simple reaction, viz., by heating bitter almond oil with acetic anhydride and acetate of soda. The Baden company at Ludwigshafen has simplified this reaction and prepares cinnamic acid directly by fusing benzyl chloride with acetate of soda. Propiolic acid itself is precipitated from the solution of the soda salt by means of an acid and is sent to the market as a paste. The only serious imperfection in the process is the loss of 40 per cent of the cinnamic acid in the form of para-nitrocinnamic ether. This imperfection will, it is to be hoped, be overcome by the employment of the compound for some useful purpose."

"The death on April 19 of Mr. Darwin has called forth in England and every other civilized country unanimous and unqualified testimonies to his great merits as the leading scientific mind of his time. Both in Westminster Abbey and in St. Paul's Cathedral the preachers spoke of Mr. Darwin's life and labors, praising him for his 'pure and earnest love of truth' and his patient care and industry in its pursuit. Canon Liddon, in his sermon at St. Paul's, observed 'that when Professor Darwin's books on the origin of species and on the descent of man first appeared, they were largely regarded by religious men as containing a theory necessarily hostile to religion. A closer study has greatly modified any such impression.' It is seen that whether the creative activity of God is manifested through catastrophes, as the phrase goes, or in progressive evolution, it is still His creative activity, and the really great questions beyond remain untouched. The evolutionary process, supposing it to exist, must have had a beginning. Who began it? It must have had material to work with; who furnished it? It is itself a law or system of laws; who enacted them?"


IT'S A NEW KIND
OF CADILLAC...
...for a new kind of Cadillac owner. Compare Cimarron to the likes of the BMW 320i, Audi 5000 and Saab 900s. Spec for spec, feature for feature. Only Cimarron has front-wheel drive, air conditioning and genuine leather seating areas with lumbar support, all standard. Plus Cadillac's exclusively tuned touring suspension. Take a Cimarron test drive today, at the Cadillac dealership nearest you.
GM
CIMARRON
BEST OF ALL...IT'S BY CADILLAC.


This year,
remember Father's Night.
COURVOISIER
For the ultimate gift, consider Courvoisier VOC in the Baccarat decanter, around $150. Call (800) 327-5702.


COURVOISIER
COGNAC
Admiré pour sa perfection, sa qualité a conquis le monde entier
VSOP
Fine Champagne
COURVOISIER
The Brandy of Napoleon
COGNAC
Produced and Bottled by Courvoisier, Jarnac, France
The Cognac of Napoleon

# THE AUTHORS

RICHARD P. SHELDON ("Phosphate Rock") is a geologist with a particular interest in sedimentary mineral deposits. Graduated with a B.S. from Yale University in 1950, he went on to get his Ph.D. in geology from Stanford University in 1956. In 1957 and 1958 he worked as a geologist for the Lion Oil Company. In 1958 he joined the U.S. Geological Survey, for which he worked until his retirement this year. In 1964 and 1965 he was visiting professor at Yale, and in 1968 and 1969 he was visiting professor at the Colorado School of Mines.

ANTHONY C. S. READHEAD ("Radio Astronomy by Very-Long-Baseline Interferometry") is professor of astronomy at the California Institute of Technology and director of the Owens Valley Radio Observatory. A native of South Africa, he received his undergraduate education at the University of the Witwatersrand, obtaining a B.Sc. in 1968. His Ph.D., awarded in 1972, is from the University of Cambridge. From 1972 to 1977 he continued his work under a grant from the Royal Society of London, first at Cambridge and then at Cal Tech. In 1977 Readhead joined the faculty at Cal Tech, and last year he took up his current jobs.

WAI YIU CHEUNG ("Calmodulin") is member in biochemistry at St. Jude Children's Research Hospital in Memphis, Tenn. Born in Kwangtung, China, he was graduated from the National Tsung Hsing University in Taiwan with a B.S. in 1956. He continued with his education in the U.S., obtaining his M.S. at the University of Vermont in 1960 and his Ph.D. from Cornell University in 1964. From 1964 to 1967 he was research fellow at the University of Pennsylvania. In 1967 he moved to St. Jude. Since 1976 he has also been professor of biochemistry at the University of Tennessee Center for the Health Sciences in Memphis. Cheung's main research interest is the molecules that serve as messengers between cells in the regulation of biological processes, including calcium ions and cyclic AMP.

MARVIN L. COHEN, VOLKER HEINE and JAMES C. PHILLIPS ("The Quantum Mechanics of Materials") are physicists who share an interest in the structure of solids and its relation to the physical properties of the solids. Cohen is professor of physics at the University of California at Berkeley. Born in Canada, he moved to the U.S. for his university education. His B.S. (1957) is from the University of California at Berkeley; his M.S. (1958) and Ph.D. (1964) are from the University of Chicago. In 1963 and 1964 he worked at Bell Laboratories. In the latter year he joined the faculty at Berkeley; he was appointed professor in 1969. Heine is professor of physics at the University of Cambridge. Born in Germany, he emigrated to New Zealand in 1939. He began his education in physics there, obtaining his bachelor's degree (1950) and his master's degree (1953) at the University of Otago. Thereafter he moved to England; he received his doctorate from the University of Cambridge in 1956. After a year of postdoctoral work at the University of California at Berkeley he returned to Cambridge, where he has remained. Phillips is a member of the technical staff of Bell Laboratories. His Ph.D., granted in 1956, is from the University of Chicago. In the same year he joined the staff of Bell Laboratories. In 1958 and 1959 he was a research fellow in physics at the University of California at Berkeley; the following two years he worked in the same capacity at Cambridge. From 1960 to 1968 he was on the faculty of the University of Chicago, returning in 1968 to Bell Laboratories.

ARLETTE LEROI-GOURHAN ("The Archaeology of Lascaux Cave") is director of the laboratory of palynology (the study of spores and pollens) at the Musée de l'Homme in Paris. She has a degree in Oriental archaeology from the École du Louvre. After obtaining it she spent two years in Japan with her husband, the anthropologist André Leroi-Gourhan. On returning to France she worked for three years at the Muséum d'Histoire Naturelle studying methods of applying botanical knowledge to archaeology. Her work at the Musée de l'Homme concerns the climate and biological environment of Europe over the past 100,000 years and its relation to human settlement.

BRIAN L. PARTRIDGE ("The Structure and Function of Fish Schools") is assistant professor of biology at the University of Miami. He received his undergraduate education at Stanford University and at the University of British Columbia; his B.Sc. was given by the latter institution in 1974. He went on to obtain a D.Phil. from the University of Oxford in 1978. He then moved to the Scripps Institution of Oceanography as a research fellow, studying the neurological mechanisms that enable weakly electric fish to locate objects. He joined the faculty of the University of Miami in 1979. He has recently been accepted at medical school; after finishing his medical studies he plans to do work on human neurophysiology.

ROBERT H. WURTZ, MICHAEL E. GOLDBERG and DAVID LEE ROBINSON ("Brain Mechanisms of Visual Attention") are colleagues at the Laboratory of Sensorimotor Research at the National Eye Institute who have a common interest in visual perception. Wurtz is chief of the laboratory. He received his bachelor's degree in 1958 at Oberlin College and his doctorate in 1962 from the University of Michigan. From 1962 to 1965 he was research associate at Washington University. Wurtz moved to the National Institutes of Health (NIH) in 1965 and was appointed to his present position in 1978. Goldberg is chief of the section on basic neuro-ophthalmologic mechanisms at the laboratory. His A.B. (1963) is from Harvard College; his M.D. (1968) is from the Harvard Medical School. In 1969 he joined the staff of the NIH as associate neurologist at the National Institute of Mental Health. He left to serve a residency and to work at the Armed Forces Radiobiological Research Institute, returning to the NIH in 1978. Since 1976 he has also been clinical associate professor of neurology at Georgetown University. Robinson is research physiologist at the laboratory. He obtained his B.S. (1965) at Springfield College, his M.S. (1968) at Wake Forest University and his Ph.D. (1972) from the University of Rochester. From 1971 to 1974 he was research fellow in neurophysiology at the National Institute of Mental Health, leaving in 1974 to become research physiologist at the Armed Forces Radiobiological Research Institute. He returned to the NIH in 1978.

LORD RITCHIE-CALDER ("The Lunar Society of Birmingham"), who died at the age of 75 in January, had a long career as a writer on science and technology and their social consequences. Born in Scotland, he quit school at 15 and went to work as a police court reporter, first in Dundee and then in Glasgow. In 1926 he went on to London, where as a reporter successively for *Daily News, Daily Chronicle* and *The Daily Herald* he came to Fleet Street fame for his reporting of the social turmoil of the Depression and for his pioneering coverage of science. During World War II he ran the political-warfare bureau in the Foreign Office. After the war, as science editor of *The News Chronicle* and *The New Statesman,* he brought the new technical agencies of the United Nations and the economic development of the former colonial regions to public attention. From 1961 to 1967 he occupied a chair in international relations established for him at the University of Edinburgh. Created a life peer in 1966, he served as chairman of the U.K. Metrication Board and was a founding trustee of the Open University. Calder wrote more than 30 books, most of them on topics related to science and the future of man.


She can't flaunt a fur on the Côte d'Azur.
Born out of fire and ice more than a hundred million years ago. Every diamond is unique. But a diamond of one carat or more is even more precious. A gift so rare, it can never be measured. Until you see the look in her eyes.
THE DIAMOND SOLITAIRE. A RARE GIFT.
The one and a quarter carat diamond pendant shown is enlarged for detail.
A diamond is forever. De Beers

# METAMAGICAL THEMAS

## *About Nomic: a heroic game that explores the reflexivity of the law*

by Douglas R. Hofstadter

In his excellent book *A Profile of Mathematical Logic* the philosopher Howard DeLong tells the following classic story of ancient Greece. "Protagoras had contracted to teach Euathlus rhetoric so that he could become a lawyer. Euathlus initially paid only half of the large fee, and they agreed that the second installment should be paid after Euathlus had won his first case in court. Euathlus, however, delayed going into practice for quite some time. Protagoras, worrying about his reputation as well as wanting the money, decided to sue. In court Protagoras argued:

"'Euathlus maintains he should not pay me but this is absurd. For suppose he wins this case. Since this is his maiden appearance in court he then ought to pay me because he won his first case. On the other hand, suppose he loses his case. Then he ought to pay me by the judgment of the court. Since he must either win or lose the case he must pay me.'

"Euathlus had been a good student and was able to answer Protagoras' argument with a similar one of his own:

"'Protagoras maintains that I should pay him but it is this which is absurd. For suppose he wins this case. Since I will not have won my first case I do not need to pay him according to our agreement. On the other hand, suppose he loses the case. Then I do not have to pay him by judgment of the court. Since he must either win or lose I do not have to pay him.'"

Then DeLong adds, "It is clear that to straighten out such puzzles one has to inquire into general procedures of argument." Actually, to many people it is not so clear. Quite the contrary. To many people paradoxes such as this one appear to be mere blemishes on the face of the law, which can be removed by simple cosmetic surgery. Similarly, many people who take theology seriously think that paradoxical questions about omnipotence such as "Can God make a stone so heavy that he cannot lift it?" are just childish riddles, not serious theological dilemmas, and can be resolved in a definitive and easy way. Throughout history simplistic or patchwork remedies have been proposed for all kinds of dilemmas created by loops of this sort. Bertrand Russell's theory of types is a famous example in logic. The loops will not go away that easily, however, as Russell found out. They are deep and pervasive, and attempts to unravel them lead down unexpected pathways.

In fact, reflexivity dilemmas of the Protagoras-v.-Euathlus type and problems of conflicting omnipotence crop up with astonishing regularity in the down-to-earth discipline of the law. Yet until recently their central importance in defining the nature of the law has been little noticed. In the past few years only a handful of specialized papers on the subject have appeared in law journals and philosophy journals.

It was with surprise and delight, therefore, that I learned that an entire book on the role of reflexivity in the law was in preparation. I first received word of it—*The Paradox of Self-Amendment: A Study of Logic, Law, Omnipotence, and Change*—in a letter from its author, Peter Suber, who identified himself as a philosophy Ph.D. attending law school in Chicago. This was more than a year ago; since then Suber has finished law school and has taken a job teaching philosophy at Earlham College in Richmond, Ind. He hopes *The Paradox of Self-Amendment* will be out soon.

In correspondence with Suber I have found out that he has an even more ambitious book in the works, tentatively titled *The Anatomy of Reflexivity.* It is a study of reflexivity in its broadest sense, encompassing, as he says, "the self-reference of signs, the self-applicability of principles, the self-justification and self-refutation of propositions and inferences, the self-creation and self-destruction of legal and logical entities, the self-limitation and self-augmentation of powers, circular reasoning, circular causation, vicious and benign circles, feedback systems, mutual dependency, mutual causation, and reciprocity."

In his original letter Suber not only gave a number of interesting examples of self-reference in law but also presented a game he calls Nomic (from the Greek *nomos,* "law"), which is presented in an appendix in *The Paradox of Self-Amendment.* I found reading the rules of Nomic to be a mind-opening experience. Much of this article will be devoted to Nomic, but before we tackle the game itself I should like to set the stage by mentioning some other examples of reflexivity in the political arena.

My friend Scott Buresh, himself a lawyer, described the following perplexing hypothetical dilemma, which he first heard posed in a class on constitutional law. What if Congress passes a law saying that henceforth all determinations by the Supreme Court shall be made by a 6–3 majority (rather than a simple 5–4 majority, as is currently the case)? Imagine that this law is challenged in a court case that eventually makes its way up to the Supreme Court itself, and that the Supreme Court rules that the law is unconstitutional—and needless to say the ruling is by a 5–4 majority. What happens? This is a classic paradox of the separation of powers, and it was nearly played out, in a minor variation, during the Watergate era, when President Nixon threatened he would obey a Supreme Court ruling to turn over his tapes only if it were "definitive," which presumably meant something like a unanimous decision.

It is interesting to note that conservatives are now trying to limit the jurisdiction of the Supreme Court over issues such as abortion and prayer in the schools. Constitutional scholars expect that a showdown might ensue if Congress passes such a statute and the Supreme Court is asked to review its constitutionality.

Conflicts that enmesh the Supreme Court with itself can arise in less flashy ways. Suppose the Supreme Court proposes to build an annex in an area environmentalists want to protect. The environmentalists take their case to court, and it gets blown up into a large affair that eventually reaches the level of the Supreme Court. What happens? Clearly the reason this kind of thing cannot be prevented is that any court is itself a part of society, with buildings, employees, contracts and so on. And since the law deals with things of this kind, no court at any level can guarantee that it will never get ensnared in legal problems.

If self-ensnaredness is a rare event for the Supreme Court, it is not so rare for other arms of government. An interesting case came up recently in San Francisco. There had been a large number of complaints about the way the police department was handling cases, and so an introverted internal-affairs committee was set up to look into such matters as police brutality. Then, inevitably, complaints arose that the internal-affairs committee was whitewashing its findings, and so Mayor Dianne Feinstein


POWER
STYLING
DATSUN 200-SX
Now that gas mileage is the hot button, few cars boast about their styling. The new Datsun 200-SX, Hardtop SL is a brilliant exception. Look at the integrated, wrap-around bumpers. The new louvers on the hood. The classic blacked-out grille. Inside: real comfort and the ultimate in sophistication. This car actually "talks" to you. A voice calls your attention to six vital functions like… "Right door is open." On the road a more powerful 2.2 liter fuel-injected engine propels you along without sacrificing economy: EPA estimated 26 MPG, 38 estimated highway. Use MPG for comparison only. Your mileage may differ depending on speed, trip length and weather. Highway mileage will probably be less. The thrilling 200-SX: proof positive that no one marries performance and styling like Datsun.
DATSUN
WE ARE DRIVEN
Product of NISSAN
NISSAN
© 1981 NMC-U.S.A.

# "The data from this experiment could affect national defense. That's why we use Scotch® Brand Disk Cartridges."


**Ken Bish, Engineer,**
**Systems Research Laboratories,**
**Inc., Dayton, Ohio**

Every Scotch Disk Cartridge is tested and certified error-free before it leaves the factory. Because, for the defense of your data, nothing less than perfection is acceptable.

The disks in Scotch Disk Cartridges are defended by 3M's exclusive CRASHGUARD® protective disk coating. It greatly minimizes the possibility of a head crash, and minimizes the damage, should one occur.

Scotch front-loading Disk Cartridges also feature our exclusive Living Hinge air door, which reduces the possibility of damage to the disk from contact with the air door stop during shipping and handling.

You can get Scotch Disk Cartridges in front or top-loading models. To find out where you can find Scotch Disk 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 Hears You...**

**3M**

set up a doubly introverted committee, again internal to the police department, to investigate the performance of the internal-affairs committee. The last I heard was that the report of this committee was unfavorable. What finally resulted I do not know.

Parliamentary procedure too can lead to the most tangled of situations. For example, there are several editions of *Robert's Rules of Order,* and a body must choose which set of rules will govern its deliberations. The latest edition of *Robert's Rules* states that if no specific edition is chosen as the governing one, then the most recent issue holds. A problem arises if one has not adopted the latest edition, since one cannot then rely on its authority to tell one to rely on it.

In some ways parliamentary procedure, which deals with how to handle simultaneous and competing claims for attention, bears a remarkable resemblance to the way a large computer system must manage its own internal affairs. Within such a system there is always a program called an operating system with a part called the scheduling algorithm that weighs priorities and decides which activity will proceed next. In a "multiprocessing" system this means determining which activity gets the next "time slice" (lasting for anywhere from a millisecond to a few seconds, or possibly even for an unlimited time, depending on the activity's priority and numerous other factors). But there are also "interrupts" that come and interfere with—oops, just a moment, my telephone's ringing. Be right back. There. Sorry we were disturbed. Someone wanted to sell me a telephone-answering system. Now what would—ah, ah, just a sec—ahchoo!—sorry—what would I do with one of those things? Now where was I? Oh, yes—interrupts. Well, in a way they are like telephone calls that take the store clerk away from you, annoying you in the extreme, since you have come to the store in person whereas the telephone caller has been lazy and yet is given higher priority.

A good scheduling algorithm strives to be equitable, but all kinds of conflicts can arise, in which interrupts interrupt interrupts and are then themselves interrupted. Moreover, the scheduler has to be able to run its own internal decision-making programs with high priority, yet not so high a priority that nothing else ever runs. Sometimes the internal and external priorities can become so tangled that the entire system begins to "thrash." This is the term used to describe a situation where the operating system is spending most of its time bogged down in "introverted" computation, deciding what it should spend its time doing. Needless to say, during periods of thrashing very little "real" computation gets done. It sounds quite like the cognitive state a person can get into when too many factors are weighing down all at once and the slightest thought on any topic seems to trigger a rash of paradoxical dilemmas from which there is no escape.

Operating systems and courts of law cannot, for better or worse, go to sleep. Their tangles are very real, and some means of dealing with them has to be invented. It was considerations such as this that led Suber to invent his tangled game of Nomic.

He writes that he was struck by the often heard cynicism that "government is just a game." Now, one essential activity of government is lawmaking, and so if government is a game, it is a game in which changing the laws (or rules) is a move. Moreover, some rules are needed to structure the process of changing the rules. Yet no legal system seems to have any rules that are absolutely immune to legal change. Suber's main aim, he wrote, was "to make a playable game that models this particular situation. But whereas governments are at any given moment pushed in various directions in their rule changing by historical realities and the ideology of their people and existing rules, I wanted the game to start with as 'clean' an initial set of rules as possible." Nomic is such a game, and its rules (or rather, its Initial Set of rules) will be presented below. Most of the following description is in essence by Suber himself. I have simply interspersed some of my own observations.

In legal systems, statutes are the paradigmatic rules. Statutes are made by a rule-governed process that is itself partly statutory; hence the power to make and change statutes can reach some of the rules governing the process itself. Most of the rules that govern the making of statutes, however, are constitutional and are therefore beyond the reach of the power they govern. For instance, Congress may change its parliamentary rules and its committee structure, and it may bind its future action by its past action, but through mere statutes it cannot alter the fact that a two-thirds "supermajority" is needed to override an executive veto, nor can it abolish or circumvent one of its houses, start a tax bill in the Senate or even delegate too much of its power to experts.

Although statutes cannot affect constitutional rules, the latter can affect the former. This is an important difference of logical priority. When there is a conflict between rules of different types, the constitutional rules always prevail. This logical level distinction is matched by a political level distinction, namely that the logically prior (constitutional) rules are more difficult to amend than the logically posterior (statutory) rules.

It is no coincidence that logically prior laws are harder to amend. One purpose of making some rules more difficult to change than others is to prevent a brief wave of fanaticism from undoing decades or even centuries of progress. This could be called self-paternalism: a deliberate retreat from democratic principles, although one chosen for the sake of preserving democracy. It is our chosen insurance against our anticipated weak moments. That purpose will not, however, be met unless the two-tier (or multitier) system also creates a logical hierarchy in which the less mutable rules take logical priority over the more mutable rules; otherwise the more mutable rules could by themselves undo the deeper and more abstract principles on which the entire system is based. If supermajorities and the concurrence of many bodies are necessary to protect the foundations of the system from hasty change, that protective purpose is frustrated if those foundations are reachable by rules requiring merely a simple majority of one legislature.

Although all the rules in the American system are mutable, it is convenient to refer to the less mutable constitutional rules as immutable and to the more mutable rules below them in the hierarchy as mutable. The same is true in Nomic, where at least initially no rule is literally immutable. If Nomic's self-paternalism is to be effective, then, its "immutable" rules, in addition to resisting easy amendment, must have logical priority.

Many designs could satisfy this requirement. Nomic has adopted a simple two-tiered system, modeled to some extent on the U.S. Constitution. In principle a system could have any number of degrees of difficulty in the amendment of rules. For instance, Class *A* rules, the hardest to amend, could require unanimity of a central body and the unanimous concurrence of all regional bodies. Class *B* rules could require 90 percent supermajorities, Class *C* rules 80 percent supermajorities and so on. The number of such categories could be indefinitely large. Indeed, if appropriate qualifications are made for the informality of custom and etiquette, a strong argument could be made that normal social life is just such a system of indefinite tiers. Near the top of the "difficult" end of the series of rules are actual laws, rising through case precedents, regulations and statutes to constitutional rules. At the bottom of the scale are rules of personal behavior that individuals can amend unilaterally without incurring disapprobation or censure. Above these are rules for which amendment is increasingly costly, starting with costs on the order of furrowed brows and clucked tongues and passing through indignant blows and vengeful homicide.

In any case, for the sake of simplicity and to make it easier to learn and play, Nomic is a clean two-tier system rather than a nuanced or multitier system like the U.S. Government, with its intermediate and substatutory levels such as parliamentary rules, administrative regulations, joint resolutions, treaties, executive agreements, higher and lower court decisions, state practice, judicial rules of procedure and evidence, executive orders, canons of professional responsibility, evidentiary presumptions, standards of reasonableness, rules establishing priority among rules, canons of interpretation, contractual rules and so on. This is not to say that nuanced, intermediate levels may not arise in Nomic through game custom and tacit understandings. In fact, the nature of the game allows players to add new tiers by explicit amendment as they see fit, and it is easier to add them to a simple game than it is to subtract them from a complex one.

Nomic's two-tier system embodies the same self-paternalistic elements as the Federal Constitution. The "immutable" rules govern more basic processes than the "mutable" ones and thus shield them from hasty change. Since in the course of play the central core of the game may change (and the minor points *must* change), after a few rounds the game being played by the players may in a certain sense be different from the one they were playing when they started. Yet whatever results from compliance with the rules is by definition the game Nomic. The "feel" of the game may change drastically even as, at a deeper level, the game remains the same.

In a similar way human beings undergo constant development and self-modification and yet continue to be convinced that it makes sense to refer, with such words as "I," to an underlying stable entity. The more immediately perceptible patterns change, whereas deeper and more hidden patterns remain the same. From birth to maturity to death, however, the changes can be so radical that one may sometimes feel that in a single lifetime one is several different people. Similarly, in law many have acknowledged that a clause defining how a constitution may be amended, even a clause limited to piecemeal amendment, could through repeated application create a fundamentally new constitution.

The fact that Nomic has more than one tier prevents the logical foundation of the game—the central core—from changing radically in just a few moves. Such continuity is a virtue both of games and of governments, but players of Nomic have an advantage over citizens in that, whenever they are so motivated, they can adjust the degree of continuity and the rate of change rather quickly, using their wits, whereas in real life the mechanisms by which such change could be effected are barely known and partially beyond reach.

Standard games have the continuity of unchanging rules, or at least of rules that change only between games, not during them. Nomic's continuity is more like that of a legal system than that of a standard game: it is a rule-governed set of systems, directives and processes undergoing constant rule-governed change. If, however, one wants a specific entity to point to as being "Nomic itself," the Initial Set of rules given below will do. Yet Nomic is equally the product, at any given moment, of the dynamic rule-governed change of the Initial Set. The continuing identity of the game, like that of a nation or of an individual, is due to the fact (if fact it is) that all change is the product of existing rules properly applied and that no change is revolutionary. (One could even argue that revolutionary change is just more of the same: in a revolution rules that have been assumed to be totally immutable simply are rendered mutable by other rules that are more deeply immutable but that previously had been taken for granted and hence had been invisible, or tacit.)

In its Rule 212, Nomic includes provision for judgment (as in a court of law), not merely to imitate government in yet another aspect but for the same reasons that compel government itself to make provisions for judgment: rules will inevitably be made that are ambiguous, inconsistent or incomplete, or that require application to individual circumstances. "Play" must not be interrupted; therefore some agency must be empowered to make an authoritative and final determination so that play can continue.

Judgments in Nomic are not bound by rules of precedent, since that would require a daunting amount of record keeping for each game. But the doctrine of *stare decisis* (namely that precedents should be followed) may be imposed at the players' option, or it may arise without explicit amendment, as successive judges feel impelled to treat "similarly situated" persons "similarly." (Admittedly the meanings of these terms in specific cases may well require further levels of judgment. This fact is one of the most dangerous sources of potential infinite regress in real court cases.) Without *stare decisis* the players are constrained to draft their rules carefully, make thoughtful adjudications, overrule poor judgments and amend defective rules. This is one way Nomic teaches basic principles and exigencies of the law even as it vastly simplifies.

The Initial Set must be short and simple enough to encourage play, yet long and complex enough to cover contingencies likely to arise before the players get around to providing for them in a rule, and to prevent any single rule change from disturbing the continuity of the game.

One contingency deliberately left to the players to resolve is what to do about violations of the rules. The players must also decide whether old violations are protected by a statute of lim-

Windsor's taste beats V.O.!
Over 7,000 drinkers across the country have compared the taste of Imported Windsor Canadian to the taste of Seagrams V.O.
And they chose the mild, smooth taste of Windsor over V.O.—57% to 43%.
If you're a Windsor drinker, that will come as no surprise. But if you're not, try the taste of Windsor and find out for yourself why it's the Canadian that stands alone.
You can't beat the taste of Windsor. And you certainly can't beat the price.
WINDSOR
ONE CANADIAN STANDS ALONE
CANADIAN
IMPORTED
WINDSOR
Supreme
CANADIAN
Blended Canadian Whisky
BOTTLED IN CANADA

itations or whether they may still be punished or nullified. Whether the likelihood of compliance and the discretionary power of the judge suffice to deal with a crisis of confidence or to delay it until a rule can take over, and whether in other respects the Initial Set satisfactorily balances the competing interests of simplicity and complexity, can best be determined by playing the game.

Nomic affords a curious twist on one common and fundamental property of games: it allows the blurring of the distinction between constitutive rules and rules of skill, that is, between rules that define lawful play and those that define artful play. In other words, in Nomic there is a blurring between the permissible and the optimal.

Most games do not embrace nonplay and do not become paradoxical by seeming to embrace it. Interestingly, however, children often invent games that provide game penalties for declining to play, or that incorporate or extend game jurisdiction to all of "real life" and end only when the children tire of the game or forget they are playing. ("Daddy, Daddy, come play a new game we invented!" "No, sweetheart, I'm reading." "That's 10 points!") Nomic carries this principle to an extreme. A game of Nomic can embrace anything at the vote of the players. The line betweeen play and nonplay may shift at each turn, or it may apparently be eliminated. Players may be governed by the game when they think they are between games or when they think they have quit.

For most games there is an infallible decision procedure to determine the legality of a move. In Nomic, on the other hand, situations may easily arise where it is very hard to determine whether or not a move is legal. Moreover, paradoxes can arise in Nomic that paralyze judgment. Occasionally this will be due to the poor drafting of a rule, but it may also arise from a rule that is unambiguous but mischievous. The variety of such paradoxes is truly impossible to anticipate. Rule 213, nonetheless, is designed to cope with them as well as possible without cluttering the Initial Set with too many legalistic qualifications. Note that Rule 213 allows a wily player to create a paradox, get it passed (if the rule seems innocent enough to the other players) and thereby win.

So much for a general prologue to the game itself. We can now move on to a description of how a game of Nomic is played. To reiterate, Nomic is a game in which changing the rules is a move. Two can play, but three or more make for a better game. The gist of Nomic is to be found in Rule 202, which should be read first. Players will need paper and pencil, and (at least at the outset) one die. Instead of sheets of paper, players may find it easier to use a set of index cards. All new rules and amendments are to be written down. How the rules are positioned on paper or on the table can indicate which are then "immutable" and which "mutable." Amendments can be placed on top of or next to the rules they amend. Inoperative rules can simply be deleted. Alternatively, for more complex games players may prefer to transcribe into their own notebooks the text of each new rule or amendment and to keep a separate list by number of the rules still in effect. Ideally, perhaps, all rules should be entered in a computer, with a terminal for each player; amendments then could be incorporated instantly with a corresponding adjustment to the numerical order.

## INITIAL SET OF RULES

### *Immutable Rules*

101. All players must always abide by all the rules then in effect, in the form in which they are then in effect. The rules in the Initial Set are in effect whenever a game begins. The Initial Set consists of Rules 101–116 (immutable) and 201–213 (mutable).

102. Initially rules in the 100's are immutable and rules in the 200's are mutable. Rules subsequently enacted or transmuted (that is, changed from immutable to mutable or vice versa) may be immutable or mutable regardless of their numbers, and rules in the Initial Set may be transmuted regardless of their numbers.

103. A rule change is any of the following: (1) the enactment, repeal or amendment of a mutable rule; (2) the enactment, repeal or amendment of an amendment, or (3) the transmutation of an immutable rule into a mutable rule or vice versa. (Note: This definition implies that, at least initially, all new rules are mutable; immutable rules, as long as they are immutable, may not be amended or repealed; mutable rules, as long as they are mutable, may be amended or repealed; no rule is absolutely immune to change.)

104. All rule changes proposed in the proper way must be voted on. They will be adopted if and only if the required number of votes are received.

105. Every player is an eligible voter. Every eligible voter must participate in every vote on rule changes.

106. All proposed rule changes must be written down before they are voted on. If they are adopted, they must guide play in the form in which they were voted on.

107. No rule change may take effect earlier than the moment of the completion of the vote that adopted it, even if its wording explicitly states otherwise. No rule change may have retroactive application.

108. Each proposed rule change must be given a rank-order number (ordinal number) and must be referred to by that number. The numbers must begin with 301, and each rule change proposed in the proper way must receive the next successive integer, whether or not the proposal is adopted.

If a rule is repealed and reenacted, it receives the ordinal number of the proposal to reenact it. If a rule is amended or transmuted, it receives the ordinal number of the proposal to amend or transmute it. If an amendment is amended or repealed, the entire rule of which it is a part receives the ordinal number of the proposal to amend or repeal the amendment.

109. Rule changes that transmute immutable rules into mutable rules may be adopted if and only if the vote is unanimous among the eligible voters.

110. Mutable rules that are inconsistent in any way with some immutable rule (and that can be made consistent with it only by transmuting it into a mutable rule) are wholly void and without effect. They do not implicitly transmute immutable rules into mutable rules and at the same time amend them. Rule changes that transmute immutable rules into mutable rules will be effective if and only if they explicitly state their transmuting effect.

111. If a rule change as proposed is unclear, ambiguous, paradoxical or destructive of play, is held by a player to consist of two or more rule changes compounded or to be an amendment that makes no difference, or is otherwise held to be of questionable value, then the other players can suggest amendments or argue against the rule change before the vote. The proponent, however, decides the final form in which the proposal is to be voted on and chooses the time to end debate and vote.

112. The state of affairs that constitutes winning may not be altered from achieving *n* points to any other state of affairs. The magnitude of *n* and the means of earning points may, however, be altered, and rules that establish a winner when play cannot continue may be enacted and (when they are mutable) be amended or repealed.

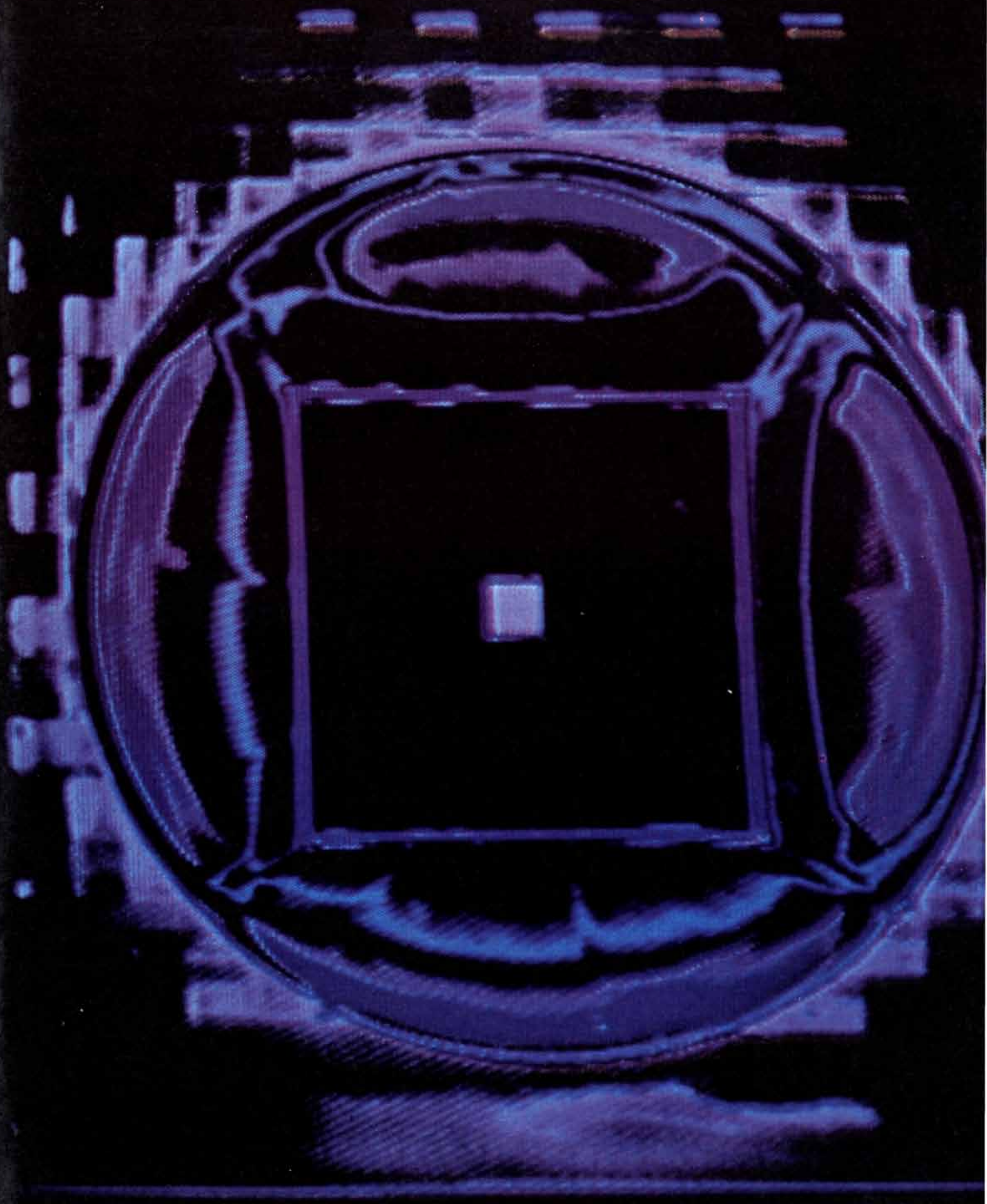
113. A player always has the option of forfeiting the game rather than continuing to play or incurring a game penalty. (No penalty worse than losing, in the judgment of the player incurring the penalty, may be imposed.)

114. There must always be at least one mutable rule. The adoption of rule changes must never become completely impermissible.

115. Rule changes that in any way affect rules needed to allow or apply rule changes are fully as permissible as other rule changes. Even rule changes that repeal part or all of their own authority are permissible. No rule change or type of move is or is to be impermissible solely because of the self-reference or self-application of a rule.

116. The adoption of rule changes

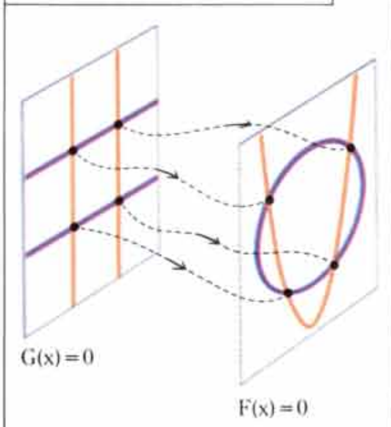
# The Continuation Method



# The Continuation Method

*The need to solve systems of polynomial equations arises in pursuits ranging from geometric optics to chemical kinetics. A practical method of solution, developed at the General Motors Research Laboratories, provides designers of mechanical parts with a new capability.*

Two equations in two unknowns




*The two pairs of parallel lines of G(x) = 0 evolve into the parabola and ellipse of F(x) = 0.*

*The three pairs of parallel planes of G(x) = 0 evolve into the paraboloid, ellipsoid and cylinder of F(x) = 0.*

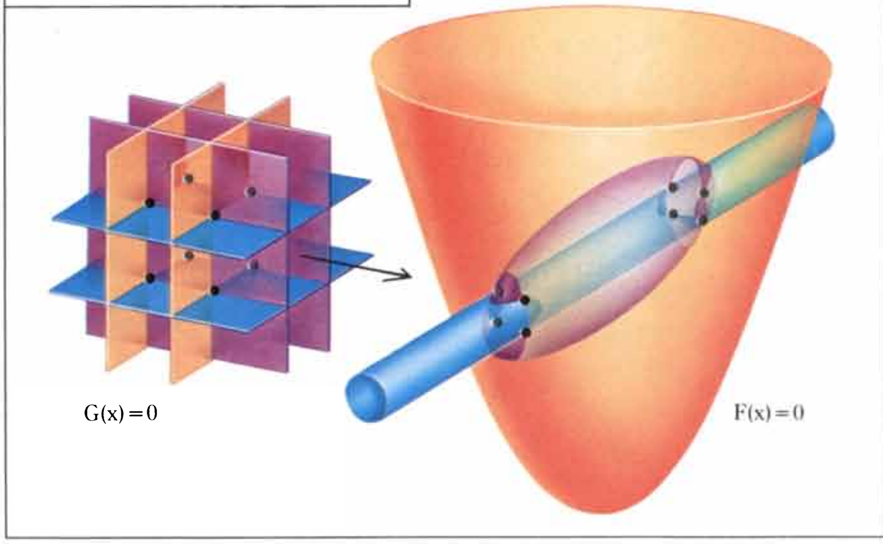
CLASSICALLY difficult non-linear equations—those made up of polynomial expressions—can now be solved with reliability and speed. Recent advances in the mathematics of continuation methods at the General Motors Research Laboratories have practical implications for a wide range of scientific and engineering problems. The immediate application at General Motors is in mechanical design. The new method finds all eight solutions to three quadric equations in a few tenths of a second—fast enough for computer-aided design on a moment-to-moment basis. Algorithms based on this method are critical to the functioning of GMSOLID, an interactive design system which models the geometric characteristics of automotive parts.

Systems of non-linear equations have been solved for many years by "hit or miss" *local* methods. The method developed at General Motors by Dr. Alexander Morgan is distinguished by being *global* and *exhaustive*. Local methods depend on an initial estimate of the solution. They proceed by iterative modifications of this estimate to converge to a solution. However, success is not guaranteed, because there are generally no practical guidelines for making an initial choice that will ensure convergence. Reliability is further compromised when multiple solutions are sought.

Global methods, by contrast, do not require an initial estimate of the solution. The continuation method, as developed by Dr. Morgan, is not only global, but also exhaustive in that, assuming exact arithmetic, it guarantees convergence to all solutions. The convergence proof rests on principles from the area of mathematics called differential topology.

Here is the way continuation works. Suppose we want to solve a system F(x) = 0. We begin by generating a simpler system G(x) = 0 which we can both solve and continuously evolve into F(x) = 0. It is important that we select a G properly, so the process will converge. Dr. Morgan has devised a method for selecting G which gives rapid convergence and reliable computational behavior. He first applied a theorem established by Garcia and

Three equations in three unknowns



Zangwill to select G. However, the resulting algorithm could not achieve the speed and computational reliability necessary for several applications. Next, he utilized some ideas from algebraic geometry—"homogenous coordinates" and "complex projective space"—to prove a new theorem for selecting G. The result of Dr. Morgan's efforts is a practical numerical method based on solid mathematical principles with innate reliability.

Reliability is the critical element for mathematical methods embedded in large computer programs, because errors may not become evident until after they have ruined a large data structure compiled at great expense and effort. Speed is also important to economical real-time implementation. This method has proved to be reliable and fast in solving problems involving equations up to the sixth degree in three or four variables. However, there are obvious practical limitations on the number of equations and their degree, due to the limited precision of computer arithmetic and computer resource availability.

THE FIGURES illustrate the transition from simple $G(x)=0$ to final $F(x)=0$. In both figures, the "simplicity" of $G(x)=0$ is reflected graphically in its linear structure—seen as lines and planes. The non-linearity of $F(x)=0$ is seen in the curvature of the final shapes in each figure.

In figure 1, the four dots on the left plane represent the set of simultaneous solutions to the system of equations $G(x)=0$. The four dots on the right plane represent the set of simultaneous solutions to the system of equations $F(x)=0$. The dashed lines represent simultaneous solutions to intermediate systems whose graphs would show the evolution from one configuration to the other. With the addition of a third dimension in figure 2, the number of dots representing simultaneous solutions doubles. Representation of the transitional points, as in figure 1, would require a fourth dimension.

"Continuation methods, although well known to mathematicians," says Dr. Morgan, "are not widely used in science and engineering. Acoustics, kinematics and non-linear circuit design are just a few fields that could benefit immediately. I expect to see much greater use of this mathematical tool in the future."

## THE MAN BEHIND THE WORK

Dr. Alexander Morgan is a Senior Research Scientist in the Mathematics Department at the General Motors Research Laboratories.

Dr. Morgan received his graduate degrees from Yale University in the field of differential topology. His Ph.D. thesis concerned the geometry of differential manifolds. Prior to joining General Motors in 1978, he taught mathematics at the University of Miami in Florida and worked as an analyst at the Department of Energy's Savannah River Plant in South Carolina.

While serving in the U.S. Army, Dr. Morgan participated in the development and analysis of simulation models at the Strategy and Tactics Analysis Group in Bethesda, Maryland.

Dr. Morgan's current research interests include the qualitative theory of ordinary differential equations and the numerical solution of non-linear equations.

is permissible only when a rule or a set of rules makes it permissible. Otherwise whatever is not explicitly prohibited or regulated by a rule is allowed and unregulated (as opposed to the maxim "All is forbidden except what is explicitly allowed").

*Mutable Rules*

201. Players must alternate in clockwise order, taking one whole turn apiece. Turns may not be skipped or passed, and parts of turns may not be omitted. All players begin with zero points.

202. One turn consists of two parts, in this order: (1) proposing one rule change and having it voted on, and (2) throwing one die once and adding the number of points on its face to one's score.

203. A rule change is adopted if and only if the vote is unanimous among the eligible voters.

204. If Initial Rule 203 is amended or repealed, then whenever rule changes are adopted without unanimity, the players who voted against such rule changes receive 10 points apiece.

205. An adopted rule change takes full effect at the moment of the completion of the vote that adopted it.

206. If any player's proposed rule change is voted down, that player loses 10 points.

207. Each player always has exactly one vote.

208. The winner is the first player to achieve 100 (positive) points.

209. At no time are there to be more than 25 mutable rules.

210. Players may not conspire or consult on the making of future rule changes unless they are teammates.

211. If two or more mutable rules conflict with one another, or if two or more immutable rules conflict with one another, the rule with the lowest ordinal number takes precedence.

If at least one of the rules in conflict explicitly says of itself that it defers to another rule (or type of rule) or takes precedence over another rule (or type of rule), such provisions must supersede the numerical method of determining precedence.

If two or more rules claim to take precedence over one another or to defer to one another, the numerical method must again govern.

212. If players disagree about the legality of a move or the interpretation or application of a rule, the player to the right of the one moving is to be the Judge and decide the question. (Such a process is called invoking Judgment.) The Judge's Judgment may be overruled only by a unanimous vote of the other players, taken before the next turn is begun. When Judgment has been invoked, the next player may not begin his or her turn without the consent of a majority of the other players. If a Judge's Judgment is overruled, the player to the right of the Judge becomes the new Judge for the question, and so on, except that no player is to be Judge during his or her own turn, or during the turn of a teammate. Unless a Judge is overruled, one Judge settles all questions arising from the game until the next turn is begun, including questions as to his or her own legitimacy and jurisdiction as Judge. New Judges are not bound by the decisions of old Judges. New Judges may, however, settle only those questions on which the players currently disagree and that affect the completion of the turn in which Judgment was invoked. Disagreement, for the purposes of this rule, may be created by the insistence of any player.

213. If the rules are changed so that further play is impossible, or if the legality of a move is impossible to determine with finality, or if by the Judge's best reasoning, not overruled, a move appears equally legal and illegal, then the first player who is unable to complete a turn is the winner.

This rule takes precedence over every other rule determining the winner.

There you have the rules of Nomic. A friend of mine commented after reading them, "It won't ever replace Monopoly." I grant the truth of that, but it is certainly more interesting than Monopoly to contemplate playing! To make such contemplation even more intriguing, Suber, who has actually played this crazy-sounding game, offers a wide variety of suggestions for interesting types of rule changes. Here are some samples:

Make mutable rules easier to amend than immutable rules by repealing the unanimity requirement of Initial Rule 203 and substituting (say) a simple majority. Add new tiers above, below or in between the two tiers with which Nomic begins. Make some rules amendable only by special procedures ("incomplete self-entrenchment"). Devise "sunset" rules that automatically expire after a certain number of turns. Allow private consultation between players on future rule changes ("logrolling"). Allow secret ballots. Allow "constitutional conventions" (or "revolutions") in which all the rules are more easily and jointly subject to change according to new, temporary procedures. Put an upper limit on the number of initially immutable rules that at any given time may be mutable or repealed.

Allow the ordinal numbers of rules to change in certain contingencies, thereby changing their priorities. Or alter the very method of determining precedence; for example, make more recent rules take precedence over earlier rules, rather than vice versa. (In most actual legal systems the rule of priority favors recent rules.)

Convert the point-earning mechanism from one based on chance to one based on skill (intellectual or even athletic). Apply a formula to the number on the die so that it will increase the number of points awarded to any player whose proposal gets voted down or whose judgment gets overruled but will decrease the number of points awarded to a player who votes nay, who proposes a rule change of more than 50 words, who takes more than two minutes to propose a rule change, who proposes to transmute an immutable rule to a mutable rule or who proposed a rule that was enacted but is later repealed.

Introduce a second or third objective, for example a cooperative objective to complement the competitive objective of earning more points. Hence each player might on each turn contribute a letter to a growing sentence, a line to a growing poem, a block to a growing castle and so on, the group as a whole trying to complete the thing before one of them reaches the winning number of points. Or introduce a second competitive objective, such as having each player make a move in another game, with the winner (or winners) of the game that is finished first obtaining some predetermined advantage in the game that is still being played. Or make any aspect of the game conditional on the outcome of a different game, thus incorporating into Nomic any other game or activity that can muster enough votes. Similarly, leave Nomic pure but add stakes or drama (such as psychodrama).

Institute team play. Require permanent team combinations or allow alliances to shift according to procedures (informal negotiation, an algebraic formula applied to scores or systematic rotation of partners). Create "hidden" partners (for example, the points a player earns in a turn are also added to the score of another player, or split with one, selected by a mechanism).

Extend the aptness of the game as a model of the legislative process by inventing an index that goes up and down according to events in the game and that measures "constituency pressure" or "constituency satisfaction"; use the index to constrain permissible moves (for example through a system of rewards and penalties). Allow a certain number of turns to pass before a proposal is voted on, giving the players the opportunity to see what other proposals may be adopted in its place.

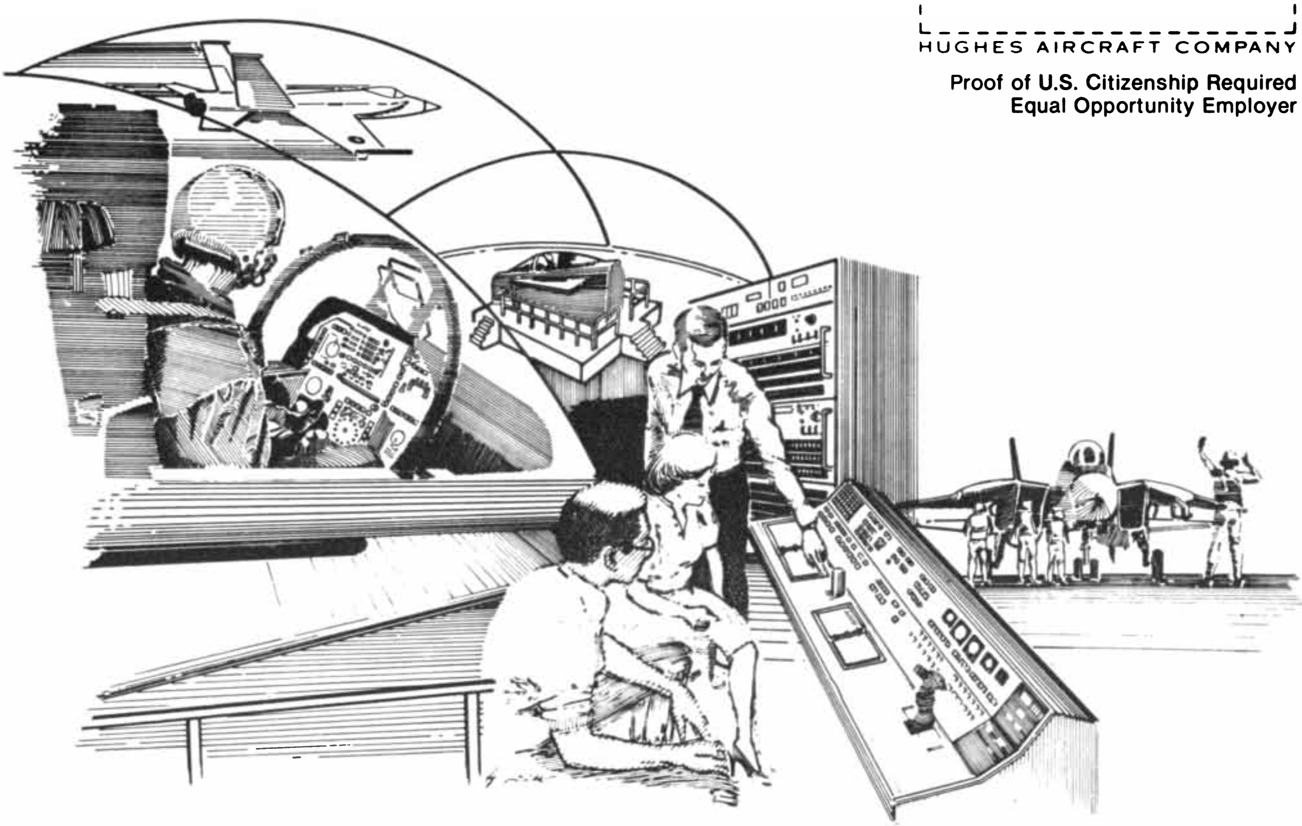
Suber is interested in hearing from readers about their experiences in playing Nomic, as well as any suggestions for improvement or comments on reflexivity in law generally. His address is Department of Philosophy, Earlham College, Richmond, Ind. 47374.

Suber's ultimate challenge to players of Nomic is this: to ascertain whether any rules can be made genuinely immutable while preserving some rule-changing power, and whether the power to change the rules can be irrevocably and completely repealed.

The richness of the Nomic universe is abundantly clear. It certainly meets every hope I had when in my book *Gödel, Escher, Bach: An Eternal Golden Braid* I wrote about self-modifying games. It was my purpose there to describe such games in the abstract, never imagining that anyone would work out a game so fully in the concrete. It had been a dream of mine for a long time to devise a system that was in some sense capable of modifying every aspect of itself, so that even if it had what I referred to as inviolate levels (corresponding roughly to Suber's "immutable" rules), they could be modified as well.

I vividly remember how the dream came about. I was a high school student when I first heard about computers from the late George E. Forsythe, who was then professor of mathematics at Stanford. (There was no such thing as a department of computer science at the time.) In his talk to our mathematics class he emphasized two things. One was the notion that the purpose of computing was to do anything people could figure out how to mechanize. Thus, he pointed out, computing would inexorably make inroads on one new domain after another, as we came to recognize that an activity that had seemed to require ever fresh insights and mental imagery could be replaced by an ingenious and subtly worked-out collection of rules, the execution of which would then be a form of glorified drudgery done at the speed of light. For me one of Forsythe's most stunning illustrations of this notion was the way computers had in some sense been applied to themselves, namely with compilers, programs that are designed to translate programs from an elegant and human-readable language into the cryptic strings of 0's and 1's of machine language.

The other notion Forsythe emphasized—and it was closely related to the first one—was the fact that a program is just an object that sits in a computer's memory and as such is no more and no less subject to manipulation by other programs (or even by itself!) than mere numbers are. The fusion of these two notions was what gave me my inspiration to design an abstract computer. (Playing on the names of the ENIAC, ILLIAC, JOHNNIAC and other computers I had heard of, I called it IACIAC.) I hoped IACIAC not only could manipulate its own programs but also could redesign itself, change the way it interpreted its own instructions and so on. I quickly ran into difficulties and never completed the project, but I have never forgotten the fascination. It seems to me that although it is a game and not a computer, Nomic comes closer in spirit to the goal I sought than anything I have ever encountered. That is, except for itself.



# First the bad news.

Our profits will be down in 1982 after two record-setting years.

The recession is one of the reasons why.

# Now the good news.

The other reason is we're giving up some of our earnings to invest more heavily in future growth.

We could add more than $2 a share to 1982 earnings just by holding research and development and oil and gas exploration spending at 1981 levels.

But we're not going to do that. We're going to spend more on research and development than ever before. About $200 million. 32% more than last year.

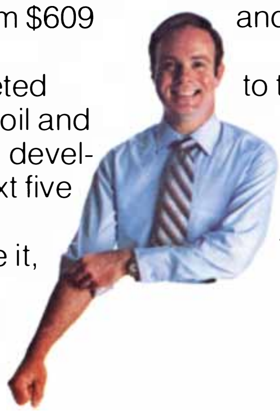
We're also increasing our capital spending to about $700 million. Up from $609 million in 1981.

And we've budgeted almost $3 billion for oil and gas exploration and development over the next five years.

The way we figure it, the choices we're making now are going to pay off later.

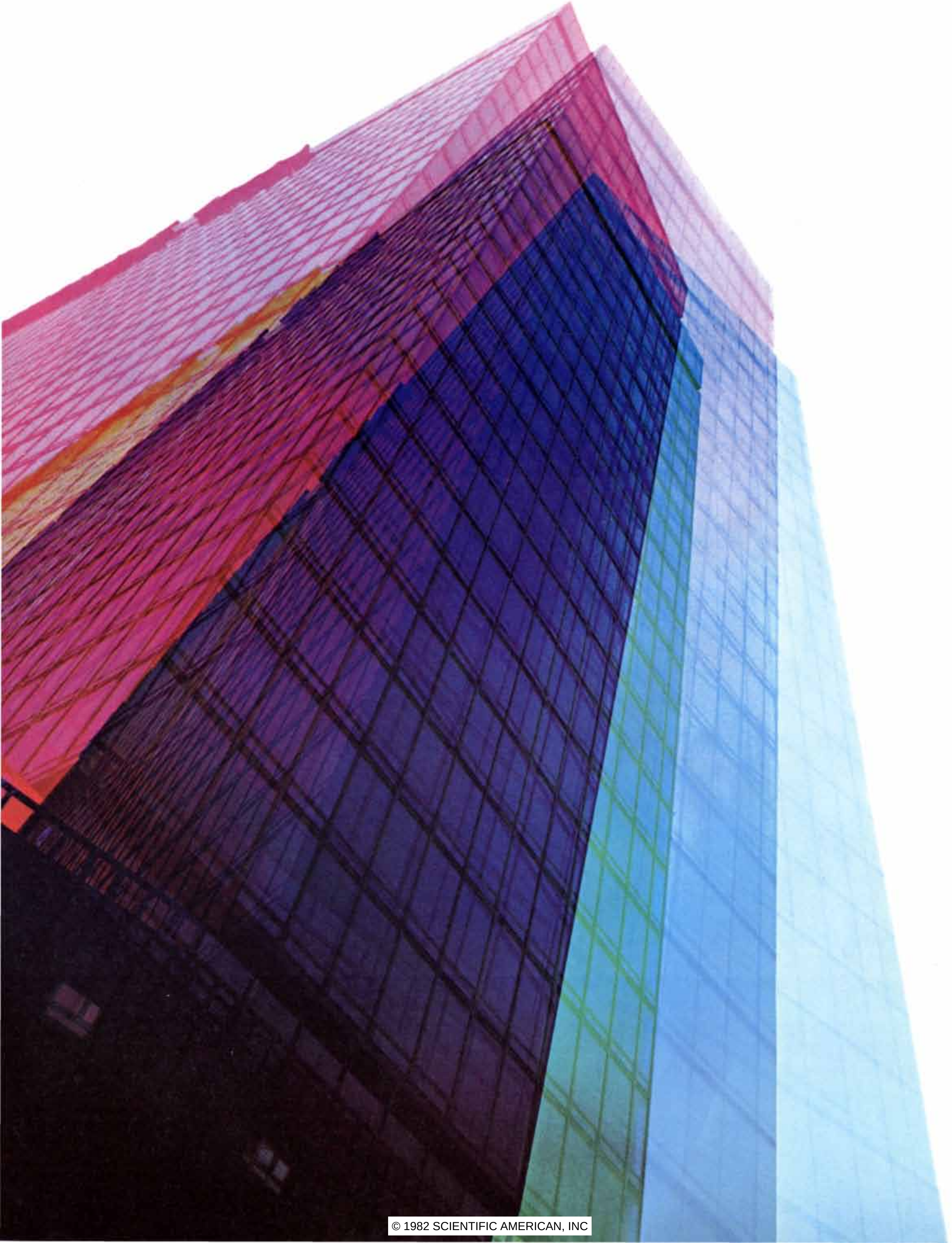
In the next decade, we expect our growth to be more and more profitable.

Then we'll have nothing to tell you but good news.



ALLIED Corporation

**We mean business.**



# BOOKS

## *Inexhaustible Newton, chip manufacture, the technology of cheese, marine geology*

by Philip Morrison

NEVER AT REST: A BIOGRAPHY OF SIR ISAAC NEWTON, by Richard S. Westfall. Cambridge University Press ($49.50). It is 125 years since the appearance of the "reigning biography" of Isaac Newton, Sir David Brewster's two volumes of Victorian adulation and scholarship. Now we have this hefty but lively modern volume, the fruit of 20 years of work by a distinguished Indiana historian of science. The book is a scientific biography, with Newton's career as a man of learning and research at the core. It is the product of a post–World War II industry of Newton studies, a phenomenon that has brought seven volumes of Newton's letters into print, together with eight volumes of his mathematical manuscripts and a number of fine books on particular Newtonian themes, well documented, perceptive and fresh.

There is still no shortage of raw material: Professor Westfall provides a footnote to establish his own firsthand estimate of the quantity of Newton's manuscripts on alchemy. They alone amount to six or eight substantial books from the writing desk and the laboratory bench; one book-length study has appeared, without new laboratory work. In Jerusalem there is a large collection of Newton's theological and historical papers; in all there are five major collections in his hand. Newton worked steadily over a long lifetime, never at rest. The title of this work is Newton's, taken from a letter he wrote at the age of 51: "A Vulgar Mechanick can practise what he has been taught or seen done, but if he is in an error he knows not how to find it out and correct it...; Whereas he that is able to reason nimbly and judiciously about figure, force, and motion, is never at rest til he gets over every rub."

The newer studies of Newton traverse three main roads. Most of the books have sought to tease out method and influence, assumptions and philosophy. How did he come by this idea or that? The painstaking study of texts lies along this path. We have a two-volume version of the *Principia* comparing the variations among the three editions Newton saw into print. The notorious controversies over priority are encountered along the way. A second path has brilliantly sought out the psychology of the man, so solitary, work-possessed, guarded, fearful, jealous, at times tyrannical. But the route traveled here is taken at a more leisurely pace, with an eye on the landscape; it is a contextual study, setting Newton in his time. What did others say of his claims? Did Cambridge come to a halt while the plague closed down the colleges? So regarded, myths seem a little reduced in strangeness, the inner life a little less hyperbolic, the lonely spirit nearer balance.

Schoolboy Newton was remembered by everyone for his "strange inventions." Sundials he had made filled the house where he lodged. (Even in old age he would "look at the shadows instead of the clock to give the time.") He tied a paper lantern to a kite tail to frighten the neighbors; he built a model windmill powered by a mouse treadmill and a little cart he could run by cranking as he sat in it. "We know now that Newton found many of these contrivances in a book called *The Mysteries of Nature and Art* by John Bate." That antique volume fed his fertile interest as its equivalents in every generation since have nourished the boys and girls who make their adult careers around the laboratory.

Newton was admitted to Trinity College in 1661. The universities had by then become reservoirs of patronage; between them they controlled through their degrees the benefices of the state church. "Cambridge University, like Oxford, drifted toward the status of a degree mill exploited without conscience by those fortunate enough to gain access to it." True, Cambridge had been the center of English Puritanism, and after the Restoration it plainly had to show that it had purged itself of error. Most of the heads of colleges were replaced; their jobs were the juicy morsels of political spoils. The student body, however, grew apace, reaching again the peak size of the days of James I. The life of the mind was all but absent. There is a story that when Newton left home, his educated uncle gave him a book of logic, saying it would be the first book his tutor would read to him, as an earlier tutor had done for the uncle as an undergraduate 30 years before. The prediction was accurate, but there was no vigor left in Cambridge's formal Aristotelianism. It was rote. The disputations had become rituals; even the lectureships were sinecures. No one came to the lectures, and often the lecturers themselves did not appear.

The degree statistics show no signs of decrease during the dispersal of the university for two plague years. Newton never finished his assigned texts; no one cared. Examinations were lax, although for Trinity fellowships they continued. Newton somehow converted his wretched subsizarship (the lowest of grants, requiring its poor student holders to work as valets and waiters for the fellows and wealthier students) into a four-year fellowship in 1664. It seems sure that this preferment came about through the intercession of some powerful advocate.

Isaac Barrow was the one man at Trinity who had the competence to assess Newton's mind and work. Five years later Barrow would write to identify the author of a paper on infinite series: "His name is Mr Newton, a fellow of our College, & very young (being but the second yeest Master of Arts) but of an extraordinary genius & proficiency in these things." Barrow that year resigned the well-endowed Lucasian chair he held in order to obtain it for young Newton. Barrow was eager for, and soon received, other preferment; within three years he became master of the college. It was almost surely Barrow who in 1675 secured a royal dispensation exempting, beginning with incumbent Newton, the Lucasian professors in perpetuity from the need to take holy orders.

That requirement would have been all but impossible for Newton then; he had become in secret totally disaffected from the Trinitarian creed of the Church of England. His refusal not only would have cost him his job but also would have left him vulnerable to the exposure of a kind of heresy that was socially damning in that time of religious strife. Because of a misunderstanding during an oral examination, however, Newton believed that in 1664 Barrow had "but an indifferent opinion of him." Perhaps Newton was misled about Barrow's judgment; perhaps a shadowier senior fellow, a relative of the household where Newton had boarded in his school days, was responsible for that first recognition. We do not know.

Between 1664 and 1667 (with one year off campus) Newton found and mastered Descartes's new analytic geometry, set down questions in his notebooks that presage his entire experimental future in optics and mechanics and made his way self-taught to the "resolution of problems by motion," the idea of the calculus. For Westfall these are three miraculous years, not the traditional single annus mirabilis. They do not leave the 24-year-old Newton in command of "the results that have made his reputation deathless, not in mathematics, not in mechanics, not in optics. What he had done in all three was to lay foundations." What of the falling apple? It seems real, but the record does not support the notion that he carried the full idea of universal gravitation in secret for 20 years. His own notes deny the possibility. That he had some inkling, some half-contradictory hint of magnificent generality, remains credible. Universal gravitation is not, after all, some mere spark of insight; there would be "time to think on it."

The long life unfolds before the reader, always against a full stage. In his London career the now-great worthy is a serious civil servant but no longer a profound original, either in religion or in mechanics. He changed status with the Glorious Revolution of 1688, in which he took some public risk. The *Principia* was then newly set before the world, thanks largely to tireless Halley. One can read of Newton's niece, a young woman of beauty and wit (Swift became her intimate friend, and he once wrote to Stella "I love her better than any one here") who was celebrated in 1703 at the Kit-Kat Klub in baroque verse diamond-scribed on a toasting glass. Such were the pleasures of the great families and their friends who made up the ruling Whig Junto of the times. Worldly Voltaire held that Newton's high post had been gained through the conquest of the minister Halifax by the charmer. "Fluxions and gravitation would have been of no use without a pretty niece."

Conquer Catherine Barton had; evidence is compelling that she became either mistress or wife to Halifax in secret. The brilliant and scandalous minister had, however, already been Newton's close friend for many years; although the talented young man was no student of science, at Cambridge he had joined with Newton and a few others to form the Philosophical Society. As Chancellor of the Exchequer he first named Newton to a post at the Mint, but there is no sign that the "famous witty Mrs. Barton," then a mere girl, had had anything to do with an appointment for which Newton was plainly suited and on which he had some public claim by political service. The affair came years later, with the Newtons already rich and soundly set in London society.

Here is the older Newton as expert parliamentary witness on finding the longitude. He was not very hopeful about good clocks, but he listed all the possibilities we have today, except inertial guidance. (The time signals then proposed were acoustic and of rather too short a range. Still, it is near the coast that most danger lies.)

Newton died at 86. For 15 years he had laundered his most unorthodox views out of the theological works he intended for publication. He died, however, with a gesture; he refused the final sacrament of the Anglican church. A covert man, all his triumphs must have seemed to him incomplete. "He was greatly concerned to leave his image behind him." No one save royalty sat so often for portraits. There are some 20 of them in this book, with about 30 other interesting images, mostly pages and drawings from Newton himself. (We miss any likeness of Mrs. Barton.) This ample volume somehow does not resolve the drama. Not the legacy of Newton's own words by the million, not the careful piecings of the scholars nor their psychoanalytic insights can make clear this unequaled mind, so charged, powerful and lonely.

One impression comes clear. It is only a partial knowledge that was the true fruit of Newton's work. The savants of the time, mechanical philosophers all, lived under the sway of a universal preconception. The occult properties and tendencies of the school philosophers could no longer be tolerated; matter must move only by the mediation of subtle particles in collision. When the gifted Christiaan Huygens was told that the *Principia* was in preparation, he replied: "I don't care that he's not a Cartesian as long as he doesn't serve us up conjectures such as attractions." But attractions we got, and with them profound quantitative and genuine—but incomplete—understanding, widely serviceable unaltered for 250 years. It could be argued that Newton's long hours among the firebrick furnaces he built with his own hands, seeking in the crucible the occult wisdom of the alchemists, showed him that the material world was subtler than the naive models of the Continental thinkers. Until he lost confidence in the soundest metaphysics of his time he could not make real progress.

INTEGRATED CIRCUIT FABRICATION TECHNOLOGY, by David J. Elliott. McGraw-Hill Book Company ($27.50). In our rough times hardly a princess remains to discern the uncomfortable pea under her mattress. Just such an intrusion on a much smaller scale, however, is the occasion for an explicit warning in this detailed engineering treatise on the nicest production technology in history. The silicon wafer clamped down by suction against its vacuum chuck must be dust-free not only on its receptive lapped, polished and etched mirror surface but also on its backside. A speck a couple of micrometers across caught there might bend the front surface just enough to defocus the image that shapes the 100 or 200 intricate little chips the three- or four-inch crystalline disk is expected to yield when it is finally cut apart along neat diamond- or laser-scribed lines.

From an initial outline of the steps between raw silicon and microprocessor chip through detailed chapters on the hows and whys of all the remarkably sophisticated processes this book can carry a reader who has a broad if nonmathematical background in applied physics to a working position along the microconstructional route. This is the world of hardware and its skillful application; there is hardly a word on computer architecture or design and few on the physics of the semiconductor. The tone is high and steady: Give us the design on paper or magnetic tape and we shall finish the work, with a sharp eye to the improvement in the yield of satisfactory chips, the result of tireless quality control and quick error diagnosis. It is there that economy lies. The little city streets so carefully mapped out in patches of doped semiconductor silicon, in lattice-matched thin layers, in more layers of dielectric silicon dioxide and in aluminum conducting circuitry must support a well-controlled electronic traffic.

This production is of course dominated by its tiny scale. Nowadays a good chip may bear 1,000 or 2,000 circuit elements on each square millimeter. The error one can tolerate on each side of every design line is perhaps 15 percent: the best production resolution for such a line today is about 1.8 micrometers. The limit is currently fundamental enough; it is set by the wavelength of the photolithographic radiation. Masks of chromium on glass exposed in the near ultraviolet fix the present possibilities: strong mercury lamps, tended by complex fiber-optics collection systems hungry for photons in any direction, allow reduced projections to expose the photoresist, stepping out hundreds of multiple images at rates not far from a wafer per minute. To be sure, improvements lie ahead. The factor of two in size that is optimistic engineering realism might almost come from exposure in the far ultraviolet; quartz masks will replace glass, and mercury discharges might be supplanted by xenon (or deuterium) lamps.

Electron beams and X rays are obvious next steps. Indeed, the scanning electron beam is already very popular for mask making, but so far it is too slow for chip production. X rays and ion beams are in development; the natural resolution limit is then transcended, although more practical limits arise. The author soberly sums it up as "unlikely that it will be economically feasible to manufacture [integrated circuits] with... line widths of 0.5 to 0.8 [micrometer]." That is a circuit density about 10 times better than the best present production.

Much of the detail given here is directed at the key photolithographic steps. The best resists today are called positive, that is, the developer acts to remove those areas that are exposed to light. The mask pattern is formed by unexposed material.

Precision is limited by isotropy; it is hard to etch away a surface and leave a steep wall. Many scanning electron micrographs here show test circuits with marvelously steep walls and deep-cut grids. Plasmas and ion beams as etchants have made possible quite new approaches. The etched material is most often a micrometer-thick layer of silicon dioxide, formed either by vapor decomposition or by slower thermal growth in hot oxygen at pressure. The patterns produced by etching this amorphous layer are generally used as masks for doping the needed transistors, diodes and resistors on patches of bulk crystalline silicon, either by diffusion or by direct implantation of boron or phosphorus ions with scanning ion beams. Aluminum is etched in the dry way with a plasma of carbon tetrachloride and helium; the product is the powdery chloride.

The standards for this technology are clearly high. The very wafer on which it all rests is a sawn slice of a single-crystal ingot of ultrapure zone-refined silicon. It is routine enough (a neat automatic optical polarimetric machine is offered for the purpose) for the producer to map out the surface of every wafer to micrometer accuracy. The sensitive "front end" of the process, where the photoresist is spun onto the much-cleaned wafer, is of course secluded in clean rooms, their air and water lines filtered and irradiated against dust and bacteria. The employees wear hat, gloves, coat and boots, and yet they still shed lint and particles. Smokers may give off two million particles after the smoking break; nonsmokers shed some tens of thousands per hour.

Lasers and scanning electron microscopes, pure dry nitrogen and helium, reliable automated vacuum systems, ion- and electron-beam chambers are commonplaces in this industry; its production machines are the output of a specialized and innovative associated industry stretching from Osaka to Geneva. Although not every operation can as yet be made continuous, the wafers are often passed along automatically two dozen at a time in special cassettes. There are efforts at computer control of all processes, based on formal models of the measured effects of the myriad interacting parameters that define the entire long chain. It is not much of a trick, it seems, to make a few chips at high cost. It is generating them by the million that is a feat, which seems to be managed best with a step-by-step extension of rational and measured control. This is a glimpse behind the scenes of a major industry at the highest innovative level. Meant for the responsible engineer, it is a good keyhole for the general reader, even though it is not intended for us mere construction watchers.

CHEESEMAKING PRACTICE, by R. Scott. Applied Science Publishers Ltd., London. Distributed in the U.S. by International Ideas Inc., 1627 Spruce Street, Philadelphia, Pa. 19103 ($49). The heavy chemical industry can produce simple molecules in bulk under superb control; the fabrication of integrated circuits, artifacts under close control at the micrometer level, has passed beyond art. But the cheesemakers of the entire Temperate Zone (their output has the same order of value as that of the chip makers) must deal both with living macromolecular complexity and with a product whose acceptance depends on the perceptions of human senses. Their activity is far from rationalized; craft and art still reign, based on a deal of experience handed down from Sumer and before.

This is not to say that we know nothing. This interesting volume is the work of a man with 50 years' experience at the research and administrative levels in the United Kingdom; he knows enzymes, ultrafiltration and bacteriophages no less than he knows the big milk pan on the stove in the upland farmhouse kitchen. Cheesemaking includes all that and much more. Dr. Scott offers an up-to-date overall summary of the art, science and trade of making cheese, with ample references for those who would read more. Of course, the school of experience is strongly recommended.

The coming of laboratory science into this old craft is a phenomenon of the late 19th century. Louis Pasteur initiated the purposeful heat treatment of raw milk against bacterial growth in about 1870; pure cultures gradually took over the role of the natural souring starter from the 1880's; the long-traditional rennet from the calf stomach was extracted and standardized first in Denmark (1870), and F. J. Lloyd pioneered the titration of acid content in cheddar production at the end of the century. Today there is even a diplomacy of cheese, with regulations adopted for international trade, a machine-making industry that produces stainless-steel heat exchangers and mixing vessels in large sizes and wide variety, and a body of studies of the kinetics and interactions among the many determinants that part a Brie from baker's cheesecake curd or from a mechanically kneaded fibrous mozzarella.

It starts, of course, with milk, from cow, goat, sheep, yak, buffalo, camel, horse and ass. "Milk is not, as some people imagine, a defined substance M-I-L-K"; it is an entire ecology, varying in time and place, loaded with its own microflora. Cheese takes on the cultural patterns of the herdsman and the herd; there are some 2,500 named varieties of cheese, among which Scott reckons there might be 60 distinct types of curd. Cheese begins as curd: the coagulated casein proteins of milk, with varying amounts of milk fat, soluble proteins, salts, lactose and its ferment, lactic acid. That coagulation can be induced by mere electrolyte change—by its own souring or by added ions—or more subtly by enzymes that link the milk proteins at specific sites. Nowadays the natural souring of milk is usually preempted by adding a controlled bacterial culture to make what happens surer (be wary of virus-infected starter); the coagulant too can be either the preferred extract of the calf stomach, the enzyme chymosin or the faster-acting and temperature-tolerant enzymes derived from bacteria and now widely sold.

Curd is not yet cheese; it must ripen. This stage depends on the prior mechanical treatment, on kneading, pulling and pressing, on wrappings and storage conditions. While a rich flora of bacteria and mold grows slowly within the cool stored cheeses, proteolytic enzymes can change the tough curd to thick aromatic cream within a rind itself host to other varieties of microorganisms. Gas bubbles, blue-veined colonies of molds, added salt, brine soaking, smoke, grapeseed coatings—the taxonomy is rich.

Cheeses can be made even from whey, the milk fraction remaining after the protein curd has been removed. (Miss Muffett took both.) This idea seemed to have escaped legal notice for a time; whey, dilute to be sure, nonetheless holds plenty of values. Whey cheese is made by a craft long known in Norway; nowadays vacuum evaporation may speed the work and save fuel. There are many soft cheeses based on lactic acid coagulation. They include a very popular European product made out of curd from which the whey has been centrifugally separated. It is called *quark* in a number of European tongues; very likely they knew it in Trieste when James Joyce lived there, even though the theoretical physicists of California saw in the word only a Joycean triplicity.

MARINE GEOLOGY, by James P. Kennett. Prentice-Hall, Inc. ($34.95). MINERAL DEPOSITS AND GLOBAL TECTONIC SETTINGS, by A. H. G. Mitchell and M. S. Garson. Academic Press ($49.50). A representative spot on the surface of the earth is naturally a seascape. With that mineral fluid the hammer of the geologist is of no service, nor can he walk out the formations. Geology was truncated until this deficiency could be remedied. Most of what we know about marine geology is less than 30 years old; *Marine Geology* is the first comprehensive survey at a nonmathematical and introductory level. Few other broad texts in science draw the bulk of their data from original publications less than a decade old; we recognize here the consolidation of a victorious revolution. The geologists went to sea.

Harry Hess appears here in a photograph as a navigation officer on a troopship in World War II. He was in the early 1960's a pioneer of the concept of sea-floor spreading and destruction that added the essential element to Alfred Wegener's long-rejected idea of conti-

You're not just moving in,
you're moving up.
IMPORTED
Canadian Club
C.C.
Blended Canadian Whisky
Distilled and Bottled under
Canadian Government Supervision by
Hiram Walker & Sons
Limited
Walkerville, Ontario
THIS WHISKY IS 6 YEAR
86.8 U.S. PROOF
Canadian Club®
"The Best In The House"®
6 Years Old. Imported in Bottle from Canada by Hiram Walker Importers Inc., Detroit, Mich. 86.8 Proof. Blended Canadian Whisky. © 1982

# You should be thinking about the problem, not about the computer.

Thirty years ago, engineers and scientists found a workhorse computer to meet their needs: the IBM 701.

Today, the descendants of the 701 are the most effective IBM systems ever: they help you get answers quickly, cost-effectively, and so easily that you'll almost forget you're using a computer.

For example, IBM's Virtual Machine/Conversational Monitor System (VM/CMS) lets you set up and enter a job and monitor its progress. And intervene, if necessary, to change a parameter to drive the whole process toward better results, faster.

Under VM/CMS, even the IBM 4341 super-mini can support as many as 200 interactive users at a time.

### VM/CMS Easily Learned

One user of VM/CMS on a 4341 is the University of Pennsylvania, where Roy Marshall directs the Physics Department Computer Facility. "Users learn it very easily," Marshall says. "They do simple things the first day. They flow with the problem-solving: A physicist does physics, not computer science."

"The editor is the most powerful I have ever used," says Dr. Richard Steinberg, who is pursuing a proton decay experiment. "I can get any file–data, programs, text–with two keystrokes. With the prompting system, I can enter a big job and *know* it will run. I won't find out the next morning that there was a job-entry error."

### A Super-Mini with Punch

The IBM line of upwardly compatible computers extends from the small 4331 to a giant system. The 4341 is a true super-mini in size and cost, yet it has every feature that makes its larger counterparts ideally suited as engineering and scientific systems: for example, the 64-bit data paths and the rich set of 51 floating-point instructions.

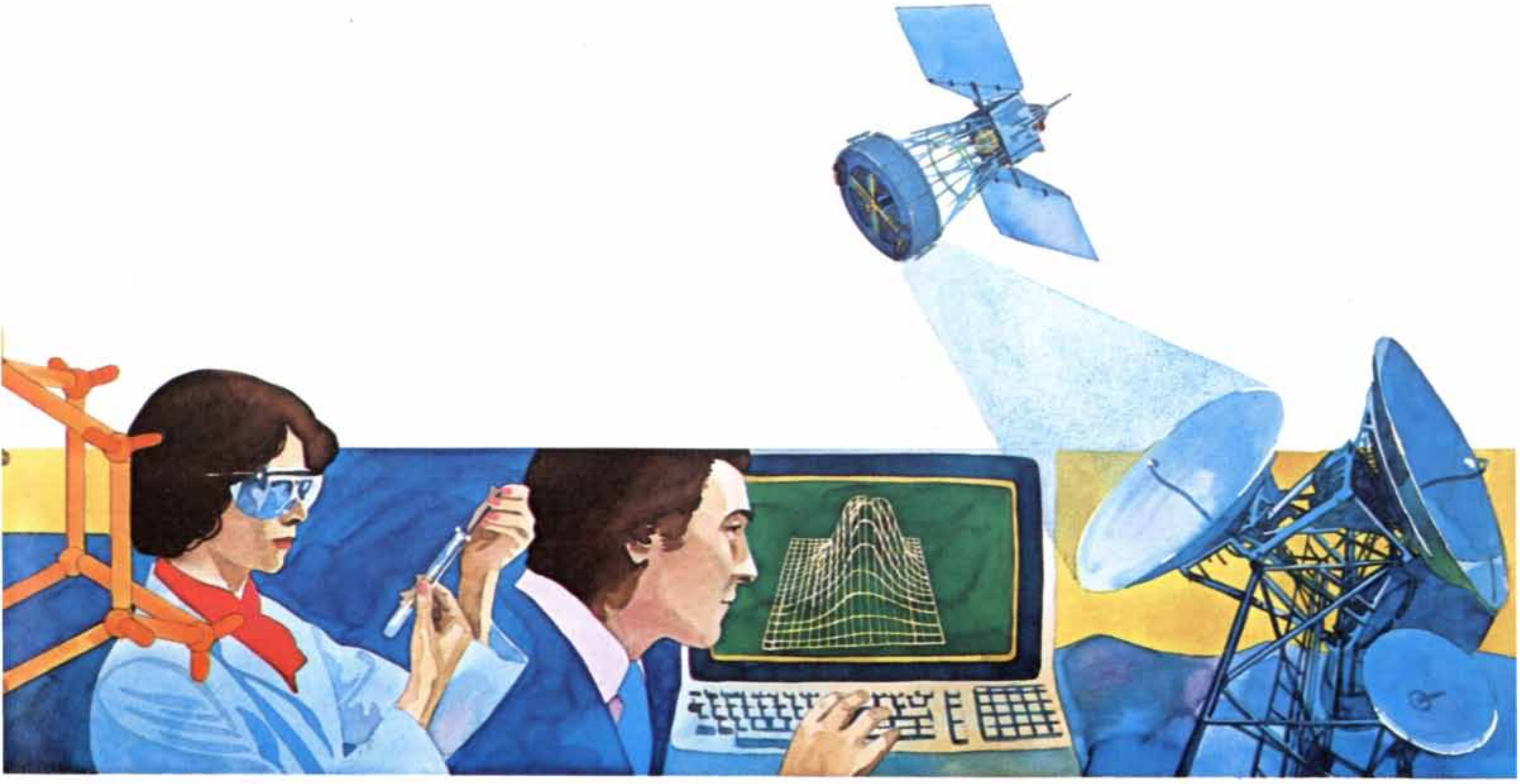
### The Top of the Line

IBM's largest computer is the 3081. With its processing speed, memory size, and the speed and number of its I/O channels, the 3081 is designed to handle the massive problems presented by such disciplines as elementary particle physics.

At the Stanford Linear Accelerator Center (SLAC) near Palo Alto, a 3081 reduces a flood of recorded sensor data.

"CPU power is critical to us, in terms of saving the scientist's time," says Charles Dickens, director of computing services at SLAC. "Under VM, he can look at intermediate results and–if necessary–change the physical experiment or the calculation.

"And the ability to move sensor data rapidly from our instrument tapes is vital. We need



the fast channels and high-performance peripheral devices of the IBM system."

**This One Grows on You**

The IBM 3033, upgradable over a 4:1 range of computing power, covers the span from the super-mini to IBM's biggest.

At Western Geophysical Company, nine 3033 systems with attached IBM array processors are absorbing a torrent of data from geophysical exploration all over the world. From a jumble of seismic echoes, they derive the hidden contours of underground rock layers.

At Western's headquarters in Houston, senior vice president Carl H. Savit explains: "To improve the signal-to-noise ratio and arrange the data for the required series of calculations, we perform massive data sorts. The rapid channel rates of the 3033 are essential to us.

"Our product is data," Savit adds. "The computer is our production machinery. We depend critically on continuous operation and quick response. We need close support from our computer vendor, and we get it from IBM."

**The Mini You Get Attached To**

IBM's minicomputer for sensor-based processing is the Series/1: versatile, powerful, compact–and low in cost. Designed specifically for such applications as instrument control and online recording, it offers simple attachability for a wide variety of devices and excellent software for real-time control.

**A Personal Computer with Color**

A new addition to the IBM line is the Personal Computer: a desk-top machine with a remarkable full-color display system and superior graphics. It can communicate with any IBM host system, to access a data base or program library, or to submit a job remotely, as easily as it solves problems as a free-standing computer.

**The Support You Expect**

IBM offers the engineering and scientific user extensive support: consultants, educational programs, and access to professionals.

Tap into this 30 years' accumulation of experience. Choose an IBM system that meets your needs today and lets you grow tomorrow. A system accompanied by the service and support you expect from the leader.

For more information, contact Dr. Jack W. Hugus, IBM, Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, N.Y. 10604.

IBM®

nental drift. The photograph is more than a souvenir: it symbolizes the war-born technology that underlies marine geology today. If the geologists cannot see the bottom at all, and if they can feel for it only laboriously here and there with a long wire cable, as they did aboard the *Challenger* a century ago, what perception is left? Nowadays they listen. It was the ping of antisubmarine sonar that gave birth to a now mature technology.

First came continuous records of the sea depth under the keel made as the ship was under way. Now it is multichannel seismic-reflection methods that map the strata below the bottom, down through the sediments into the bedrock. "A half-pound block of TNT was dropped over the ship's side every three minutes" during 10 years of work by the pioneers; these days the echoes of steadily repeating high-pressure airgun bursts are picked up by hydrophones strung out along the wake of the ship in a plastic tube kilometers long.

This 800-page book, crammed with maps and diagrams, is divided into four parts. First come the structure of the ocean crust and the watery currents, next the decisive active continental margins and shelves; the sediments of the entire sea bed are described (perhaps the most recent of all the results); finally the history of the oceans is treated overall, its rhythms and changes, mostly within the past tenth of geologic time.

The geometry and physical nature of the great natural lines and arcs that sculpture the sea floor witness and trace out the motions of the rigid plates, which all rotate from time to time about some shifting points. The account given here is clear and full: the steep, hot ocean ridges, the plates submerging and earthquake-beset at collision, the deviations that are the neat arcs of fracture zones cutting across the ridges. On that geographic basis ocean structure and history are built, of course with due attention to the rich magnetic and seismic inferences of the geophysicists. Each of the three great oceans is followed in time. The Atlantic is now mature, after 200 million years of growth. The Pacific, once a world ocean around a single island continent, is dwindling now; although still mighty, it will close entirely in another 200 or 300 million years.

The melodrama of India's collision at full speed with southern Asia is outlined. First came a soft encounter with an island arc; then came the hard collision that shortened the crust, folding up the highest mountains on the planet. Those dizzy heights are thus strangely part of this chronicle of the lowest lands; the depths certainly belong as well. The level undersea plains far from shore, beyond the reach of the silt-laden river flows and the sudden slides, fill very slowly. The *Challenger* first found red clay in the Pacific depths; it covers the bottom of half that ocean. Its fine particles are indeed airborne dust from old lands, mostly mica-derived clay, stained with iron oxide, rust brown. About half of the ocean floor is covered, however, with the carbonate debris of life, a whitish ooze of plankton skeletons, mainly of foraminifera and their kin. Most of these particles dissolve before they can settle to the greatest depths. The ooze lies like snow over the higher elevations of most ocean basins.

Since 1968 a new venture, the International Program for Ocean Drilling, has managed to play the role of the hammer of the field geologist. A specialized ship able to drill at a depth of 20,000 feet, able to core out a plug of the bottom more than 2,000 feet long, this new *Glomar Challenger,* able to grope in the darkness with its sonar scanners to reenter a little drill-hole funnel far below, has sampled the floors of all the oceans except the Arctic. Even more recently a few geologists have themselves gone to look at the spreading ridges close up, in their deep-sea submersibles. The results are powerful. The sea floor is not silent and undisturbed; its stratigraphic record, there in the ooze and the clay, is as vivid with unconformity and change as the geologic sections of many a landscape. It can be analyzed by careful and varied forms of dating, by isotope distribution, by the clear alternation of its clays and carbonate snows.

Often the records show that the sea has locally dried up, leaving salt desert flats. Just such drying has been found in the Mediterranean; a couple of kilometers of thick salt deposits are there very deep, the result of not one but many evaporated basinfuls of seawater. That furrow in the lands has dried up dozens of times. It last filled permanently over the Gibraltar sill about five million years back; oozes and clays make up the newer bottom levels.

The entire book is a treasury of such results and evidence. Any general reader who remembers the school geology will enter a new world: the same careful and clever approach yields new and marvelous narratives, with new tools, in the world of watery darkness. The tale is not ended; all the cores taken over the decade are still scanty. The deep-sea narrative is fairly to be compared with writing the geology of the U.S. from the study of half a dozen drilled samples taken as representing the entire country.

The authors of the second book, *Mineral Deposits and Global Tectonic Settings,* a volume almost as terse and technical as any botany, are two imaginative and experienced economic geologists. Long collaborators, A. H. G. Mitchell and M. S. Garson work now for United Nations agencies in the field, Mitchell posted near Manila, Garson in Tanzania. A cosmopolitan view is characteristic of the subject: gold, like phosphorus or cerium, is where you find it. The authors' theme is as straightforward as its details are intricate: plate tectonics is on the way to becoming an applied science of real predictive power. The deposits of useful minerals worldwide can be seen now in their settings as understood from plate tectonics. Is a mineral province found at an old hot spot, on a passive continental margin, in an ancient ocean ridge or fracture zone, at the locus of a plate's submergence or where two plates collide? Setting by setting, Mitchell and Garson classify and describe the chief known—and not so widely known—ore deposits, from Abakiliki through the Sea of Okhotsk to Zambia, and from antimony to jadeite and zirconium.

This detailed collection of succinct accounts of the ore finds of the world is difficult but fascinating reading for the nonspecialist and a treasury for reference. Through all the detail there rises the sense of a new order out of chaos. The first sign was perhaps given a decade ago by study of the rich copper porphyries that lie at the Pacific edge of the American cordillera, from Chile to British Columbia. Half of the world's copper comes from these ore bodies, whose location and nature pretty surely represent the gas and hot salt water bringing pyritic copper up from the depths, concentrating the scarce but useful ions for a brief spell of geologic time within an active thermal plume. That plume with its gases and water and heat arises during the difficult burial of the ocean floor at a destructive plate edge.

On the other hand, the carbonaceous outpourings of the East African Rift, diamond and washing soda alike, seem to come from convective hot spots that remain for a long time in place. Out of them the slow volcanoes grow and diamond pipes suddenly erupt. It was the fluxlike ash from the white-capped Tanzanian soda volcano depicted here (a third kind of snow?) that preserved long ago the footprints of our hominid ancestors. But the rich ore complex of southern Africa is as tangled geologically as the region now is in human terms.

It looks as though a new worldwide predictive discipline is slowly forming, always modulated by local structures and by the complexities of mineral chemistry. It is probably no accident that the admiring foreword was prepared by an exploration geologist associated with one of the most audacious and successful mining companies in the world. Petroleum belongs to another world of experts; the volume also sets aside the very oldest provinces, those formed before about 2,700 million years ago, because the changing relations of ocean and continent are not clear before that time. Many of the archival periodicals of economic geology are a century old; it is of some interest that there is now one periodical, whose first volume is dated 1980, called *Global Tectonics Metallogeny.*


# Lockheed knows how.

Advances in research tend to make headlines, but breakthroughs in manufacturing can be as important. The scene above depicts a manufacturing breakthrough that can result in easily fabricated buildings in space. That ingenious advance and others more Earthbound are reported below.

## Making building blocks for space.

Someday, astronauts will build a structure in space, and they may well use a new Lockheed manufacturing process.

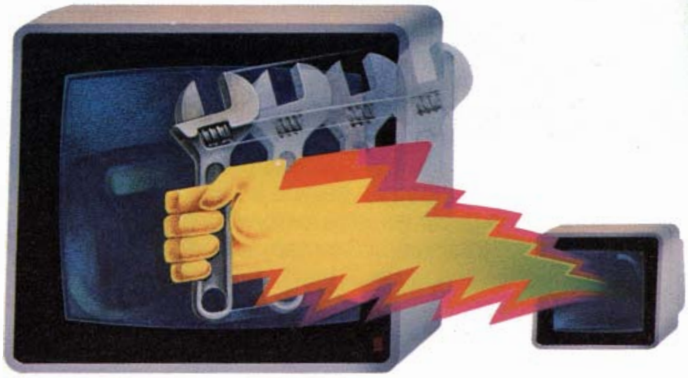
Light and very strong, composites—fibers of graphite embedded in epoxy—are gaining increasing use in aircraft and spacecraft. But making composites usually involves costly "cooking" in a huge oven called an autoclave. However, Lockheed now eliminates that process. It winds graphite fibers around a tube that tapers from 4 to 2 inches over its 8-foot length. Then the tube is heated—this avoids autoclaving—and epoxy is injected around the fibers. Result: an 8-foot structure weighing about 1½ pounds, ⅓ the weight of an aluminum structure of equal strength.

These tapered structures can be packaged inside one another like paper cups to save precious space when transported in the Space Shuttle. A special automated assembly machine designed by Lockheed under contract to the NASA Langley Research Center will enable astronauts to assemble them in space without tools.

Aspiration? Far from it. Lockheed has tested, proven and manufactured dozens of these remarkable structures.

## Adding a work force by wire.

To increase productivity—virtually a nationwide goal—you simply make more efficient use of existing staffs and facilities. So logical, yet so difficult. However, Lockheed has invented a remarkable new computerized system that does just that.



It begins with another unique Lockheed System—CADAM. The computer-based CADAM basically gives engineers a three-dimensional drawing board as big as 48 football fields on which to design parts. CADAM's computer then electronically transmits the design to the plant's manufacturing experts.

Now Lockheed has linked CADAM computers at divisions across the country by a telecommunications network—COMCAM.

If one plant is overloaded, it calls on COMCAM and transmits a CADAM design to another division that can free staff and tools to make the part. This eliminates the cumbersome task of physically moving actual drawings, specifications and tapes from plant to plant. Changes in the design can be made at one plant and instantly transmitted to

# make better products.

*Astronauts could assemble buildings in space without using any tools, thanks to a Lockheed process described below.*

another. Put simply, COMCAM makes better use of people, machines and time—increased productivity.

## Calling up the best brains in manufacturing.

The CADAM system, which was invented at Lockheed, enables engineers to design aircraft parts on a computer terminal with greater precision and imagination than ever before.

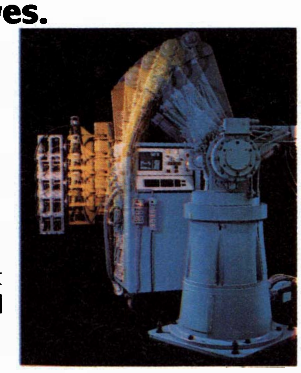
To help its manufacturing experts translate those designs into actual airframes, Lockheed has given them an even more remarkable computer-based system—GENPLAN.

Exclusive to Lockheed, GENPLAN actually makes decisions such as which tools and manufacturing processes should be used. It determines the best manufacturing sequence. It considers factors such as labor standards and factory rules. Within GENPLAN's huge data base, Lockheed has captured the accumulated knowledge of thousands of manufacturing experts. Moreover, Lockheed divisions thousands of miles apart can share this knowledge.

Working at a computer terminal, the manufacturing planner can create, review and edit the manufacturing plan.

The time savings of GENPLAN are immense. Plans that took hours now are made in minutes. The steps required to release the plan to the assembly line have been reduced by 75%. Another key benefit flows from GENPLAN. The increase in precision and reliability of manufactured parts is as dramatic as the time savings.

## Robots roll up their sleeves.



The extended "arm" that you see in motion is a robot being "trained" to place rivets with great speed and precision. Once programmed—and programming is critical in the effective use of a robot—it may cut riveting costs 50%. Other robots are at work at Lockheed, including machines that handle onerous painting duties and that can "see" and choose among different parts.

*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**


Thank Dad for believing you were very special every step of the way.
Johnnie Walker®
Black Label Scotch
YEARS 12 OLD
12 YEAR OLD BLENDED SCOTCH WHISKY, 86.8 PROOF. BOTTLED IN SCOTLAND. IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y. © 1982

Established 1845 SCIENTIFIC AMERICAN June 1982 Volume 246 Number 6

# Phosphate Rock

*This essential raw material for fertilizer is widely distributed but unevenly mined and consumed. Now its geologic origins are becoming clearer, and more deposits are likely to be discovered*

by Richard P. Sheldon

The fruitfulness of modern agriculture arises in large part from heavy applications of fertilizer. The supply of fertilizer depends in large part on a raw material one seldom hears about: phosphate rock. Yet in 1981 the amount of phosphate rock mined throughout the world was some 143 million tons. For purposes of comparison, that was somewhat more than the world's production of salt and somewhat less than the world's production of lime in the same year; setting phosphate at 100, salt was 80 and lime 106. The U.S. mines about 40 percent of the world's supply of phosphate rock per year. In 1981 the U.S. output was valued at $1,539 million.

Phosphate (mainly the ion $PO_4^-$) is an essential nutrient of both plants and animals. Phosphorus is a major constituent of key molecules in all the processes of life, for example the genetic material DNA and the energy-rich compound ATP. Together with nitrogen and potassium, phosphate is a major constituent of all modern chemical fertilizers. It is present naturally in water and soil in sufficient amounts to sustain a variety of biological communities. In the thousands of years since the advent of agriculture, however, the soils of cultivated lands have been more or less depleted of their natural phosphate, with the result that phosphate fertilizer must be applied in order to sustain and increase their fertility.

The consumption of phosphate is increasing at a rate of about 6.3 percent per year, which raises the question of whether the supply is likely to run out. The answer is no. The fact that the developed countries consume far more phosphate than the developing ones raises the question of whether anything can be done to achieve a better balance. The answer is yes, provided the potential phosphate deposits of developing countries are discovered and enough attention is directed to the appropriate technology, that is, to processes that will enable the developing countries to increase their supply of phosphate without inordinate capital investment.

The main source of the phosphate in fertilizer is phosphate rock of marine origin. A smaller source is phosphate rock of igneous origin. At one time guano (seabird droppings) and rock derived from it were significant sources, but they are now nearly exhausted.

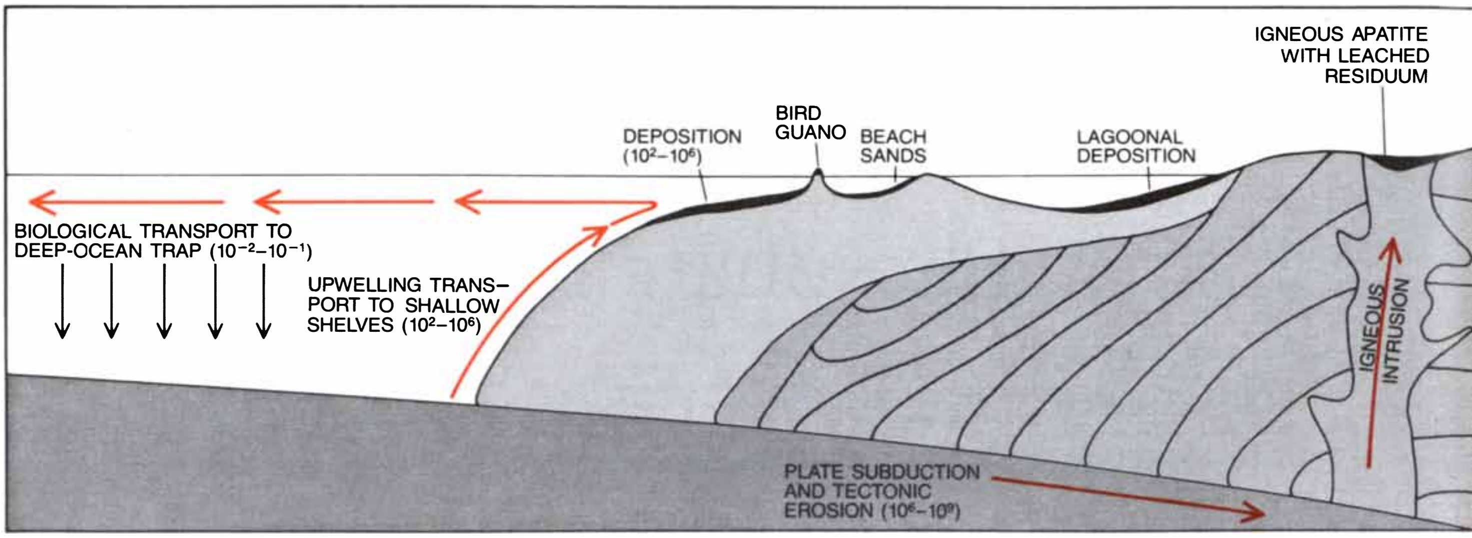
Phosphorus moves as phosphate in a grand geochemical cycle. Phosphate in minerals is released into the biosphere by chemical weathering. The phosphate in the biosphere is continuously utilized by plants and animals on its way to the ocean. In the ocean much of it is withdrawn from the biological part of the cycle as the oceanic plants and animals that take up phosphate die and sink into deeper water. There the organic matter is broken down by oxidation or bacterial action and the phosphate is liberated back into the water.

The ocean is stratified according to the density of its waters, with the denser, colder and more saline waters at the bottom. This water becomes enriched in phosphate and other nutrients. The deep-ocean water circulates slowly, the time required for a complete turnover of the ocean being tens of thousands of years. Small amounts of the phosphate dissolved in the deep ocean are continuously incorporated into marine sediments such as red deep-sea muds, carbonate rocks and organic matter, but this material does not form significant amounts of phosphate rock. At certain times in the earth's history, however, large amounts of phosphate were deposited as phosphate rock by processes that were long a mystery.

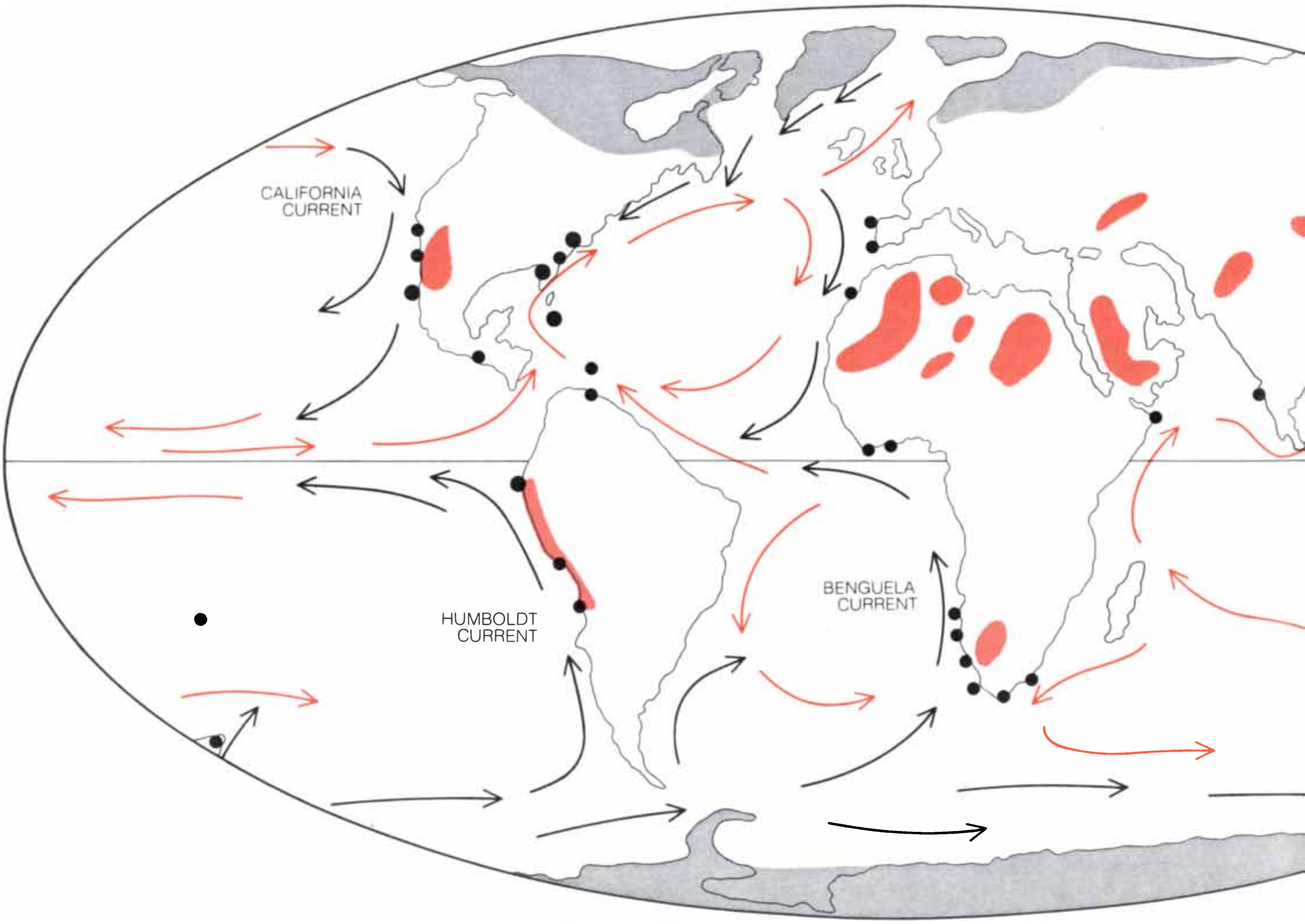
In 1937 A. V. Kazakov of the U.S.S.R. advanced the hypothesis that marine phosphate rock is deposited in areas of upwelling ocean currents. Such currents are generated by the wind-driven gyral circulation of the upper few hundred meters of water along the edges of the oceans. In the Northern Hemisphere the movement is clockwise, in the Southern Hemisphere counterclockwise. The gyres extend from the Equator to a latitude of about 50 degrees.

Where the eastern boundary current sweeps the west coasts of North America, South America and Africa it tends to move offshore to the west, in part blown by the strong trade winds and in part diverted by the force of the rotation of the earth (the Coriolis force). The water thus transported is replaced by colder water that wells up along the shore. These cold currents are so distinctive that they have names: the Benguela Current off southwestern Africa, the Humboldt Current off western South America and the California Current off western North America.

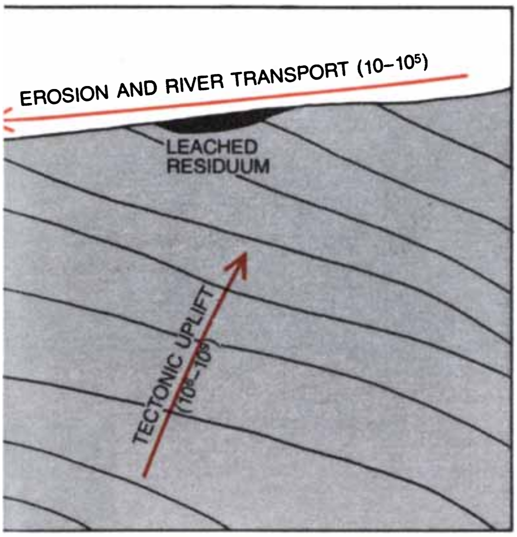
The upwelling of cold, nutrient-rich water to the surface along shallow continental shelves in the sunny low latitudes of the trade winds sets up a unique environment. The highly fertile waters of a tropical climate nourish huge quantities of plant life (primarily diatoms, which are floating one-cell plants) that feed a chain of life extending from microscopic animals to small crustaceans to fishes to birds to man. Organic debris in these waters (plant and animal remains and



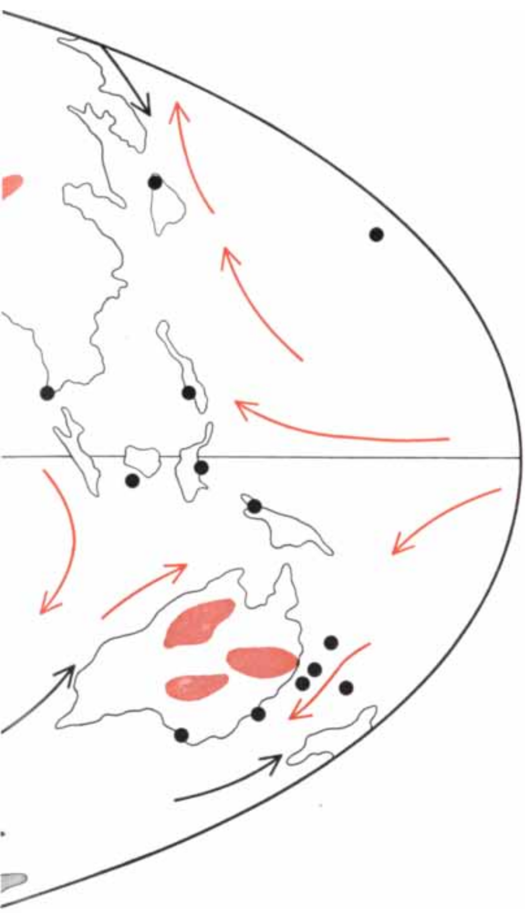

**GEOCHEMICAL CYCLE** of phosphate, which is an oxygen anion of the element phosphorus, rotates slowly from the land to the sea and back to the land. On the land chemical weathering releases the phosphate from minerals. The phosphate is carried gradually to the oceans, being utilized by plants and animals on the way. In the ocean it is taken up by marine plants and animals. When they die and sink to the bottom, the breakdown of the organic matter leaves phosphate pellets in the sediment. Upwelling currents deposit phosphate on the




**PHOSPHATE WAS DEPOSITED** on a large scale in five geologic periods by the mechanism depicted here for the most recent one, the Late Tertiary to Recent period of some 10 million years ago. The positions of the continents are shown as they are thought to have been at the time. The key mechanism was the wind-driven gyral circulation of the top few hundred meters of the ocean. The gyres, moving clockwise in the Northern Hemisphere and counterclockwise in the Southern, included warm currents (*colored arrows*) and cold ones (*black arrows*). Where the eastern boundary current swept the west coast of North America, South America and Africa it was blown offshore to the west by the trade winds and was replaced by upwelling cold currents such as the three named. They were rich in phosphate and sup-




**continental shelves. Some is returned to the land by sea-floor spreading and plate tectonics. The numbers show (in years) the estimated residence time of phosphate in the cycle.**



**ported an abundance of plants and animals that sank onto the shelves when they died. Much of their phosphate remained there, creating major deposits (*dots*). The trade-wind processes, related to polar glaciation, also built large sand deposits on land (*dark color*).**

fecal pellets) sinks to the bottom. Where it is not swept away by bottom currents it decays through bacterial action, releasing phosphate into bottom water that is already rich in it. G. Baturin of the U.S.S.R., by examining the geochemistry of modern sediments on the continental shelf of southwest Africa, has shown that they contain scattered buckshot-size pellets composed of the mineral apatite: a calcium phosphate plus fluorine and carbonate. Fishbones and shark teeth, also consisting of apatite, are common in these sediments, as are tiny planktonic animal shells of calcium carbonate that has been replaced by apatite.

A likely hypothesis, although it has not been conclusively demonstrated, is that concentrations of apatite pellets form on the ocean floor and make up widespread beds of phosphate rock. The other sediments associated with phosphate rock on the ocean floor are diatomite and mud rich in organic material. The resulting rock assemblage consists of phosphate rock, chert (a rock composed of silica from organisms such as diatoms and siliceous sponges) and black shale.

Huge quantities of such assemblages were deposited on continental shelves at low latitudes by the upwelling of eastern boundary currents five times over the past 560 million years (in the Ordovician, the Mississippian, the Permian, the Jurassic and the Late Tertiary to Recent periods). The ancient climates and ocean currents were much the same in those periods as they are now, with polar ice caps, tropical low-latitude seas, a low sea level and strong earth-circling winds. Phosphate rock of this origin is known as trade-wind-belt phosphate. It is found on dry land because it has been placed there by the processes of sea-floor spreading and plate tectonics.

A second family of economically important phosphate-rock deposits is called the equatorial group. Such deposits are made up of the same rock assemblage as the trade-wind deposits, but the similarity stops there. Whereas the trade-wind deposits are found in the trade-wind latitudes of eastern boundary currents and tend to run north and south, the equatorial deposits appear in low latitudes and tend to run east and west. The equatorial assemblages were deposited at times when (1) the sea level was unusually high, flooding large continental areas with shallow seas; (2) the polar regions were open to the ocean and warm, and (3) strong upwelling currents were flowing in equatorial regions rather than in the trade-wind belts. It seems likely that at such times the oceanic gyres were flowing weakly or perhaps not at all.

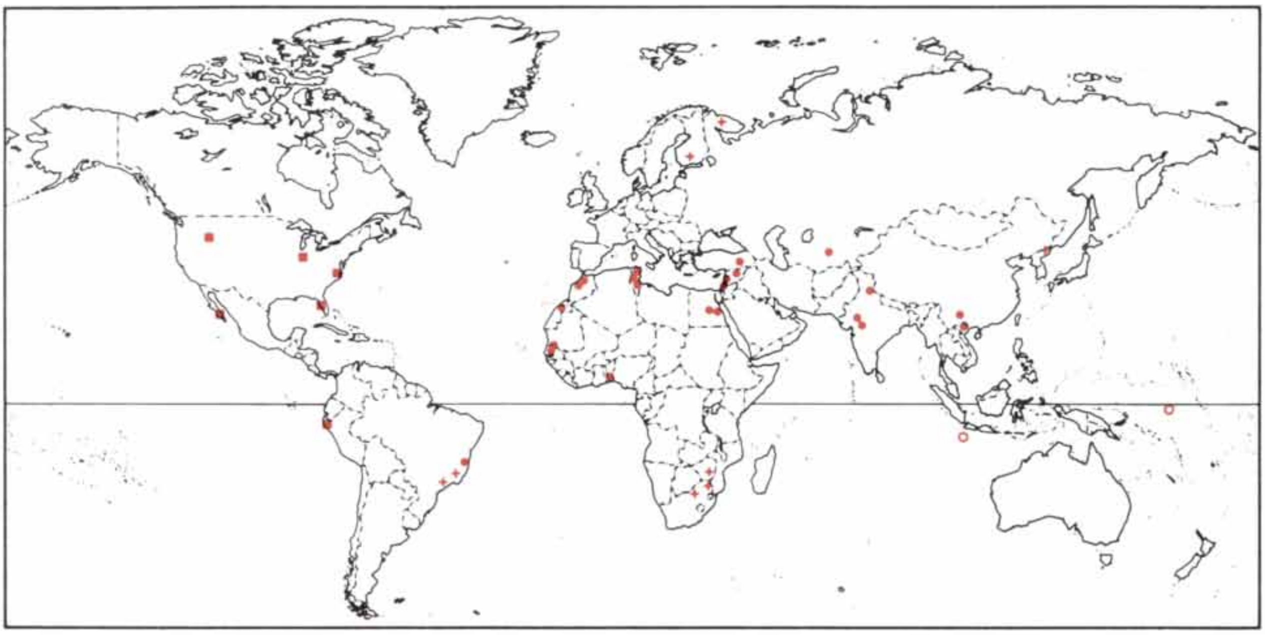
The trade-wind and equatorial deposits dominate the current world production of phosphate rock. Major trade-wind deposits are in southern Baja California in Mexico, the Sechura Desert of Peru, the Caribbean Sea, the southeastern U.S. and the northwestern U.S. Major equatorial deposits are in Morocco, the western Sahara and elsewhere in North Africa, in the Middle East, Venezuela and Colombia; together they make up the largest phosphate province in the world. It seems likely that more deposits of both types will be found, in particular because their origin is becoming better understood.

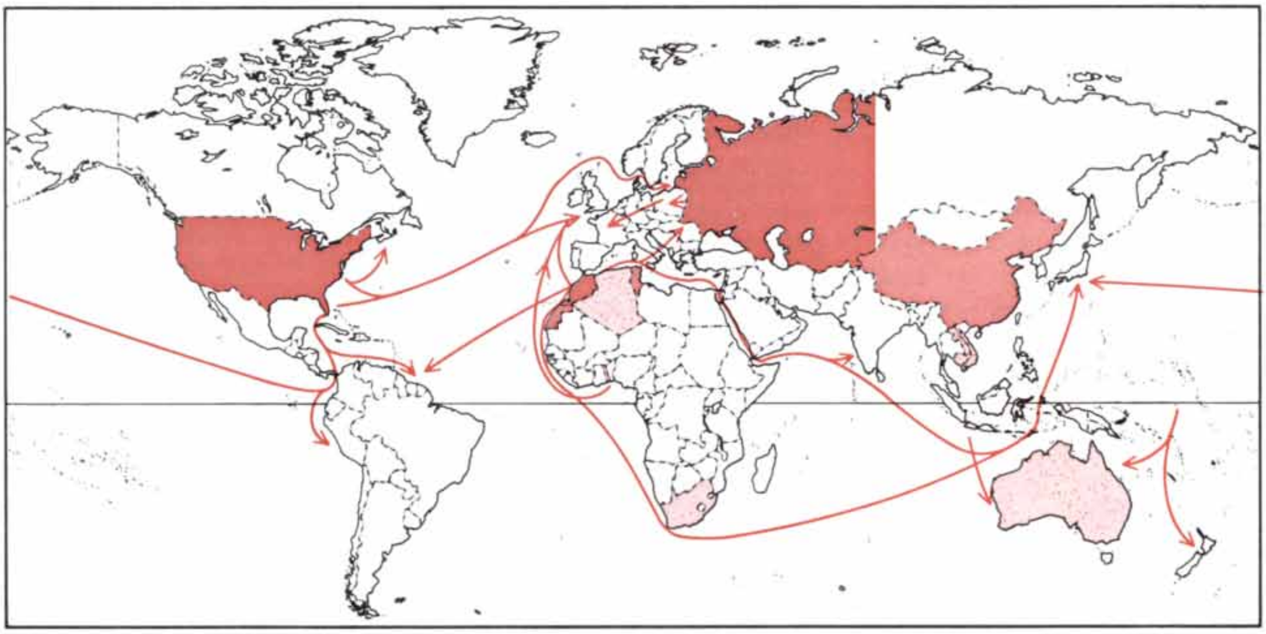
The igneous deposits consist of apatite, most of which is calcium hydroxyfluorophosphate: $Ca_5(PO_4)_3(OH,F)$. They are derived from rocks that melted in the earth's crust, flowed upward and solidified as intrusions into other formations. Some small, circular intrusions (ring dikes) hold concentrations of apatite as high as 60 percent, but most deposits are not as rich. Major sources of igneous phosphate rock are the Palabora deposit in South Africa and the Khibiny deposit of the Kola Peninsula in the U.S.S.R. Igneous apatite deposits now contribute somewhat more than 15 percent of the world's production of phosphate; production from this source is expanding at about the same rate as production from marine sedimentary phosphate rock. The geologic origin of igneous deposits is relatively obscure, so that the theoretical basis on which exploration for new igneous sources can be made is limited. The circular form of typical igneous deposits can often be distinguished in images made from satellites, however, and so the rate of discovery has been increasing.

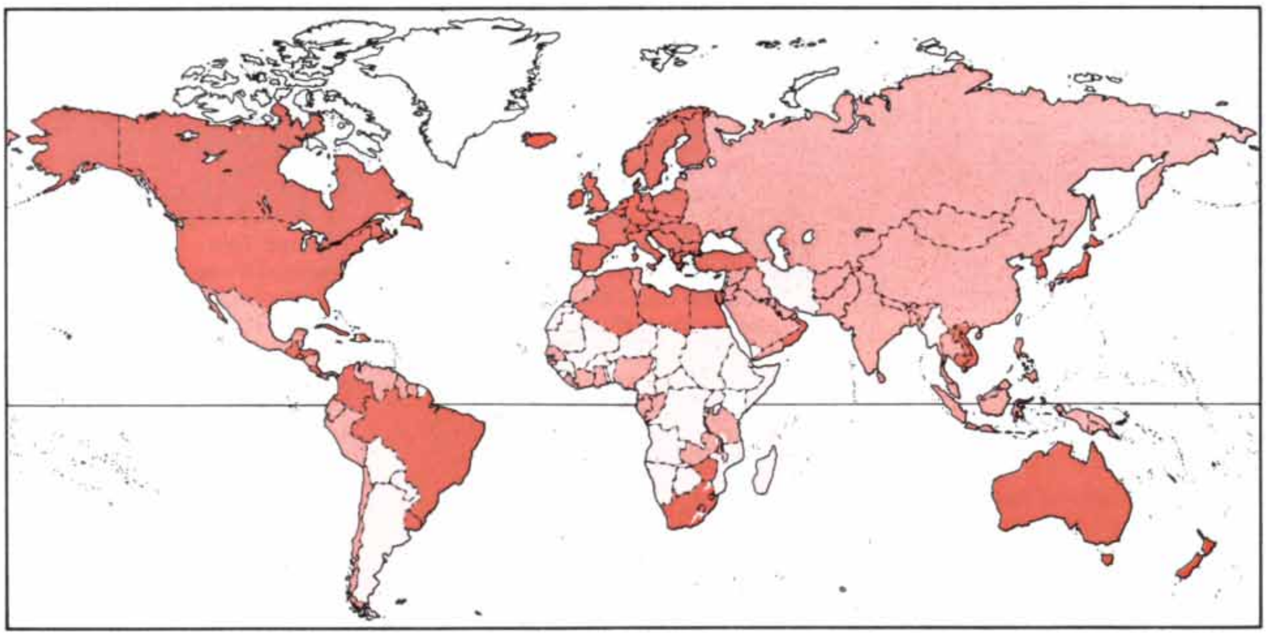
It should be noted that many other deposits of phosphate rock of different origin are known. They are smaller and of lower grade than the three major kinds I have described. So far they cannot be mined economically, and they have therefore not been well studied. Still, their quantity is impressive. I have seen a map of Asia on which all known phosphate deposits were plotted with red dots; it looked as though the continent had measles. It is likely that similar plots for other regions would yield similar results.

The economic phosphate deposits of the world appear to have been distributed by a vengeful god. The most populated regions, which have the greatest need for phosphate, have the least of it. China, India, Indonesia, Brazil, Bangladesh, Pakistan, Nigeria, Vietnam, the Philippines, Thailand and Turkey, the 11 most populous developing countries, have 56 percent of the world's population but produce 7.6 percent of the phosphate rock and import only an additional 3 percent.

On the other hand, the developed countries, together with China and a few sparsely populated developing countries, have most of the phosphate depos-







EQUATORIAL PHOSPHATE ROCKS
TRADE-WIND-BELT PHOSPHATE ROCKS
IGNEOUS PHOSPHATE ROCKS
INSULAR PHOSPHATE ROCKS

**MAJOR ECONOMIC SOURCES** of phosphate are mostly marine deposits that have been returned to the land by geologic processes. Igneous phosphate is derived from rocks that melted in the earth's crust, flowed upward and solidified as intrusions into other formations. Equatorial deposits were created when the trade-wind gyres were weak and upwelling currents flowed near the Equator.

PRODUCTION
(MILLIONS OF METRIC TONS)
1–3 4–10 20–60

**WORLD TRADE** in phosphate rock flows mainly from the major producing countries to the major hard-currency markets. The developing countries, many of which need large amounts of phosphate fertilizer, cannot afford to import the phosphate rock that would enable them to make it. Most of the phosphate-rock exports of the U.S.S.R. go to eastern Europe and do not figure in the world market.

CONSUMPTION OF PHOSPHATE ROCK
(100 GRAMS PER HECTARE)
1–10 100–1,000
10–100 1,000–10,000

**PATTERN OF CONSUMPTION** of phosphate rock as phosphate fertilizer varies widely. The figures in the key represent units of 100 grams of phosphorus pentoxide ($P_2O_5$) per hectare (2.47 acres) of arable land per year. The poorest countries use the least fertilizer, which is one of the main reasons for the low productivity of their agriculture and their inability to feed their population adequately.

its and produce an overwhelming proportion of the phosphate rock. The three leading producers—the U.S., the U.S.S.R. and Morocco-Sahara—mine 73 percent of the world's phosphate rock. With the output of China, Tunisia, Jordan and South Africa added the total reaches 86 percent.

The picture for the developing countries brightens somewhat when one looks at the reserve base of phosphate rock. It consists of known deposits that could be mined in more or less the present economic circumstances. Several developing countries are said to have reserve deposits of significant size. In most cases, however, the figures are highly unreliable. Even so, deposits that are not now being mined await development in many countries of the Third World.

In considering how the data on production and the reserve base relate to the future supply of phosphate one must take account of how fertilizer is made and of the world trade in phosphate. Farming in the developed countries is highly mechanized, and the machines made to spread fertilizer call for a granulated product that does not become sticky in contact with moist air. As a result the manufacturers ordinarily turn out chemically complex mixed fertilizers such as diammonium phosphate. Moreover, the fertilizers tend to be highly concentrated in order to cut down the amount of material that must be handled in the delivery system and by the farmer.

These practices impose rigorous standards of quality on the raw materials utilized in the manufacture of fertilizer. For example, small amounts of iron oxide, alumina and magnesium in the phosphate rock cause the formation of gels that clog the filters and pipes of the fertilizer manufacturing plant. Phosphate rock with excessive amounts of those elements cannot be utilized, and such deposits cannot be mined economically.

Another constraint is that the main processes for manufacturing phosphate fertilizer call for the dissolution of the phosphate rock by sulfuric acid, which is an energy-expensive commodity. The treatment of phosphate rock containing excessive amounts of calcium or magnesium carbonate wastes sulfuric acid; therefore such deposits cannot be mined economically. A third constraint is that much phosphate rock is shipped by sea, and so the grade of rock must be kept high to hold down the shipping cost per unit of phosphate.

The effect of such constraints is to place economic limits on phosphate rock. This is the prime reason low-quality deposits in many countries are not being developed. Efforts are being made to modify the fertilizer manufacturing processes so that phosphate deposits of lower grade can come into play, but the fact that the manufacturing plants now in existence were designed for high-quality rock makes change slow.

The world trade in phosphate rock is dominated by the deposits of the southeastern U.S. and Morocco, which together account for 56 percent of the export trade. The U.S.S.R. exports 14 percent of the world flow, but its shipments go primarily to the countries of eastern Europe and therefore are outside most of the world market.

The flow of phosphate rock is hugely distorted. One side of the globe is supplying much of the phosphate rock to the other. Because shipping costs have risen greatly since 1974 the price of phosphate rock is about 50 percent higher on the importing side of the globe than it is in the places of origin, and the importing side is precisely where the countries tend to be the poorest and the most populous. As one might suppose, the higher cost has raised a major barrier to the importation of phosphate rock for the manufacture of fertilizer in those countries. The result is that the world trade is primarily structured to sell phosphate rock in the large hard-currency markets. An important exception to the pattern is found in the centrally planned Asian countries, which import and export little phosphate but have vigorously developed their deposits to supply their own needs.

The world consumption of phosphate rock per hectare (2.47 acres) of arable land is to some extent correlated with the per capita gross national product of the countries. It should also be recognized that the major exporters of grain (the U.S., Canada and Australia) are in effect exporting a significant fraction of their phosphate consumption. By the same token the developing countries that have major exports of such crops as sugarcane, coffee, cocoa and cotton are indirectly exporting or reexporting the phosphate fertilizer they apply to those crops.

The need for phosphate fertilizer and the economic demand for phosphate fertilizer are distinctly different things. Y. H. Yang of the Resource Systems Institute of the East-West Center in Honolulu has published an analysis of the need for fertilizer from the viewpoint of the amount required to grow enough food to meet the nutritional needs of the people of Asia. He has calculated that more than 10 times the quantity of fertilizer applied in 1975 would have to be laid down in Asia annually by the year 2000. The corresponding rates of application would be far beyond the optimum levels for cereal crops, which is one measure of how much the population exceeds the carrying capacity of the land.

Robert L. Fox, Goro Uehera and R. S. Yost of the University of Hawaii's College of Tropical Agriculture have jointly studied the phosphate needs of tropical soils. Uehera noted that in such soils the clay minerals on surfaces long exposed to chemical weathering processes are partly broken down, so that they adsorb phosphate on the degraded faces of the clay crystals. The phosphate is then unavailable to plants as a nutrient.

This adsorptive capacity of tropical soils must be satisfied before normal applications of fertilizer can be effective. Fox and Yost calculated that 540 million tons of phosphorus pentoxide ($P_2O_5$), the equivalent of about 1.7 billion tons of high-grade phosphate rock, would have to be added to the currently cultivated soils of the world, mostly in the tropical regions, to bring them up to a moderate level of fertility. Thereafter the soils could be maintained with annual applications of about 20 million tons, which is about half the 1980 consumption. This one-time-only investment in phosphate fertilizer would go a long way toward meeting the phosphate demand of the overpopulated tropical regions. Although the cost would be several tens of billions of dollars, it would not be burdensome to the world economy if it were spread out over several decades.

The major part of the unsatisfied need for phosphate is on subsistence-level farms in developing countries. The typical subsistence farm occupies less than one hectare and is managed by a single family. It produces only enough to keep the family fed in good years. The family usually has no cash crop, and so the farm does not generate the money needed to buy the fertilizer that would be needed to increase productivity.

It seems clear that the continuation of the present system for distributing phosphate fertilizer will not lower its cost to the point where subsistence farmers will be able to buy it. Indeed, the depletion of phosphate-rock deposits of high grade and the increase in energy costs have made fertilizer more expensive. Another problem is inducing the farmers to improve their conservative agricultural technology in accordance with the changed circumstances that larger inputs of fertilizer would bring about.

Since it is unlikely that the fertilizer supply and marketing system developed for large-scale commercial farming can be expected to also furnish fertilizer to subsistence-level farmers, it seems appropriate to consider alternative technology. Workers at the Kettering Research Laboratories in Yellow Springs, Ohio, have been doing just that. R. William Treharne and his colleagues have been studying the old electric-arc process of making nitric oxide (NO) to produce nitrate fertilizer. In this process air is passed through an electric arc, and the high temperature of the arc causes the nitrogen and the oxygen in the air to combine. The process was in use until about 1900, when it began to be replaced by the more efficient Haber-Bosch process, which utilizes the reaction of hydrogen and nitrogen to produce ammonia ($NH_3$).

Treharne and his associates reasoned that if the arc apparatus were made small enough to operate on a farm or in a farm village and had a modest source of energy, nitrogen fertilizer could be made cheaply by reacting the nitric acid with limestone. If some kind of phosphate rock were available, the reaction with nitric acid would produce phosphate fertilizer too. Such a system could make the small and low-grade deposits of phosphate rock economic to mine. Indeed, the electric-arc process is not impeded by impure phosphate rock. Ferrous iron and magnesium, the elements that interfere with the standard technology for making phosphate fertilizer, do not affect the arc process and are beneficial to the final product.

Experimental stations making fertilizer in this way are operating at the universities of Nebraska, Nevada and Hawaii, powered by small generators running respectively on solar, water and wind systems. All of them are successfully making nitrogen fertilizer, and the one in Hawaii is making both nitrogen fertilizer and phosphate fertilizer. The processes are still under development, however, and an operational system has yet to prove out.

Calcium phosphate is a common compound that can be found in recoverable form in many kinds of rock. When the high-grade phosphate rock that is now being mined runs out, huge quantities from which phosphate can be extracted with somewhat more difficulty will remain. The supply of phosphate therefore seems assured for a long time. The increased difficulty of extracting it, however, will also mean that the price of phosphate fertilizer will rise. The extent of the price increase will depend on the nature of the advances made in the technology of mining and extracting phosphate and in the manufacture of phosphate fertilizer.

The geologic distribution of phosphate rock is beginning to be understood, with the result that a scientific base can be established for a worldwide assessment of major phosphate resources and a search for undiscovered ones. Since in general the underdeveloped countries also have underdeveloped resources, it can be expected that proportionately more of the discoveries of major phosphate-rock deposits will be in such areas.

It is highly unlikely that every country will have an economic phosphate deposit of substantial size, but most regions either have one or will probably discover one. On the other hand, almost every country should have many small deposits of the kind that cannot now be mined economically. It is because of human ignorance and neglect that many countries where not enough food is available per capita find phosphate too expensive on the world market and their own deposits too uneconomic to develop.

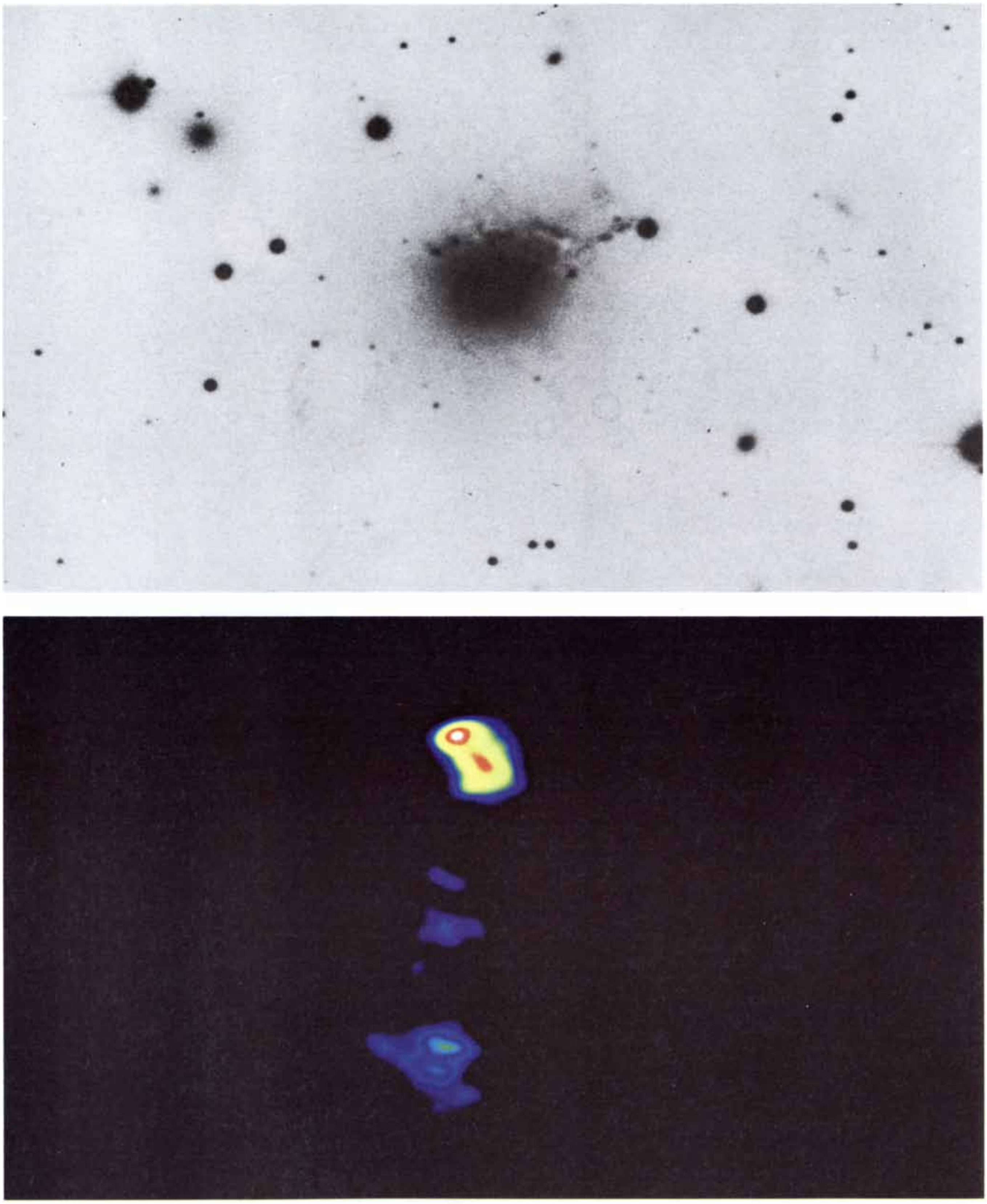
If the technology for processing low-grade phosphate rock were to be invented and developed, and if developing nations assessed their phosphate resources and discovered new ones, enough phosphate fertilizer would be available to greatly raise the fertility of tropical soils, alleviate hunger and put the agriculture of developing nations on a firmer footing. Unfortunately the development banks and agencies seem unwilling to gamble on investment in the invention and development of technology to produce phosphate from undiscovered or uneconomic deposits of phosphate rock. Their attitude is understandable, but one should remember that uncertain prospects have faced every civilization at every stage of its development. One might better ask: How can humankind afford not to take such a gamble?

| COUNTRY | 1980 PRODUCTION | RESERVE BASE |
|---|---|---|
| U.S. | 54.4 | 8,000 |
| U.S.S.R. | 26.1 | 6,300 |
| MOROCCO-SAHARA | 18.8 | 44,000 |
| CHINA | 6.7 | 10,000 |
| TUNISIA | 4.6 | 500 |
| JORDAN | 4.2 | 1,100 |
| SOUTH AFRICA | 3.3 | 700 |
| TOGO | 2.9 | 110 |
| BRAZIL | 2.9 | 800 |
| ISRAEL | 2.6 | 150 |
| NAURU | 2.1 | 20 |
| CHRISTMAS ISLAND | 1.7 | 62 |
| SENEGAL | 1.4 | 75 |
| SYRIA | 1.3 | 833 |
| ALGERIA | 1 | 1,000 |
| NORTH KOREA | .5 | 90 |
| INDIA | .4 | 108 |
| VIETNAM | .4 | 100 |
| MEXICO | .3 | 1,034 |
| ZIMBABWE | .1 | 50 |
| FINLAND | .1 | 565 |

**LEADING PRODUCERS of phosphate rock are listed in millions of metric tons. The reserve base is phosphate rock that could be mined under existing economic conditions.**

**PHOSPHATE IS MINED** at one of the large deposits of phosphate rock of marine origin near Bartow in central Florida. The dragline mines ore, which includes sand and clay as well as phosphate. The ore goes to a beneficiation plant where the phosphate is separated.

**FERTILIZER PLANT** of W. R. Grace & Co. near Bartow is where beneficiated phosphate rock is made into granulated fertilizer. The six silos at the right store the finished product; the material piled behind them is the plant's inventory of wet phosphate rock as it is delivered from the beneficiation plant. The three lagoons in the rear hold water that is cycled through the plant repeatedly for cooling.



**DISTURBED ELLIPTICAL GALAXY NGC 1275,** which dominates the rich Perseus cluster of galaxies, is a powerful source of radio waves at a distance of 300 million light-years. At the top of the page it appears as the fuzzy central object in a photograph made with the 200-inch telescope on Palomar Mountain. The diameter of the fuzzy spot is 80 seconds of arc. At the bottom the same object appears in a radio image made by very-long-baseline interferometry (VLBI). The resolution of the radio image is .0004 arc-second, revealing details 2,500 times smaller than any that can be observed optically. From edge to edge the radio image covers only .023 arc-second; on the same scale the entire galaxy in the optical image would have a diameter of 175 meters. The radio image required the correlation of simultaneous observations recorded at five locations: Owens Valley in California, Socorro in New Mexico, Algonquin Park in Ontario, Green Bank in West Virginia and Effelsberg in West Germany. The principal investigator was the author; his co-workers were Martin S. Ewing and David Hough of the California Institute of Technology, R. Craig Walker of the National Radio Astronomy Observatory and Jonathan D. Romney of the Max Planck Institute for Radio Astronomy. The smallest structure in the radio image, generated on a television screen, is one light-year long and .2 light-year across. It is thought to be associated with the central engine of the radio emission.

# Radio Astronomy by Very-Long-Baseline Interferometry

*Observations made simultaneously by radio telescopes thousands of miles apart can be combined with the aid of atomic clocks. The resolution of the observations is the highest ever achieved*

by Anthony C. S. Readhead

The most powerful objects in the universe are radio galaxies and quasars. They are now being observed in greatest detail with networks of radio telescopes extending across continents and even across oceans. The technology that makes these high-resolution observations possible is very-long-baseline interferometry (VLBI).

The remarkable images produced by very-long-baseline interferometry are governed by the same laws of physics that apply to light-gathering telescopes. In order to improve the resolving power of an imaging system the aperture over which radiation is collected must be increased in relation to the wavelength of the radiation. It follows that in principle it is much harder to discriminate between closely spaced points with a radio telescope, which works with waves ranging in length from centimeters to meters, than it is to resolve equally closely spaced points with an optical telescope, which works with waves whose length is measured in millionths of a centimeter. Resolving power, or resolution, is the minimum angular separation, measured in minutes or seconds of arc, that can be detected by an observing instrument. For a telescope the resolution is given approximately by the ratio $\lambda/D$, where $\lambda$ is the wavelength of the radiation and $D$ is the aperture of the telescope. In order to record fine details in astronomical objects $D$ should be as large as possible for the particular wavelength at which the observation is to be made.

At the end of World War II some of the young scientists who had developed radar began to apply their new skills to the nearly virgin field of radio astronomy. At that time the possibility of achieving resolving power with radio telescopes that could match the resolution of optical telescopes must have seemed an impossible dream. The resolving power of large optical telescopes under good seeing conditions is about one arc-second, roughly the angle subtended by a small coin at a distance of four kilometers or by two kilometers at the distance of the moon.

It was easy to calculate that if an equivalent resolution were to be obtained with a radio telescope operating at a wavelength of, say, one meter, the diameter of the collecting surface would have to be some 150 miles. By the same token a radio telescope with the then unthinkable diameter of 1,000 feet (subsequently the size of the largest radio dish ever built) would have a resolution no better than 800 arc-seconds, or roughly a fourth of the resolving power of the human eye when it is looking at stars in the night sky.

The skillful application of techniques of radio interferometry, coupled with the evolution of computer technology, has led to resolving powers that were inconceivable in the early days of radio astronomy. Within the past few years radio astronomers have obtained images of some of the most distant objects in the universe with resolutions approaching .0001 arc-second, about equal to the span of a human hand at the distance of the moon. This constitutes an improvement of some four orders of magnitude over the resolution of earth-based optical telescopes.

The qualification "earth-based" is necessary, because it is the turbulence in the earth's atmosphere, the phenomenon that makes stars twinkle, that degrades the resolution theoretically attainable with large parabolic mirrors. In principle the 200-inch telescope on Palomar Mountain should have a resolution of .023 arc-second. In practice its resolution is only about 10 times better than that of Galileo's best telescopes: about 10 arc-seconds. His instruments, in turn, represented a dramatic twentyfold improvement over the resolution of the unaided eye looking at the night sky. What today's great optical telescopes do brilliantly, of course, is collect large amounts of radiant energy in a small amount of time. As a "light bucket" the 200-inch Palomar telescope outperforms Galileo's telescope by a factor of about 40,000. When the Space Telescope, nearly half the diameter of the Palomar reflector, is placed in orbit in 1985 above the image-degrading atmosphere, its resolution should approach the theoretical value of .05 arc-second for an instrument of its size.

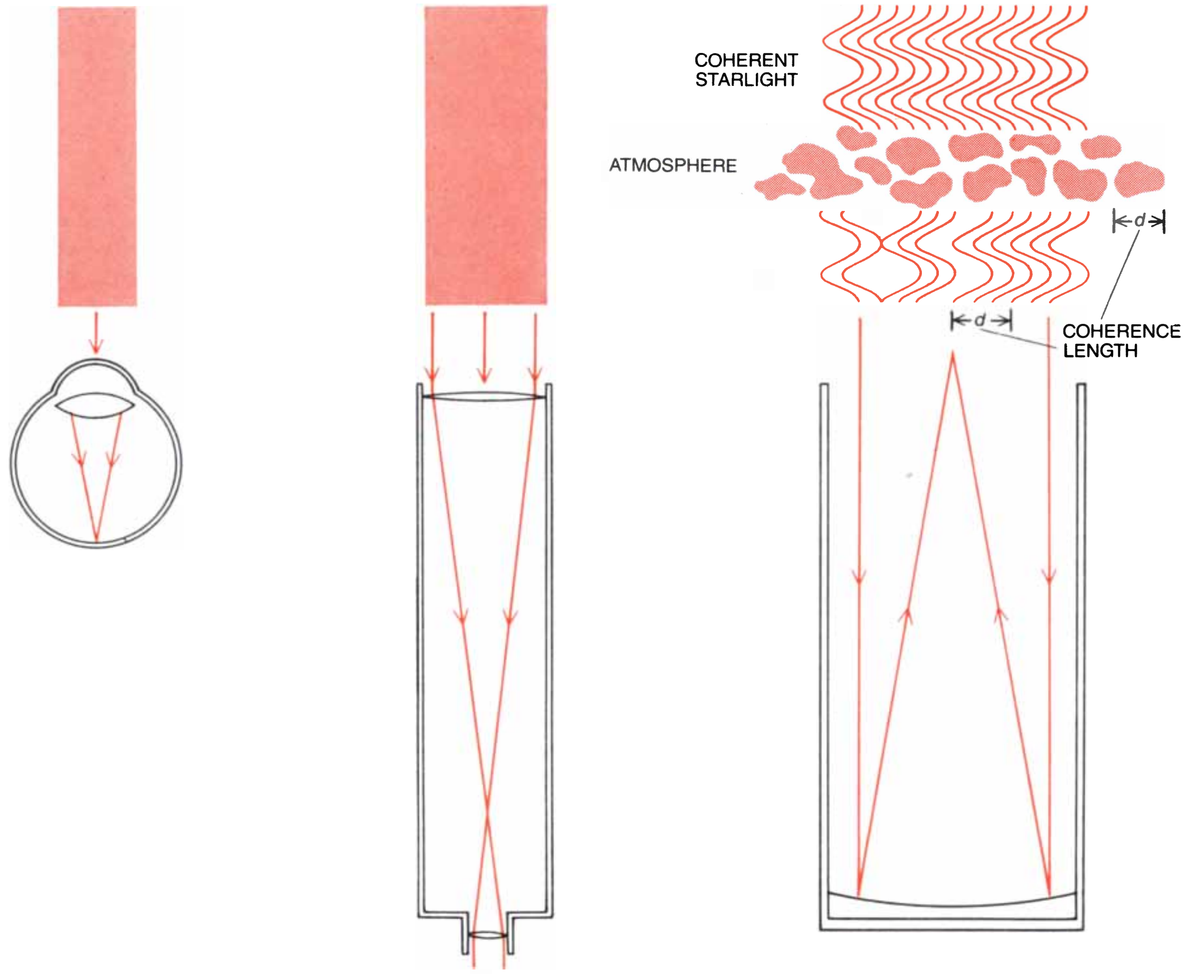
The resolution of .0001 arc-second now being realized in radio astronomy does not violate the relation $\lambda/D$. The requirement of making the aperture $D$ very large is accomplished by "aperture synthesis," a technique in which the radio waves collected by two (or more) instruments hundreds or thousands of kilometers apart are recorded simultaneously and subsequently added together. The addition must be done, of course, in a special way; it is necessary to arrange the observations precisely so that the phase relations of radio waves arriving at the telescopes participating in the synthesis are preserved. The waves are then said to be coherent. The summed intensity of the radiation will be high or low depending on whether the arriving radio waves are in phase or out of phase at the two telescopes. The waves are in phase if the wave crests reach the two telescopes simultaneously; they are out of phase if a wave trough reaches one telescope at the same instant that a wave crest reaches the other. Such a combination of instruments is called an interferometer.

Since the earth rotates, the relative position of two radio telescopes is continuously changing. As a result the signal intensity at the output of the interferometer passes rapidly through a succession of maximums and minimums as the radio waves from the celestial object being observed are successively in phase at the two telescopes and then out of phase. These intensity maximums and minimums are called interference fringes.

The actual shape of the intensity variations is a sine wave. The amplitude (size) and phase (position) of the sine wave embody information about the structure of the celestial object on a certain angular scale, which depends on the distance between the two telescopes. If the distance between the telescopes is changed, one gets a different pattern, which then contains information about the structure of the object on a different angular scale. It is therefore possible to measure the fringe pattern for a range of telescope separations and so to cover a range of angular scales on the sky.

The amplitudes and phases of the fringes, sampled for a large number of separations of the two telescopes, are called the visibility function. When this function is subjected to the mathematical operation known as a Fourier transformation, which converts a curve of amplitude v. time into a curve of intensity v. angle, one gets the image of the object directly. It is possible to obtain the fringe pattern for all separations of two telescopes out to a given separation, *D*, and hence to obtain the same image that one would get from a single giant telescope of diameter *D*. When the rotation of the earth is exploited to increase the number of separations, the resulting operation is called earth-rotation aperture synthesis.

This method of making images by radio interferometry was developed by Sir Martin Ryle and his colleagues at the University of Cambridge in the 1950's. The method has been widely applied, for example at the Owens Valley Radio Observatory of the California Institute




| | EYE | GALILEO'S TELESCOPE | 200-INCH PALOMAR TELESCOPE |
|---|---|---|---|
| APERTURE (MILLIMETERS) | ~5 | ~25 | 5,000 |
| RESOLUTION (ARC-SECONDS) THEORETICAL / ACTUAL | ~25 / ~200 | ~5 / ~10 | ~.025 / ~1 |
| RELATIVE LIGHT-GATHERING POWER | 1 | 25 | 1 MILLION |

**TWO MAJOR FUNCTIONS OF A TELESCOPE are to collect radiation over a large area and to bring together radiation from widely separated points. The first function makes it possible to observe objects otherwise too faint to be seen with the unaided eye. The second function determines the telescope's resolving power, or ability to separate sources that otherwise would be indistinguishable. The theoretical resolving power of a telescope, defined as the angular size of the smallest details that can be seen, is given by $\lambda/D$, where $\lambda$ is the wavelength of the radiation and *D* is the diameter of the aperture. On this formula the human eye, which has an aperture of five millimeters, should have a resolving power of 25 arc-seconds, some eight times higher than its actual performance when it is looking at stars in the night sky. Galileo's best instruments had a useful aperture of about an inch (25 millimeters), with which he achieved the remarkable resolution of 10 arc-seconds, about half the value theoretically attainable. This represented a twentyfold improvement over the unaided eye. Because of the turbulence of the earth's atmosphere modern optical telescopes are limited to a resolution only about 10 times higher than the one achieved by Galileo. When the light from a star reaches the top of the atmosphere, it is coherent: its waves are all in step. Then small turbulent cells in the atmosphere delay some parts of the wave front more than others, thereby destroying the coherence of the waves and corrupting the image formed by the telescope. The resolution is established by the size, *d*, of the turbulent atmospheric cells, which is typically a few inches. Therefore the resolution at the wavelengths of visible light is limited to $\lambda/d$, or to about one arc-second.**

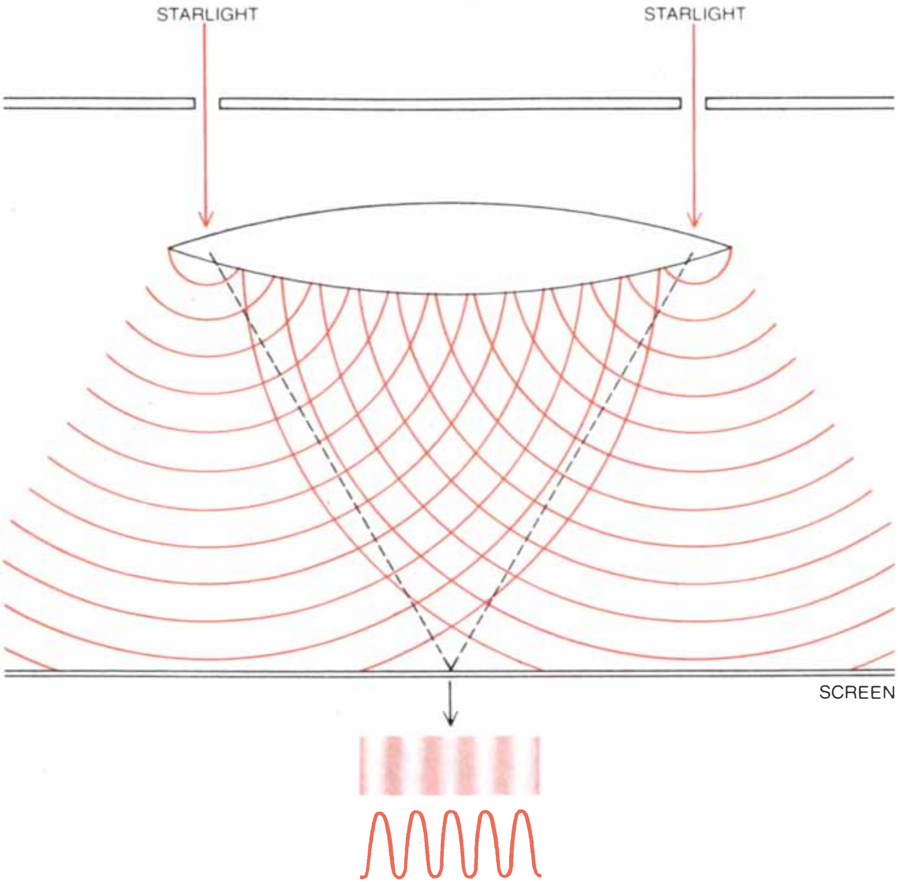
of Technology, at the Westerbork Observatory in the Netherlands and at the Very Large Array (VLA), a collection of 27 movable radio telescopes near Socorro, N.M. The radio telescopes at these observing sites are all connected directly by cable or waveguide and achieve resolutions on the order of one arc-second.

In order to achieve a resolution substantially better than one arc-second it is necessary to have telescopes hundreds or thousands of kilometers apart. Such instruments cannot simply be connected by cables. This was a problem, because the signals from the two telescopes in an interferometer must be synchronized to within a fraction of a microsecond before being added together. The advent of accurate atomic clocks in the late 1960's made it unnecessary, however, to link the two telescopes physically and led to the development of interferometers with very long baselines. In very-long-baseline interferometry the signals are recorded separately with microsecond accuracy at the two sites on the same kind of magnetic tape used for recording television images. The tapes are then shipped to a central correlator where they are synchronized and the signals are combined.

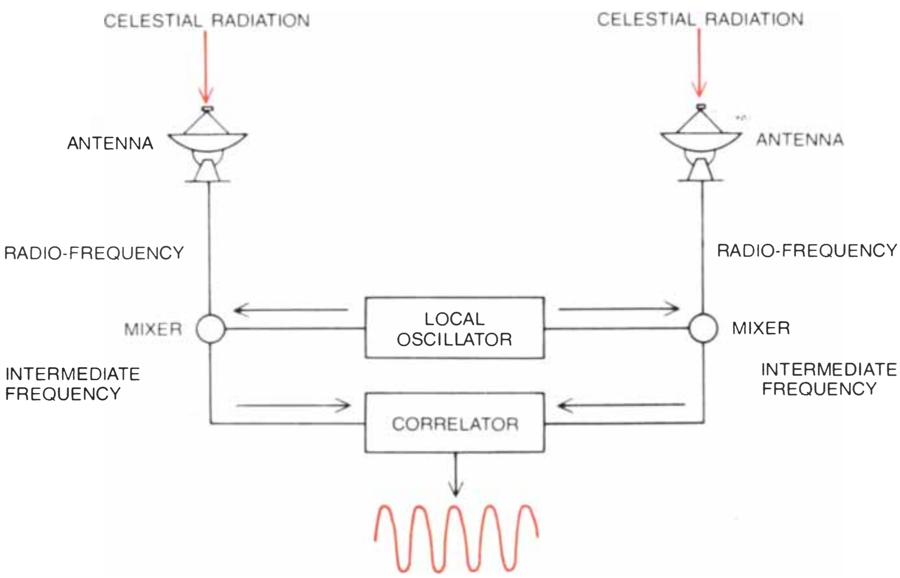
In order to obtain an image of a celestial radio source by aperture synthesis one has to measure both the amplitude and the phase of the interference fringes. The main difficulty lies in establishing the phase. Much of the effort in obtaining good images by aperture synthesis is spent in stabilizing the phase of the array of telescopes in the radio interferometer. Even after one has constructed an array with minimal instrumental errors one still has to cope with phase fluctuations caused by the atmosphere and, at the lower radio frequencies, by the earth's ionosphere and the interplanetary medium.

Phase variations due to the atmosphere and the ionosphere are not easily eliminated by calibration even when the telescopes are only a few miles apart. In very-long-baseline interferometry, where the instruments can be on different continents, the phase of the interference fringes is corrupted not only by the atmosphere and the ionosphere but also by random jumps and drifts in the atomic clocks and by ignorance of the precise distance between the foci of the two telescopes. Each of these variables or uncertainties has the same effect as slowing down or speeding up the wave front arriving at one of the telescopes, thereby causing the fringes to jitter back and forth in position and corrupting the phase measurement.

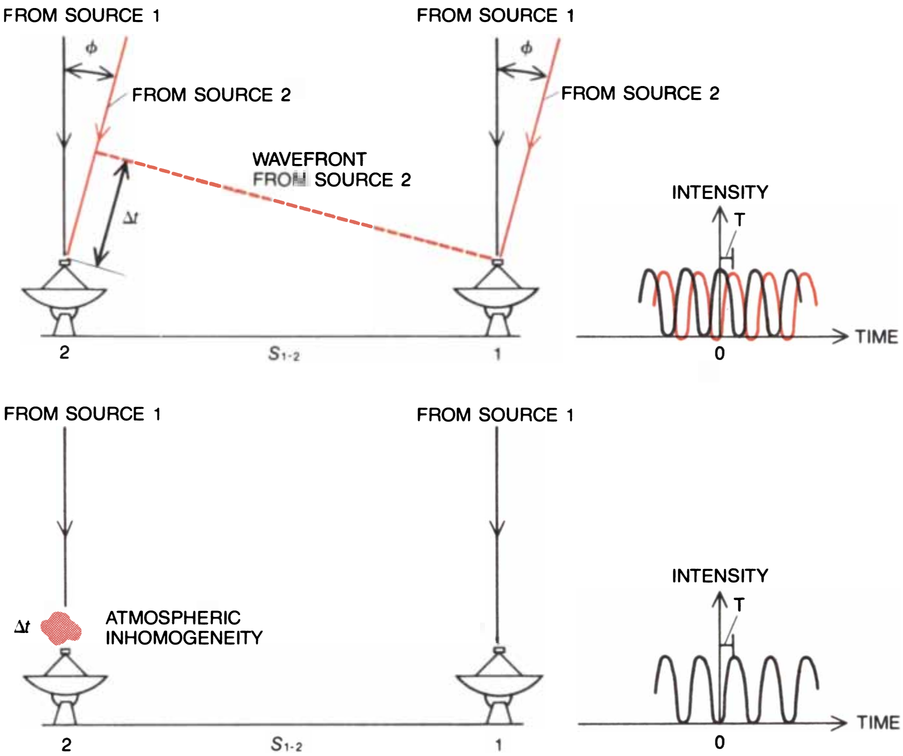
Over the past five years a number of methods have been applied to overcome this problem. Two of the methods can be used only when the celestial object being observed consists of isolated point



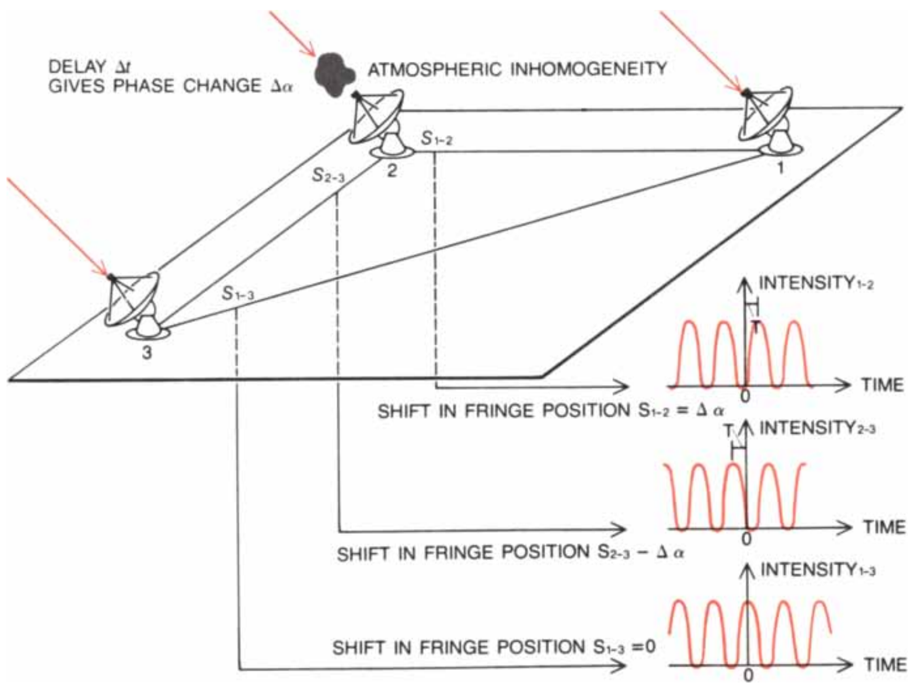

**OPTICAL INTERFEROMETER** is an instrument that adds light from two separated points originating in a common source. The waves from the two points interact to produce a pattern of interference "fringes." At positions where the crests of the two waves coincide there is constructive interference, which yields bright fringes. At positions where the crests of one wave coincide with the troughs of the other wave there is destructive interference, which yields dark fringes. Variation in intensity between the bright fringes and the dark ones follows a sine curve.




**RADIO INTERFEROMETER** is an exact analogue of an optical interferometer. If a celestial radio source is observed with a radio interferometer, the rotation of the earth causes the wave arriving at two telescopes to be successively in phase and out of phase as the two signals are added together. The signals' combined intensity again follows a sinusoidal pattern. By measuring the amplitude (size) and phase (position) for a large number of telescope separations one can "synthesize" an aperture much larger than that of the telescopes. The values of amplitude and phase define a "visibility function," which, after Fourier transformation, yields an image.




**ATMOSPHERIC TURBULENCE makes it difficult to establish both the amplitude and the phase of the interference fringes created when two radio telescopes separated by some distance $S$ are linked to form an interferometer, $S_{1\text{-}2}$. In the absence of turbulence (*top*) the output of the interferometer when it is observing source No. 1 varies smoothly with time because of the earth's rotation, passing through zenith at time $t = 0$ (*black sine curve*). The radiation from source No. 2, offset from source No. 1 by a small angle ($\phi$), produces a second fringe pattern (*color*) offset by an amount proportional to angle $\phi$. The wave front from source No. 2 arrives at telescope No. 2 at a time $\Delta t$ after it arrives at telescope No. 1. An inhomogeneity due to turbulence (*bottom*) can cause the wave front from source No. 1 to be delayed in arriving at telescope No. 2 by the same amount, $\Delta t$, simulating the same fringe shift in interferometer output.**




**SOLUTION TO THE PROBLEM OF PHASE SHIFTS, whether they are due to the atmosphere or to instrumental defects, is achieved by linking three radio telescopes to create three interferometers, $S_{1\text{-}2}$, $S_{1\text{-}3}$ and $S_{2\text{-}3}$. Here the wave arriving at telescope No. 2 has been delayed, causing an equal but opposite phase shift on interferometers $S_{1\text{-}2}$ and $S_{2\text{-}3}$. The fringes on interferometer $S_{1\text{-}3}$ are unaffected by the disturbances at telescope No. 2. When the phases around the loop are added up to obtain the "closure phase," the phase shifts introduced by the atmosphere cancel exactly. The closure phase contains only information about structure.**

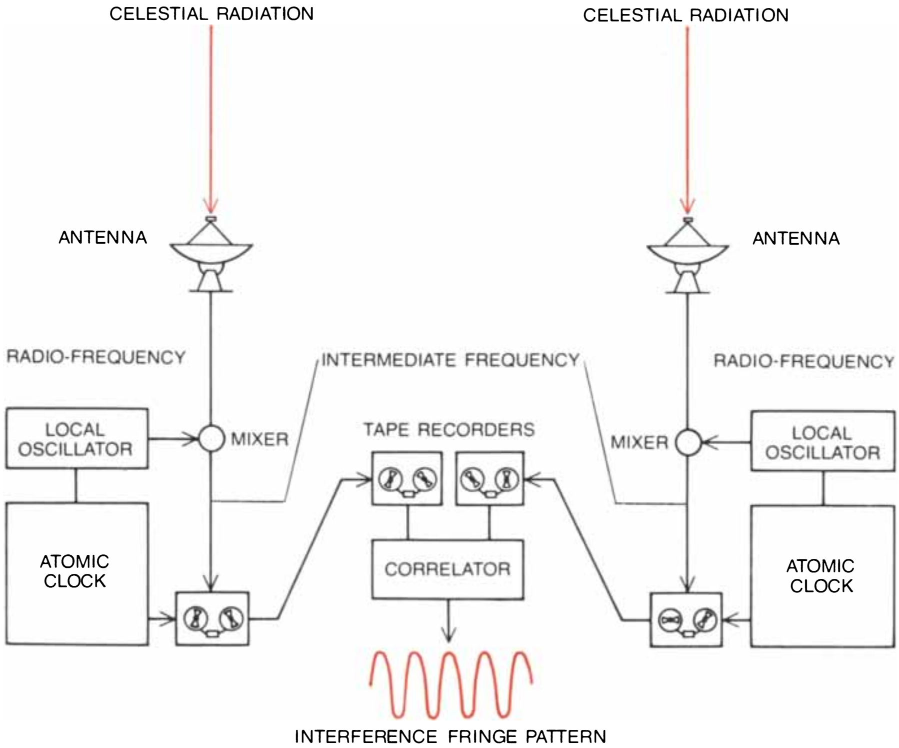
sources. They are commonly referred to as fringe-rate mapping and phase mapping, and they are not to be confused with aperture-synthesis mapping, which can be applied to any complex source. Fortunately there are some objects of great astrophysical interest that do have a number of isolated point components and can therefore be mapped by these methods. I shall briefly describe the mapping of a few such sources before proceeding to the mapping of more complex sources.

The atmospheric and instrumental effects that corrupt the determination of phase can be overcome if the object is a bright point source or if a bright point source is close to it. A point source inherently has a zero phase, and so any offset in fringe position observed for it must be due to either the atmosphere or the observing instrument or to both. Therefore all that is needed to remove the spurious effects is to subtract the phases determined for the nearby point source from the phases measured on the object. A point source exploited in this way is called a phase reference.

Some of the most remarkable bright point sources have been found in the clouds of gas around aging stars and in regions where new stars are forming. In both cases the point sources are small regions within gas clouds where molecules of water, the hydroxyl molecule (OH) or silicon monoxide (SiO) are emitting intense radiation at precise frequencies. Such sources are known as cosmic masers because they mimic the properties of a particular kind of amplifier that happens to be widely used in radio astronomy. (Maser stands for microwave amplification by stimulated emission of radiation.) Several hundred cosmic masers have been discovered since the first one was identified in 1965.

The high resolution provided by the VLBI observations has yielded structural information about cosmic masers on scales down to $10^8$ kilometers, or roughly the radius of the earth's orbit. From Doppler shifts in the frequency of the maser radiation it is possible to measure the radial velocity of the "masing" clouds. Moreover, by observing the polarization of VLBI signals from masers in star-forming regions one can determine the strength of the magnetic fields associated with clouds that are collapsing to form stars. This method was pioneered by James M. Moran, Jr., and his colleagues at the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory.

The first maps showing the disposition of water masers were made by R. Craig Walker of the National Radio Astronomy Observatory and his collaborators at the Massachusetts Institute of Technology and the Center for Astrophysics. They mapped the masers in the gas cloud designated W51 in the constellation Aquila. More recently, in a region of the Great Nebula in Orion where new stars are being born, a group working with Reinhard Genzel, who was then at the Center for Astrophysics, charted the proper motion (the actual movement across the sky) of some 40 distinct maser spots. For the Orion masers the proper motion is approximately .01 arc-second per year. The group also determined the masers' radial motion (movement along the line of sight) from the Doppler shift of each spot. These measurements, combined with the accurately mapped positions, show that the water masers are in clouds flowing at a velocity of 18 kilometers per second outward from the vicinity of two obscured sources of infrared radiation called IRC 2 and IRC 4.




**HIGHLY ACCURATE ATOMIC CLOCKS, which make it unnecessary for the two or more telescopes in the array to be physically linked, make VLBI interferometry possible. At each telescope site the radio-frequency signal from the celestial source is recorded on videotape at a rate of four million samples per second, together with periodic timing signals from a local atomic clock. The videotapes are later played back at a central processor in order to extract the pattern of interference fringes. The tapes must be synchronized to a fraction of a microsecond.**

The motion of hydroxyl masers in the region of star formation in the gas cloud W3 in the constellation Cassiopeia has recently been determined by Mark J. Reid and his collaborators at the Center for Astrophysics. With the aid of eight radio telescopes, seven in the U.S. and one in Canada, they have mapped 14 hydroxyl-maser regions in W3 with a resolution of .01 arc-second. The masers are observed projected against a bright cloud of ionized hydrogen that is energized by an obscured young star. Reid and his group have worked out a dynamical model showing that the hydroxyl masers are falling in toward the central star at a speed of five kilometers per second. The masers are evidently remnants of an accreting envelope that is still collapsing toward the newly formed star and its surrounding cloud of ionized hydrogen. With similar methods Aubrey D. Haschick and his colleagues at the Center for Astrophysics and the Haystack Observatory of M.I.T. have mapped the hydroxyl masers in the star-forming region W75N in the constellation Cygnus. They suggest that the masers have formed around a young double-star system.

One of the most exciting applications of this work is the direct determination of distances in special cases. For example, the distance to the water masers in Orion can be calculated because it appears that the spots are expanding homogeneously from a point. The expansion can therefore be determined by the "expanding cluster parallax" method. By putting the observed radial velocities, the proper motions and accurate positions into a model of homogeneous expansion one can determine the position in depth of each spot with respect to the center of expansion. Knowing the three-dimensional position of each spot, one can calculate its transverse motion from the model of expansion. The distance to the entire region can then be determined by scaling the observed proper motions, which are angular velocities, to the derived transverse motions, which are true velocities. When this has been done for each water-maser spot, the optimal solution can be secured by a least-squares fit of all the spots. By this method the distance to the Orion nebula is found to be 1,550 ± 250 light-years.

An alternative method of distance determination, the "statistical parallax" method, can be brought into play when the motion of the masers is random. One simply compares the spread in radial velocities, obtained from the Doppler shifts, with the spread in the proper motions, obtained from observations at different times. The ratio of the two spreads gives the distance directly. The method has been applied to the water masers in the star-forming region W51. Genzel, Matthew H. Schneps and their collaborators at the Center for Astrophysics have shown that this cluster of masers is at a distance of 23,000 ± 5,000 light-years.

These new methods of distance determination are accurate to within 10 to 20 percent. They are important because they bypass many of the problems in the traditional astronomical methods of distance determination, and they can be applied throughout the galaxy. Indeed, water masers have recently been detected in a nearby spiral galaxy, M33 in Triangulum, and so the new methods may be applicable to nearby galaxies as well.

Unfortunately the aperture-synthesis methods that are so successful with bright point sources cannot be applied to the mapping of more extended sources such as quasars and radio-emitting galaxies. Except in a few special cases there are no conveniently situated point sources in these objects or near them to serve as phase references. Hence until about 1975 it was thought unlikely that very-long-baseline interferometry would ever yield good images of the powerful extragalactic radio sources. As is so often the case, however, the limitations of established methods led to a new method that could have been applied with profit to conventional interferometry years earlier.

The breakthrough came with the application of an idea that had lain dormant for more than 20 years. In 1953 Roger C. Jennison, a graduate student at the Jodrell Bank Radio Observatory in England, realized one could derive useful information about the structure of a radio source even when the phase information was badly corrupted. Jennison saw that if the fringe phases around a closed loop of three or more interferometers are added together, all the spurious contributions due to propagation effects and instrumental errors precisely cancel. The resulting sum of phases contains only information about the structure of the object being observed.

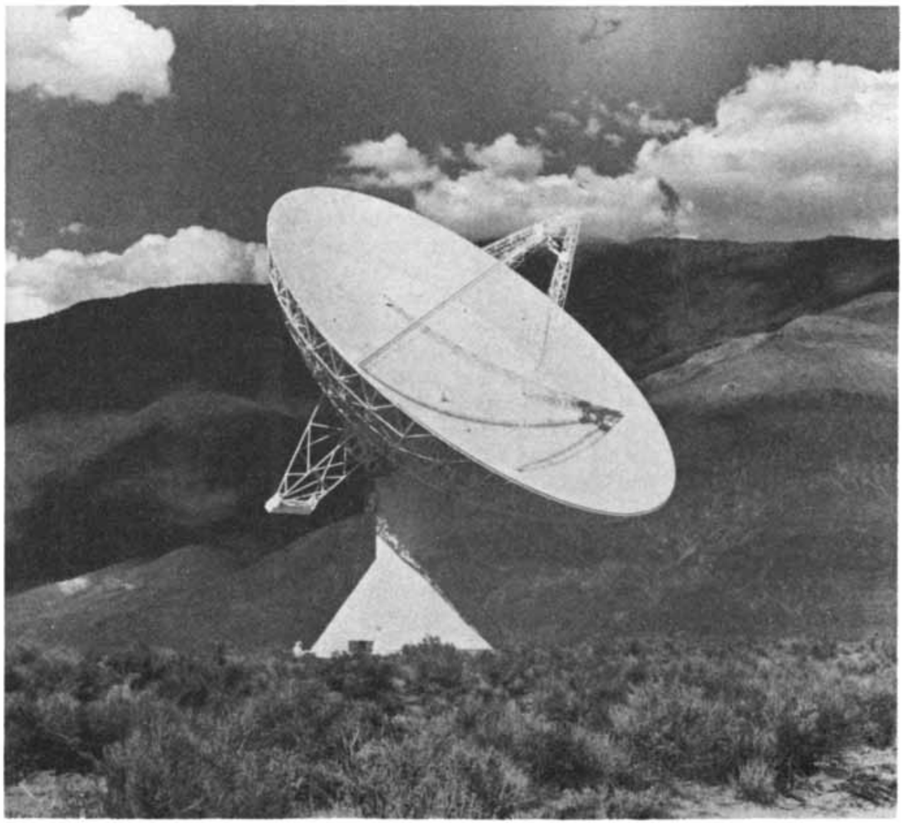
This comes about in the following

way. Consider a group of three telescopes, 1, 2 and 3, linked in a triangular system each of whose legs is an interferometer. Thus there are three interferometers—telescopes 1 and 2, telescopes 1 and 3 and telescopes 2 and 3—each generating its own pattern of interference fringes. If the atmosphere delays the wave arriving at telescope 2, this does not affect the fringes on interferometer 1–3. The fringes on interferometers 1–2 and 2–3, however, are each shifted by an equal but opposite amount. One can see that if the phases around the closed loop are added up, the phase shifts cancel exactly. The same argument applies to a delay associated with telescope 1 or telescope 3.

This idea was reconceived and applied to very-long-baseline interferometry in 1974 by Alan E. E. Rogers of the Haystack Observatory, who was initially unaware of the earlier work by Jennison. Rogers called the sum of the phases around a closed loop the closure phase. Stimulated by Rogers' paper, Peter N. Wilkinson and I, working at Cal Tech, developed a method, now known as hybrid mapping, for generating reliable images based on the closure phase from VLBI observations. A generalization of this method to a closed loop of four telescopes enables us to extract a corrected size or amplitude from four sets of fringes by measuring what is called the closure amplitude.

The closure techniques exploit the celestial object itself as a phase reference and as a means of calibrating the phase amplitude. At Cal Tech we have recently extended the method so that one can get good images out of completely uncalibrated, indeed uncalibratable, data. The new methods have revolutionized not only the making of images but also the designing of radio-telescope arrays. Moreover, they are not confined to radio observations but can be applied to any waves that give rise to detectable interference fringes.

With the advent of closure techniques it became possible to make detailed images with synthesized radio telescopes thousands of kilometers across. There is now a compelling scientific case for constructing an array of telescopes specifically designed for very-long-baseline interferometry. After intensive studies our group at Cal Tech and a group at the National Radio Astronomy Observatory have proposed construction of an array of 10 radio telescopes 25 meters in diameter extending from Hawaii and Alaska to the East Coast of the U.S. With hybrid-mapping methods these arrays would be the equivalent of a single telescope with an aperture 7,500 kilometers in diameter. So far no telescope has been expressly designed and built for very-long-baseline interferometry, but there are telescopes built for other purposes scattered around the globe that can serve to test the feasibility of the scheme. The images formed by such ad hoc networks are necessarily crude. For very-long-baseline interferometry to achieve its full potential we must await the construction of the kind of array we have proposed.



**FORTY-METER RADIO TELESCOPE of the Owens Valley Radio Observatory, of which the author is director, was completed in 1967. Unlike optical telescopes, a radio telescope can be operated around the clock, provided observations are not made too close to the sun. The Owens Valley dish often operates as one part of an interferometer array that typically includes observations made at Hat Creek in California, Fort Davis in Texas, North Liberty in Iowa, Algonquin Park in Ontario, Green Bank in West Virginia, the Haystack Observatory in Massachusetts, the Jodrell Bank Radio Observatory in England and Effelsberg in West Germany.**

Hybrid mapping was first applied in 1975 to observations of the quasar 3C 147, one of the many "quasi-stellar" radio sources that populate the farthest reaches of space. In photographs made with the largest optical telescopes quasars resemble faint ordinary stars. The first radio image, which had a resolution of .01 arc-second, revealed an object with a bright core and a curious one-sided jet [see "Cosmic Jets," by Roger D. Blandford, Mitchell C. Begelman and Martin J. Rees; SCIENTIFIC AMERICAN, May]. The following year 3C 147 was observed at the higher resolution of .003 arc-second with a network of five telescopes: two in California (Owens Valley and Hat Creek), one in Texas (Fort Davis), one in West Virginia (Green Bank) and one in West Germany (Effelsberg, near Bonn). In this image the jet was resolved into a series of irregular blobs with an overall length of .2 arc-second. In optical images 3C 147 appears as an unresolved point one arc-second across, that is, five times the size of the jet.

These results were exciting for two reasons. They demonstrated that it is technically possible to make good images with the full resolution of which very-long-baseline interferometry is capable. This means it is really possible to synthesize a telescope of global dimensions and to make images identical with those theoretically obtainable in the absence of any atmospheric effects and instrumental errors. Of greater import was the result itself, which was surprising. The large-scale radio structure of extragalactic objects is generally highly symmetrical, consisting of two large radio lobes that straddle the optical object. In the active nucleus of 3C 147, the first such nucleus to be mapped, the structure turned out to be highly asymmetric: a single jet with a bright core at one end.

Since 1976 hybrid mapping has furnished reliable images of some 50 quasars and galaxies with highly energetic nuclei. Asymmetric structure has turned out to be a common feature of such objects. Most of them also vary in brightness on time scales of about a year. Some of them appear to be expanding at speeds paradoxically greater than the speed of light. Such "superluminal" motion was demonstrated independently by Marshall H. Cohen of Cal Tech and Irwin I. Shapiro of M.I.T. as early as

1971, on the basis of theoretical models fitting interferometer fringe amplitudes. Many astronomers, however, questioned the validity of such models.

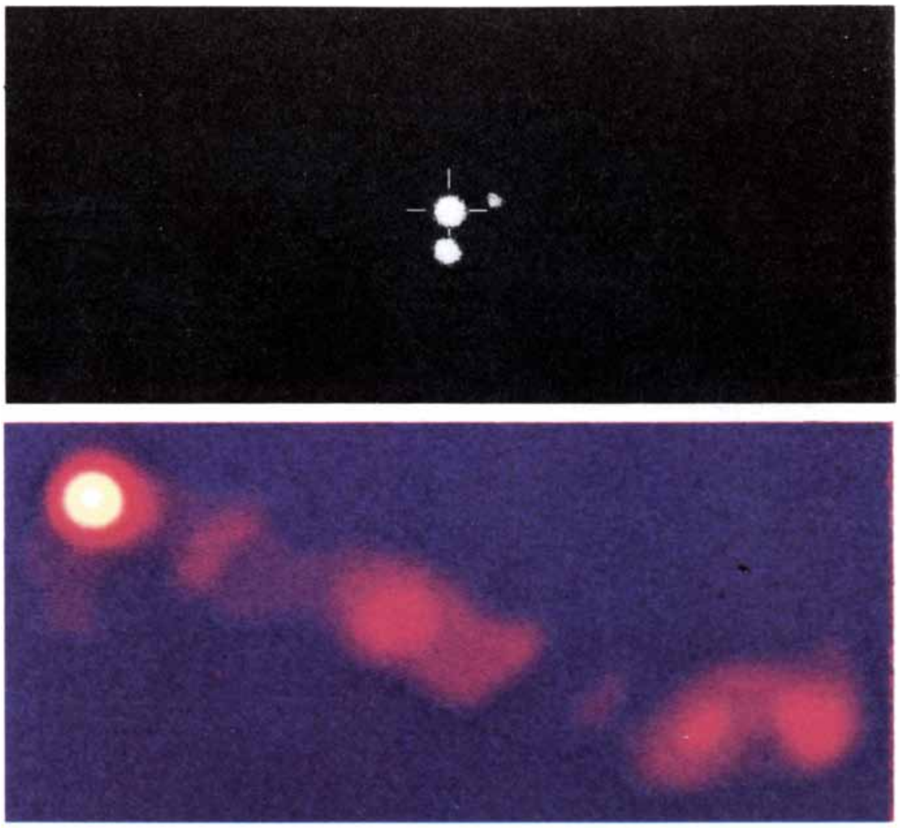
It is now clear that although the details of the models may have been incorrect, superluminal motion is real. Its reality has been convincingly demonstrated by hybrid mapping of several suspected superluminal sources. Timothy J. Pearson and his colleagues at Cal Tech have been observing the quasar 3C 273 periodically since 1977 with a VLBI array of radio telescopes in California, Texas, West Virginia, Massachusetts and, on a few occasions, Effelsberg in Germany. Between July, 1977, and July, 1980, a small blob ejected from the bright core of the quasar traveled a distance of 25 light-years and therefore seems to be moving at 9.6 times the speed of light. (In calculating the speed a small cosmological correction has been applied to the time interval.)

Now, this is only the apparent speed. The most likely explanation is that the blob is moving almost directly toward us at 99.5 percent of the speed of light. Let us suppose our instruments record two bursts of radiation that have been emitted by the blob one year apart. The bursts will seem to reach us in quick succession because the emitting object will have moved almost one light-year nearer the earth before the second burst of radiation is released. In the case of 3C 273 the second burst would arrive only 3.5 days later than the first. If one now plotted the radio image of the blob at the time of the two observations, one would find that the blob had moved laterally across the sky a distance equivalent to a tenth of a light-year and thus had seemingly traveled at 10 times the speed of light. The explanation is that the lateral displacement represents the short leg of a long, thin right triangle whose apex is pointed away from us. The blob emitted the first burst of radiation when the blob was at the far apex of the triangle and the second burst after it had actually traveled .995 light-year along the hypotenuse in our direction.

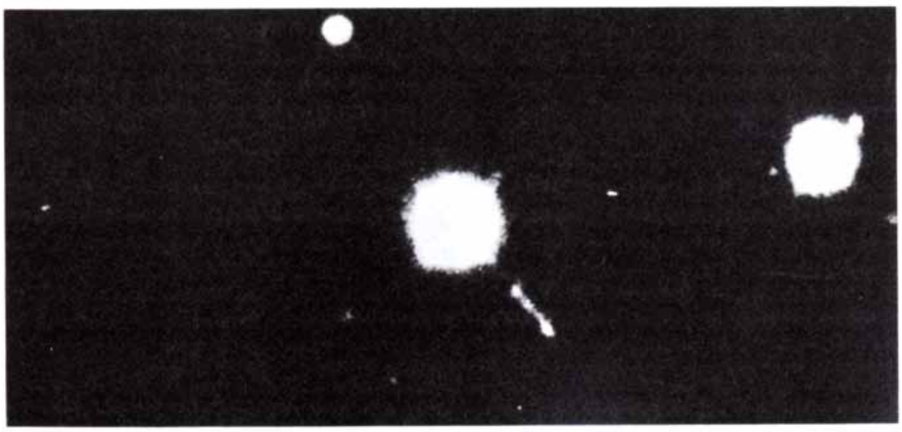
Such superluminal motion has been found in two other quasars and a galaxy of the type known as a Seyfert galaxy. It is remarkable that these active nuclei eject blobs of matter at nearly the speed of light and steadily in the same direction. It is worth noting that as early as 1967 Martin J. Rees of the University of Cambridge had suggested, on the basis of rapid variations in the brightness of such objects, that some of them are ejecting matter at speeds approaching the speed of light.

The rapid motion observed in the superluminal sources may hold the key to the one-sided structure. According to the special theory of relativity, if a radiating body is moving at nearly the speed of light, the radiation is beamed into a narrow cone in the direction of motion. It seems likely that the active nuclei in these unusual celestial sources are actually ejecting two jets in opposite directions. Only the jet beamed toward us is observed; the radiation from the jet beamed away from us is invisible. This is an attractive idea, because it serves to reconcile the large-scale symmetry observed in many extragalactic objects with the small-scale asymmetry observed in many of the nuclei. At the point where the jets plow into the intergalactic medium and are slowed down the radiation is no longer beamed but travels in all directions, so that both large blobs are observed.

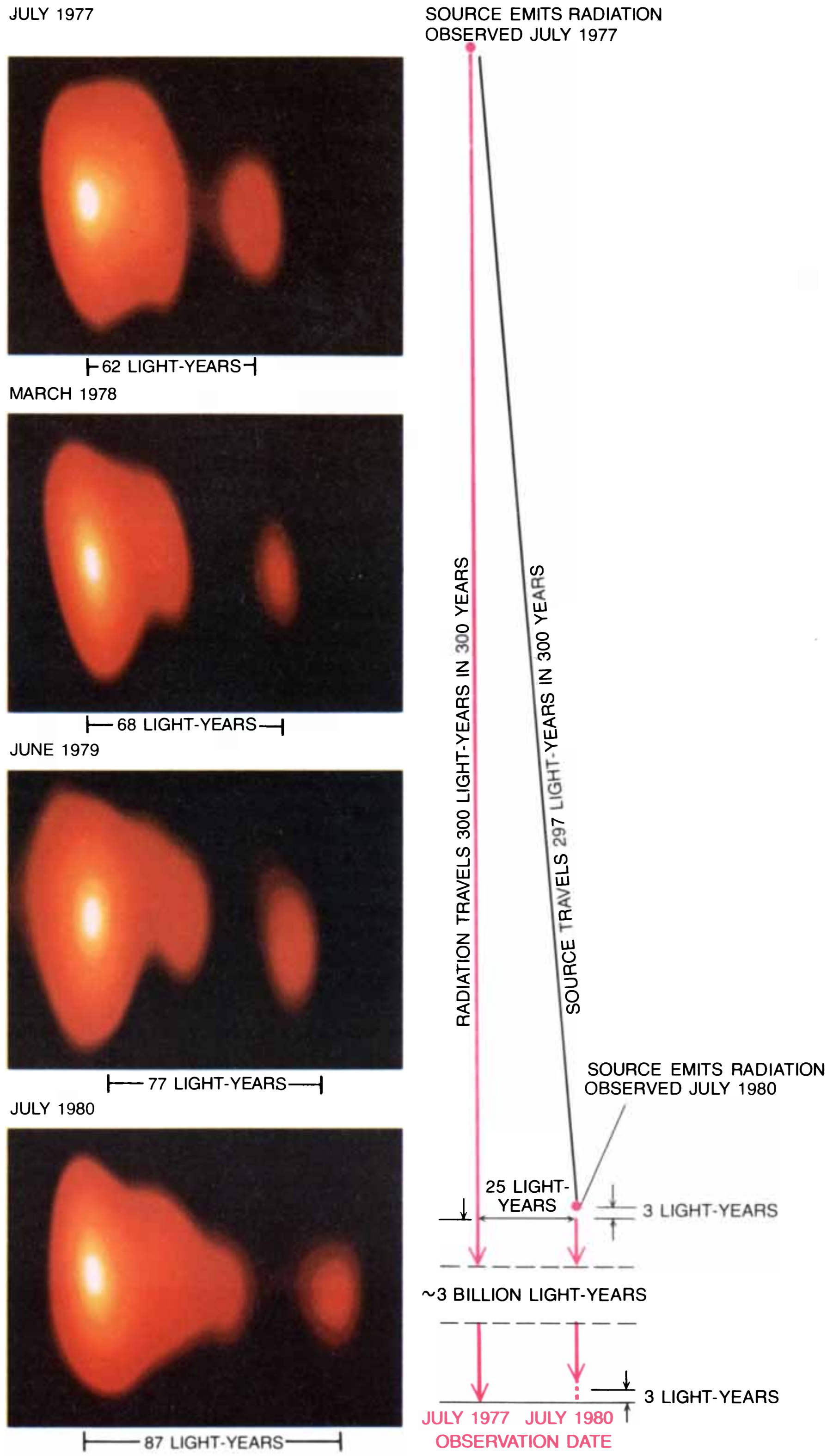
What is the nature of the energy source in quasars and radio galaxies and how is matter ejected from them?



**FIRST QUASAR TO BE MAPPED by VLBI interferometry was 3C 147, which lies at a distance of seven billion light-years. In the photograph at the top, made with the 200-inch Palomar telescope, 3C 147 looks like a typical faint star. In 1976 signals recorded by radio telescopes in Owens Valley and at Hat Creek, Fort Davis, Green Bank and Effelsberg were synthesized by the new technique of "hybrid mapping" to produce an image with a resolution of .01 arc-second (*bottom*). The arbitrarily colored radio image reveals a jet 5,000 light-years long issuing from a bright core. Length of the jet is .2 arc-second, a fifth the size of the optical image.**



**QUASAR 3C 273, the first "quasi-stellar radio source," was observed in 1963 by Maarten Schmidt of Cal Tech. He found that the light emitted by this unresolved starlike object, which coincided with a powerful radio source, was strongly shifted toward the red end of the spectrum, indicating it was receding from our galaxy with about 16 percent of the velocity of light. The VLBI mapping of 3C 273 uncovered a bizarre feature** **(*see illustration on next page*)****.**




**SERIES OF RADIO IMAGES OF 3C 273 with a resolution of .001 arc-second (*left*) were created by the synthesis of signals recorded by various combinations of either four or five telescopes in California, Texas, Massachusetts, West Virginia and West Germany. In the three-year period between July, 1977, and July, 1980, a blob ejected from the quasar appears to have traveled a distance of 25 light-years, indicating an apparent "superluminal" speed of 9,6 times the speed of light (after a small cosmological correction has been applied to the time interval). This "forbidden" velocity is actually an illusion created by the fact that the blob is moving almost directly toward us at more than 99 percent the speed of light. As is shown in the schematic diagram at the right, the radiation from the blob that reached the radio telescopes in July, 1980, was actually emitted some 300 years after the radiation that was recorded in July, 1977. In that time, however, the blob had moved to a point 297 light-years nearer the earth. Thus the two emissions reached the earth only three years apart. In the same period the blob had moved "sideways" 25 light-years (the narrow end of a thin triangle). The distance between the core and the blob in the radio images is .006 arc-second, or 1/150th the size of the optical image.**

These are two of the most fundamental questions in astronomy today. VLBI observations provide the only means of directly observing the structure in galactic nuclei. When the hybrid maps of quasars and radio galaxies are combined with radio and optical observations of lower resolution, they yield important clues to the central engines powering these dynamic objects. A particularly good example is the galaxy NGC 6251.

NGC 6251 is an elliptical galaxy in a small cluster of galaxies 400 million light-years away. At the wavelengths of visible light it appears to be a completely normal galaxy, with nothing to distinguish it from millions of similar objects. Its extraordinary properties were first revealed by observations at radio wavelengths. The illustration on the opposite page shows three radio images of NGC 6251 at three different resolutions made by three different radio telescopes. The three nested images differ in scale by a factor of more than two million. Peter C. Waggett, Peter J. Warner and John E. Baldwin made the top image with the half-mile radio interferometer at Cambridge and the middle image with the five-kilometer interferometer. The bottom image was made by Cohen, Roger D. Blandford and me with three telescopes in a VLBI observing run: one in California (Owens Valley), one in West Virginia (Green Bank) and one in Massachusetts (the Haystack Observatory).

The top image shows the overall radio structure of NGC 6251, which spans six million light-years, about 60 times the diameter of a large galaxy. The radio object is therefore one of the largest objects known. The position of NGC 6251 as recorded in optical photographs is marked by a cross. On the scale of the reproduction the optical image would be less than two millimeters in diameter. From the optical position the radio map shows a bridge extending up to the right, or northwest. A close-up of the bridge is shown in the middle image, which reveals a remarkable straight jet more than 500,000 light-years long, some five times the diameter of our own galaxy. The bright core at the eastern end of the jet is coincident with the nucleus of NGC 6251.

The bottom image shows the core enlarged by a factor of nearly 100,000. The core too is a jet, barely eight light-years long, that is aligned with the large jet to within a few degrees. The angular width of the nuclear jet is about .0001 arc-second, equivalent to the width of a human hair observed from a distance of 80 kilometers. The alignment of the two jets must have persisted for upward of 500 million years, the time it would take matter moving at close to the speed of light to travel from the nucleus to the end of the large jet.

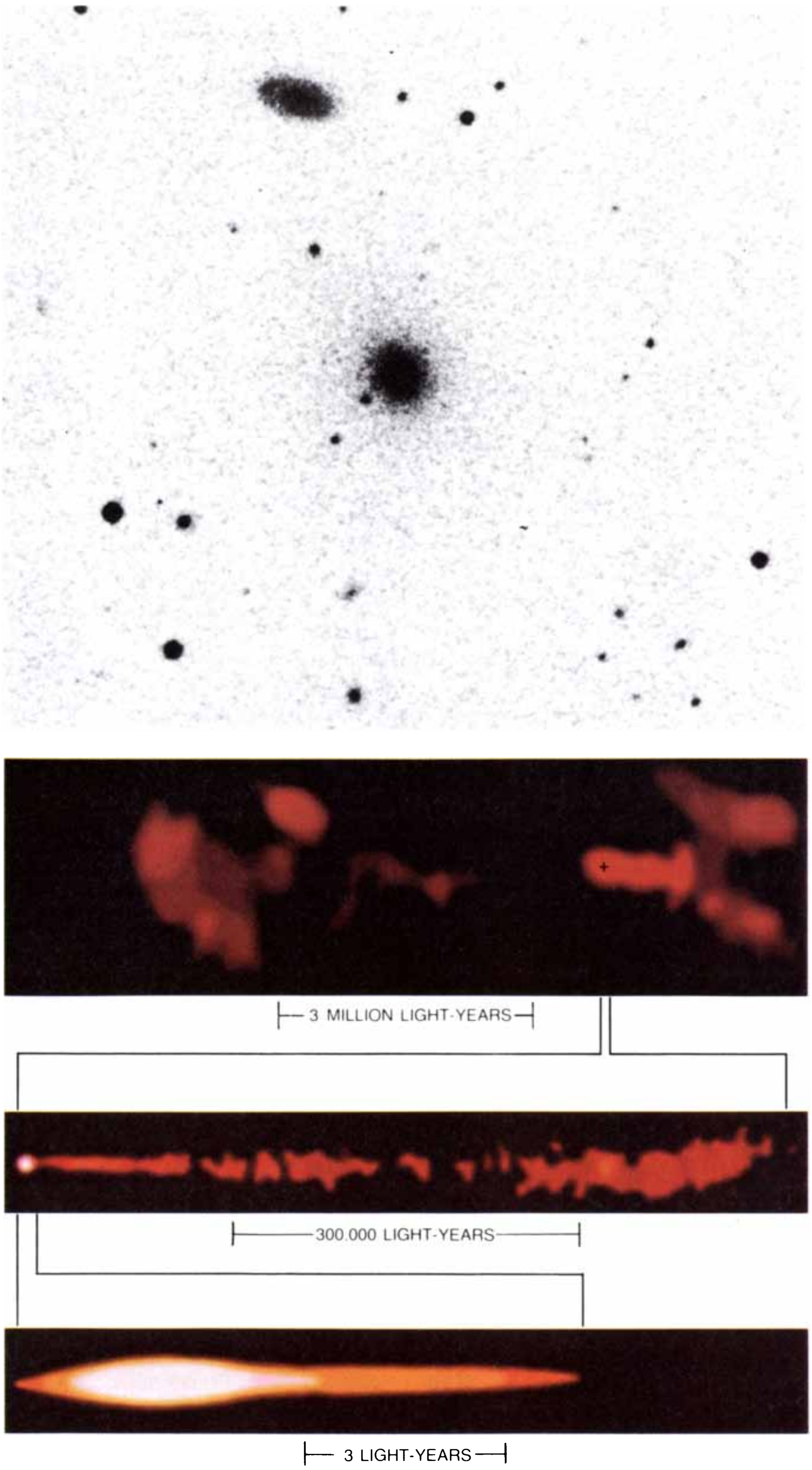
The radio properties of NGC 6251

place heavy demands on the central engine responsible for the observed activity. First, the engine must produce a vast amount of energy, equivalent to the total conversion of the mass of 10 million suns into energy, in a region much less than a light-year across. Second, it must focus the ejected matter into a cone only three degrees wide. Third, it must remain stable and keep the jet stable to within about a degree while it ejects an amount of mass equal to the mass of a small galaxy. Finally, it must be able to eject matter at close to the speed of light if it is the same kind of engine as the one responsible for superluminal sources.

Only one kind of object is known that is theoretically capable of satisfying all these requirements: a spinning supermassive black hole. The object would have to have the mass of about a billion suns, and it would have to be spinning so that gyroscopic action would keep it stable. The accurate and persistent alignment of the jets can be explained if one assumes that they are being extruded along the black hole's axis of spin. Further evidence that there may be a supermassive object in the nucleus of NGC 6251 comes from optical observations made recently by Wallace L. W. Sargent, Peter J. Young and their collaborators at Cal Tech. With the technique called speckle interferometry they have found an intense spike of radiation in the central arc-second of the nucleus. They deduce that concentrated in this small region must be a billion solar masses of matter.

If the engine in the nucleus of NGC 6251 is indeed a black hole, its power would arise from the release of gravitational energy rather than the nuclear energy that makes stars shine. The mass requirements would thereby be reduced by a factor of about 50. The ejection of matter at high pressure and at nearly the speed of light would be most easily accomplished in the deep potential well surrounding the black hole into which nearby matter was drawn at velocities close to the velocity of light. A spinning black hole of a billion solar masses would therefore account in a plausible and economical way for the observed properties of this remarkable object.

If this conclusion is correct, and if there are in fact supermassive black holes in the nuclei of galaxies, can their existence ever be demonstrated observationally? By pushing VLBI observations to higher frequencies it will be possible to improve on the resolution achieved so far by a factor of 10. At this tremendous resolution, considerably better than .0001 arc-second, one should be able to map the structure of NGC 6251 and other active galaxies on a scale not much bigger than the size of the black hole itself. We may then finally learn whether the central object is indeed a black hole or an object of an entirely new kind.




**JETS IN THE GALAXY NGC 6251,** an elliptical galaxy 400 million light-years away, have been revealed in a series of radio images made at increasing resolution. In an optical image (*top*) made with the 200-inch Palomar telescope NGC 6251 seems to be an elliptical galaxy of no particular distinction. The galaxy is, however, a strong radio source. The uppermost of the three radio images was made with the half-mile radio interferometer at the University of Cambridge by Peter C. Waggett, Peter J. Warner and John E. Baldwin. The optical position of NGC 6251 is marked by a small cross. The middle image, which was obtained by the same workers with the five-mile telescope at Cambridge, shows a jet more than 500,000 light-years long with the galactic core at the extreme eastern (*left*) end. The bottom image discloses an intense jet that is barely eight light-years long issuing from the core. The jet is only about .0001 arc-second wide. The image was made by Marshall H. Cohen, Roger D. Blandford and the author with the aid of the Owens Valley, Green Bank and Haystack radio telescopes.

# Calmodulin

*Calcium ions are important intracellular messengers. Often their message is relayed by calmodulin, a ubiquitous protein that binds calcium ions and is thereby activated to regulate various enzymes*

by Wai Yiu Cheung

For an organism to function its cells must communicate with one another. They do so through direct contact or through signals delivered by an electrical impulse or a chemical messenger. The cells are divided into various compartments and subcellular organelles, and these too must communicate. Any signal requires a receiver; for a chemical messenger the receiver is typically a protein that senses the arrival of the messenger and interprets the message by regulating the appropriate cellular activity. Perhaps the most versatile intracellular messenger is the calcium ion. Its major receptor, which appears to mediate and modulate most of the ion's manifold activities, is the ubiquitous protein called calmodulin.

The importance of the calcium ion as a cellular regulator began to be appreciated as long ago as 1883, when the British physiologist Sydney Ringer learned that muscle contraction can be maintained in an isolated frog heart only if the ion is present in the medium bathing the heart. He went on to show that many other physiological activities also require calcium. In the 1950's L. V. Heilbrunn of the University of Pennsylvania showed that injecting the ion into a muscle fiber causes the fiber to contract. Since then it has become clear that the calcium ion affects almost all aspects of cellular physiology. (It is the ionic form of calcium, $Ca^{++}$, that is active under physiological conditions because in the watery internal environment the calcium atom gives up two electrons and becomes a divalent cation: an ion with two units of positive electric charge.)

I have mentioned calcium's role in muscle contraction. It also mediates endocytosis and exocytosis (the intake and output of substances through the cell membrane), the motility of cells, the movement of chromosomes prior to cell division and perhaps the division process itself. It has a central role in the metabolism of glycogen, the storage form of glucose. And it has an influence on both the synthesis and the release of neurotransmitters, the molecules that carry a signal from one nerve cell to another. Notwithstanding these many prominent functions of calcium, however, its mechanism of action remained obscure until quite recently.

The first suggestion of how the ion works at the molecular level came in the 1950's from investigations of muscle activity. It was found that in skeletal and cardiac muscle (which is called striated muscle because of its striped appearance in a micrograph) calcium binds to a protein called troponin *c*. It was still not known how calcium causes contraction in smooth muscle (most involuntary muscle) and contractile movements in nonmuscle cells, or how it works in other tissues. The answers to many of these questions have involved calmodulin. In this article I shall first recount the story of calmodulin's discovery and then tell what is known about the structure of the protein and about its many far-ranging activities.

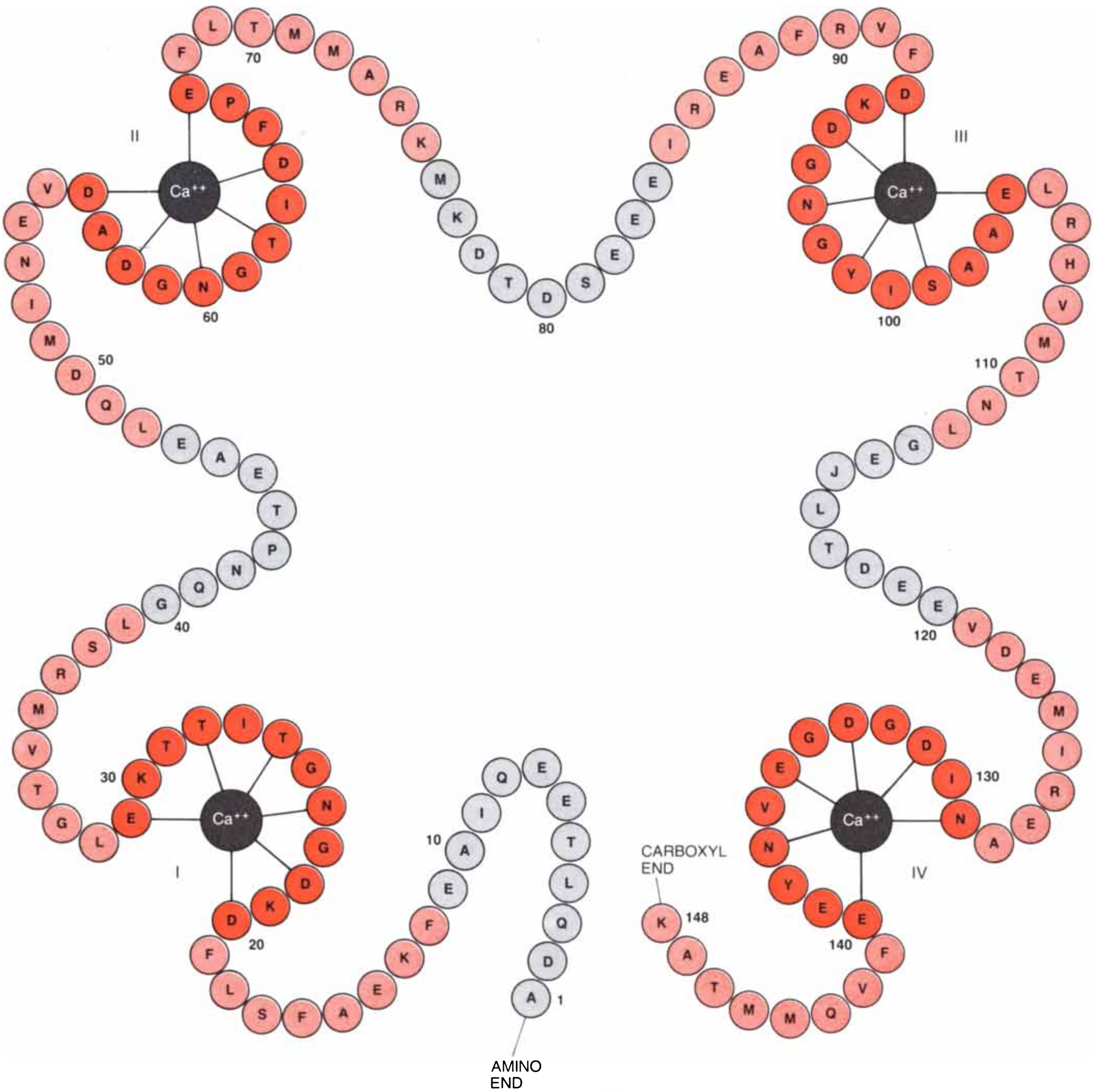
In the late 1950's Earl W. Sutherland, Jr., of the Western Reserve University School of Medicine discovered the mechanism whereby the signal delivered by a hormone to a cell is received and turns on a cellular function. The hormone was glucagon, the cells were those of the liver and the function was the breakdown of stored glycogen to supply glucose when the body is under stress. Sutherland found that the binding of glucagon to a receptor on the cell's outer membrane activates an enzyme, adenylate cyclase, embedded in the membrane and extending through it. The adenylate cyclase converts adenosine triphosphate (ATP) in the cytoplasm of the cell into the specialized nucleotide cyclic adenosine monophosphate (cyclic AMP). The cyclic AMP serves as an intracellular messenger that relays the hormone's message to the biochemical machinery of the cell, causing it to break down glycogen. Soon cyclic AMP was shown to function as an intracellular messenger in a wide variety of cells that respond to various hormones.

In 1964, as a postdoctoral fellow at the Johnson Research Foundation of the University of Pennsylvania, I was studying variations in the intracellular concentration of cyclic AMP. I needed to understand the regulatory properties of the enzyme phosphodiesterase, which breaks down cyclic AMP to the nucleotide 5'-adenosine monophosphate and thus terminates its signal. In order to purify a phosphodiesterase from the bovine brain (a rich source of the enzyme) I passed a crude brain extract through an ion-exchange chromatography column. The negatively charged resin in such a column binds different components of a crude extract more or less strongly, so that when an appropriate solution is subsequently poured through the column and successive fractions are collected, the component being sought is concentrated in certain fractions. If the component is an enzyme such as phosphodiesterase, one expects those fractions to show more enzyme activity than the others.

To my surprise the partially purified phosphodiesterase showed substantially less enzyme activity than crude brain extract. I repeated the ion-exchange procedure and the assays for enzyme activity and could find no error. It was a peculiarity of the assay procedure that led to the solution of the puzzle and to the discovery of calmodulin. I was doing the assay in two stages, first treating cyclic AMP with phosphodiesterase and then treating the 5'-adenosine monophosphate product with snake venom, which includes an enzyme that yields adenosine and phosphate. The quantity of phosphate was then measured to assess phosphodiesterase activity.

On the basis of past experience the assay results should have been the same whether the two-stage procedure was followed or the assay was done in one step, by mixing the snake venom with the phosphodiesterase. Because I was baffled by the loss of phosphodiesterase activity I repeated the assay carefully, doing it in one stage as well as in two stages. In the case of the crude brain extract there was no difference: the same level of enzyme activity was indicated by both procedures. In the case of the purified enzyme, however, there was a difference: the one-stage assay revealed significantly more activity. The two-stage assay had shown a loss of phosphodiesterase activity because snake venom was not present during the phosphodiesterase reaction. When I added snake venom to the purified enzyme, the phosphodiesterase activity increased; venom added to the crude extract had no effect.

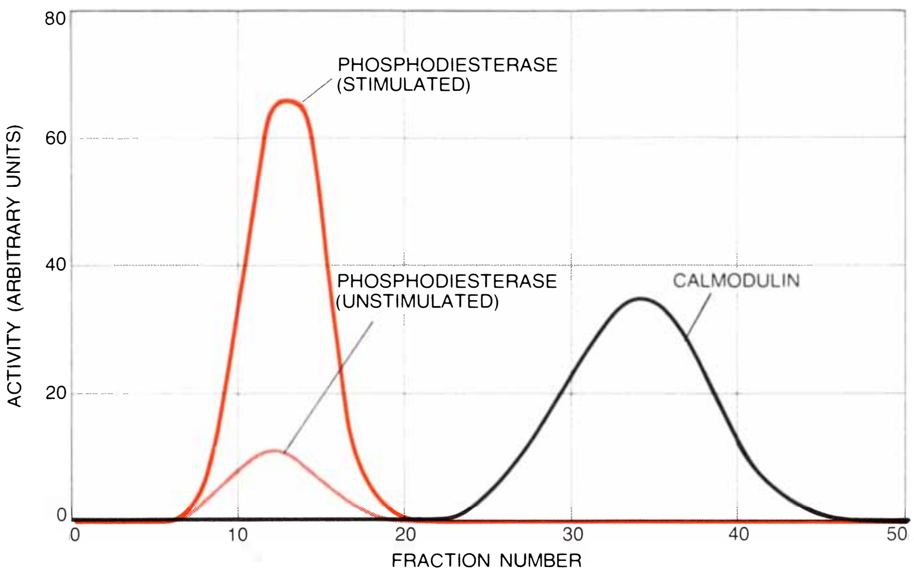
One possible explanation of these results was that an activating factor in the crude extract was removed in the purification procedure and the snake venom had some effect that made up for the factor's absence. In 1967 (by which time I had moved to St. Jude Children's Research Hospital in Memphis) I was able



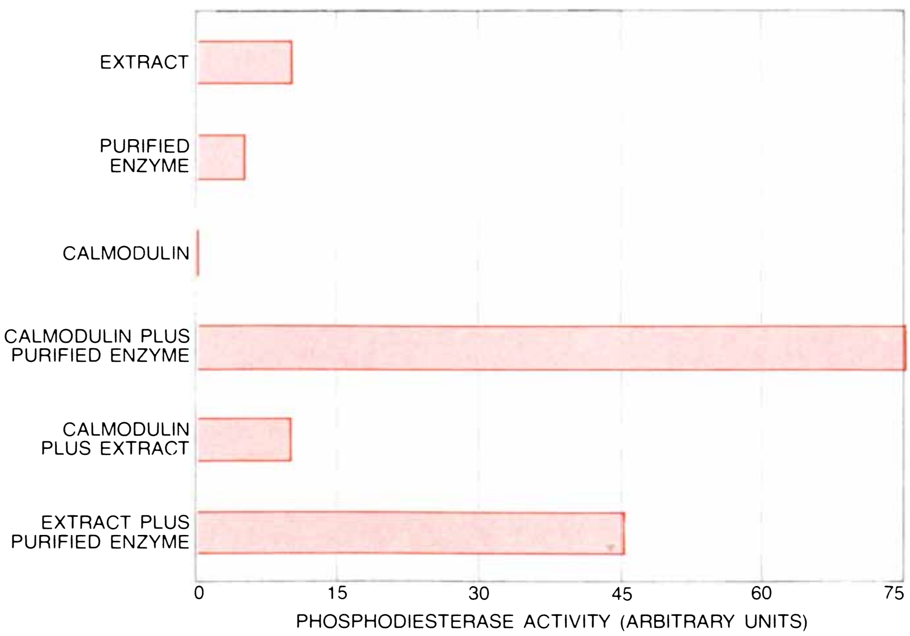

A ALANINE
D ASPARTATE
E GLUTAMATE
F PHENYLALANINE
G GLYCINE
H HISTIDINE
I ISOLEUCINE
J TRIMETHYL LYSINE
K LYSINE
L LEUCINE
M METHIONINE
N ASPARAGINE
P PROLINE
Q GLUTAMINE
R ARGININE
S SERINE
T THREONINE
V VALINE
Y TYROSINE

**CALMODULIN MOLECULE, which mediates many of the regulatory functions of calcium ions, is a single protein chain of 148 amino acid subunits. The chain has four very similar domains, each with a site that binds a calcium ion ($Ca^{++}$). The schematic diagram of the molecule emphasizes the putative structure of the four calcium-binding sites as proposed by Robert H. Kretsinger of the University of Virginia. Kretsinger based his proposed structure on an X-ray crystallographic study of parvalbumin, a calcium-binding protein with amino acid sequences closely related to those of calmodulin. Each binding site is a loop (*dark color*) flanked by a helical region (*light color*). Each calcium ion is thought to be bonded to six amino acids. The diagram gives the amino acid sequence of calmodulin from bovine brain. The sequence was established by Thomas C. Vanaman of the Medical Center of Duke University and colleagues.**

to report that one of the chromatographic fractions did indeed include a factor required for phosphodiesterase activity. The factor was a protein, but it was clearly not phosphodiesterase itself. Adding the protein to the purified, inactive enzyme restored enzyme activity; adding it to the crude brain extract, where it presumably was already present, did not increase activity. The fact that bovine brain phosphodiesterase is fully active only in the presence of a protein activator that can be removed from the enzyme was soon confirmed by other workers, notably Shiro Kakiuchi of Osaka University.




**PURIFICATION of the enzyme phosphodiesterase from an extract of bovine brain tissue separates the enzyme from calmodulin and thereby reduces the enzyme's activity. The extract is passed through an ion-exchange column. Proteins adsorbed onto the resin in the column are eluted (washed out) more or less quickly depending on their electric charge. The eluted proteins are collected in successive fractions whose activity is measured. The curves show phosphodiesterase activity (*color*) and calmodulin activity (*black*). The purified enzyme, unstimulated by calmodulin, has low activity. When either calmodulin or snake venom (a reagent in the assay for phosphodiesterase) is added to phosphodiesterase fractions, it stimulates enzyme.**



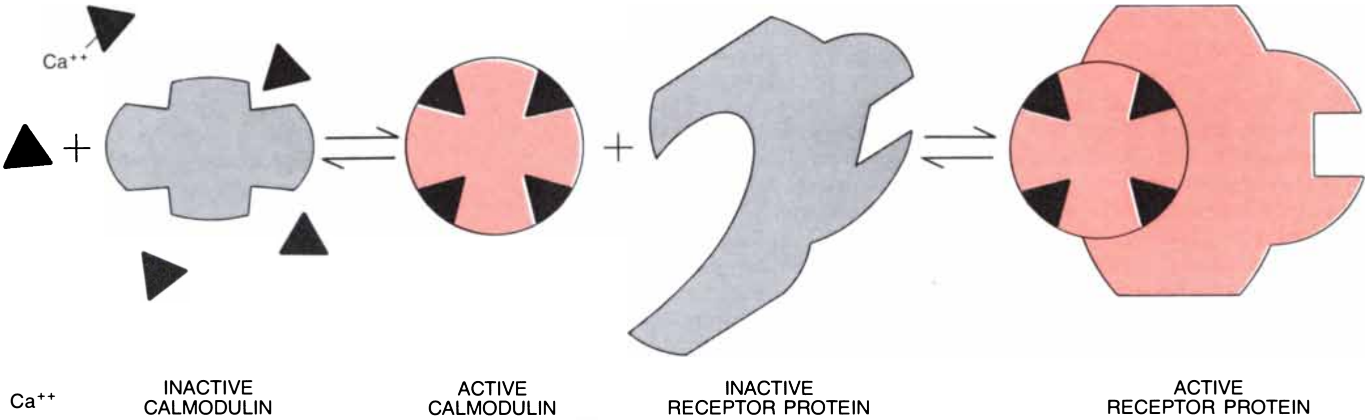

**PHOSPHODIESTERASE ACTIVITY is reconstituted with calmodulin. Phosphodiesterase purified in the chromatographic column has less activity than the crude brain extract. The calmodulin fraction has no phosphodiesterase activity. Adding calmodulin to the purified enzyme markedly increases enzyme activity. Adding calmodulin to the brain extract has no effect because the extract includes enough calmodulin to stimulate the enzyme. The presence of ample calmodulin in the extract is confirmed by mixing extract with purified phosphodiesterase. Mixture shows a level of activity much higher than sum of activities measured separately.**

The protein was calmodulin, but it got that name only much later, after its relation to calcium was well established. In the early days it was clear only that the protein has no enzymatic activity of its own, that it is remarkably stable when subjected to heat and that it interacts with phosphodiesterase stoichiometrically (in specific proportions) rather than catalytically (like an enzyme). In our laboratory and others methods were developed for purifying the protein and for characterizing it biochemically. In 1973 Jerry H. Wang of the Faculty of Medicine of the University of Manitoba found that calmodulin binds calcium. This explained some earlier observations. Kakiuchi had noted that the brain phosphodiesterase requires calcium ions for full activity and that calmodulin increases the enzyme's response to the presence of calcium; I had found a small amount of calcium to be associated with a partially purified phosphodiesterase that retained some calmodulin. Soon the relation of calmodulin and calcium was clarified. Calmodulin mediates calcium's effect on phosphodiesterase; the function of the ion is to activate calmodulin, which in turn activates phosphodiesterase.

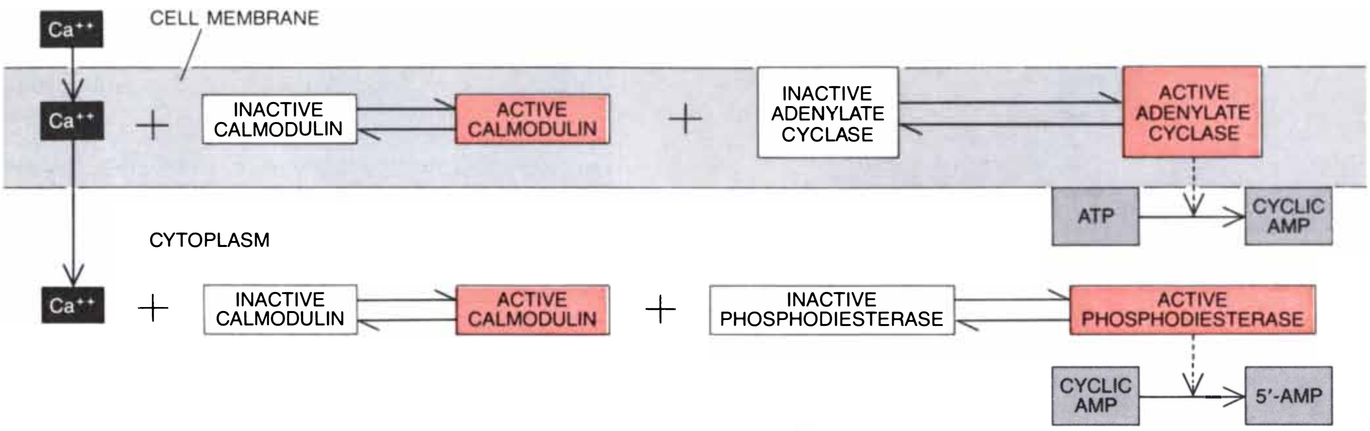
With calmodulin under investigation in a number of laboratories it was not long before the protein was characterized in detail. It is a single chain of 148 amino acid subunits, with a molecular weight of 16,700. A third of the subunits are either glutamate or aspartate, amino acids with acidic side chains that supply negatively charged carboxylate ($COO^-$) groups; it is primarily these groups that bind to the calcium ion. The protein appears to be folded into four roughly matching domains, each of which has a calcium-binding site. The affinity of the sites for calcium is such that whether the ion is bound or not depends on its concentration in the intracellular environment. That degree of affinity is an essential property for a signal-sensing receptor.

Calmodulin is a notably tough molecule. It withstands a very low (acidic) *p*H and boiling water, conditions that disable most proteins. Within cells calmodulin is found free in the cytoplasm or associated with membranes and organelles. The steady-state concentration of the protein varies with the kind of cell and with conditions. The important thing is that the concentration does not seem to limit the rate of enzymatic reactions, that is, there always seems to be more than enough calmodulin.

In addition to the high proportion of glutamate and aspartate, calmodulin is notable for the complete lack of two easily oxidized (and therefore easily degraded) amino acids, tryptophan and cysteine. The lack of cysteine (which includes a sulfur atom) means that disul-




**MECHANISM by which calmodulin mediates the biological action of calcium ions is depicted in this highly schematic diagram. Neither calcium alone nor calmodulin alone is active. The binding of four calcium ions to calmodulin changes the shape of the protein and thereby activates it. The activated calmodulin is able to interact with an enzyme (or some other protein). The interaction changes the enzyme's shape, with the result that the enzyme is activated. The flexibility of the calmodulin molecule is probably an important factor in process.**




**SEQUENTIAL ACTIVATION of the enzymes adenylate cyclase and phosphodiesterase in the brain (and perhaps in some other tissues) may proceed according to this scheme. When the arrival of a nerve impulse makes the cell's outer membrane permeable to calcium, the ions enter the membrane. They activate calmodulin in the membrane, which in turn activates adenylate cyclase; this enzyme catalyzes the conversion of adenosine triphosphate (ATP) into the intracellular messenger cyclic adenosine monophosphate (cyclic AMP). After having fulfilled some message-bearing function in the cell the cyclic AMP is degraded to 5'-adenosine monophosphate (5'-AMP) by phosphodiesterase. The breakdown enzyme, like the synthesizing enzyme adenylate cyclase, is activated by calmodulin. It is thought that calcium ions, having diffused through the cell membrane, bind to cytoplasmic calmodulin, which then activates phosphodiesterase.**

fide bridges cannot be formed between various regions of the chain. The absence of such bridges, together with the lack of hydroxyproline (a modified amino acid that often braces a turn of the protein chain), makes calmodulin very flexible. An ability to resume its normal configuration after being distorted by harsh conditions probably contributes to its stability. The flexibility is also likely to be central to calmodulin's mode of action, as I shall explain below.

The versatility of calmodulin is matched by that of its partner, calcium. The ion is abundant in all biological systems. Its value as a signal is based in part on the fact that the concentration of the ion is between 1,000 and 10,000 times as high in the extracellular fluid as it is inside the cell. Indeed, the cell cannot tolerate a high internal calcium level because the ion combines with essential organic phosphates such as ATP, sequestering them in insoluble salts. The steep steady-state differential between the extracellular and the intracellular concentrations is maintained by the cell membrane, which is ordinarily highly impermeable to the ion, and by mechanisms for disposing of excess calcium; either it is expelled from the cell by an enzymatic pump in the membrane or it is taken up by an intracellular organelle. When the cell is stimulated by an electrical impulse or some other signal, the membrane becomes momentarily permeable to calcium; the influx of the ion that results is detected as a message.

The mechanism by which the calcium ion regulates enzymes such as the calcium-dependent phosphodiesterase can now be described in some detail. The process consists of two successive activations. Phosphodiesterase is essentially inactive without calmodulin and calmodulin is inactive without calcium. When the influx of calcium ions raises their concentration in the cell above a certain threshold, four ions bind to each molecule of calmodulin. (In the case of some other enzymes the number of ions per calmodulin molecule may be smaller.)

On binding calcium the calmodulin molecule takes on a new, more compact shape and so becomes activated. More specifically, Daniel R. Storm of the University of Washington School of Pharmacology has found that a hydrophobic (water-repellent) region of the molecule is exposed and apparently interacts with the inactive phosphodiesterase (or other enzyme). The enzyme in turn takes on a new conformation and its catalytic activity is enhanced. Any excess of inflowing calcium is taken up by organelles or is pumped out of the cell. The two activation reactions are reversible. When the stimulation of the cell ends, the concentration of calcium ions falls to its steady-state level. The calmodulin molecule releases its ions and returns to its inactive shape, thereby dissociating the complex of calmodulin and phosphodiesterase. The enzyme is thereby rendered inactive and the calcium-initiated reaction ends.

Calmodulin has now been found in all organisms above the level of bacteria and in every kind of cell in such organisms; in other words, it is apparently a component of all eukaryotic (nucleated) cells. Calmodulins from phylogenetically diverse sources have identical or at least very similar biological and biochemical properties; even the sequence of the amino acids along the protein chain appears to have been highly conserved. The result is that the protein lacks both species specificity and tissue specificity: calmodulin from a unicellular protozoan will function in a test tube to activate an enzyme such as phosphodiesterase from the bovine brain.

Recognition of calmodulin's ubiquity and importance came only slowly, as it was found to be present in tissues other than the brain and in association with enzymes other than phosphodiesterase. In the early 1970's Thomas C. Vanaman of the Duke University Medical Center was studying a brain protein that seemed to be a version of troponin *c*, the calcium receptor in striated muscle. When my colleagues and I published the amino acid composition of calmodulin, Vanaman noticed that it was strikingly similar to the composition of his protein, and he suspected that the two molecules might be related. He sent me a sample of the troponinlike protein to see if it would stimulate phosphodiesterase. It worked about as well as our own calmodulin; indeed, Vanaman's protein was calmodulin. This finding and the results of other studies led to the gradual realization that earlier reports locating troponin *c* in the brain and in other tissues such as blood platelets were in error. Troponin *c* is probably found only in cardiac and skeletal muscle; the troponinlike protein in other tissues is calmodulin.

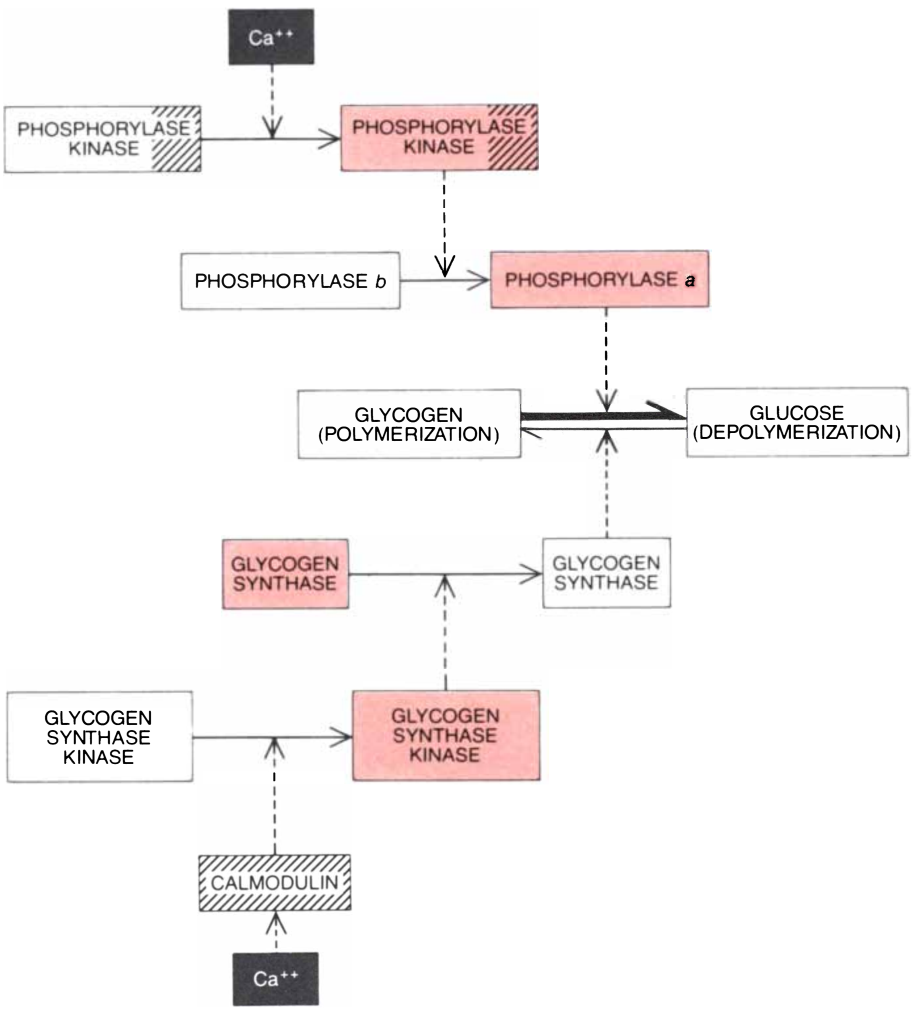
In my laboratory at St. Jude we found that the concentration of calmodulin in various tissues does not correspond to the distribution of phosphodiesterase. This raised the question of whether calmodulin has functions other than the activation of phosphodiesterase. Laurence S. Bradham of the University of Tennessee reported in 1972 that brain adenylate cyclase, the enzyme that catalyzes the synthesis of cyclic AMP, requires calcium for maximum activity. Charles O. Brostrom and Donald J. Wolff of the Rutgers Medical School and also Bradham in collaboration with our group went on to do experiments demonstrating that in the brain calmodulin activates adenylate cyclase as well as phosphodiesterase.

The regulation of the synthesizing enzyme and that of the breakdown enzyme by calmodulin seem to be nicely complementary. Adenylate cyclase is associated with the cell membrane, whereas the calcium-dependent phosphodiesterase is in the cytoplasm. As inflowing calcium passes through the outer membrane it binds to calmodulin molecules in the membrane and thus activates adenylate cyclase; the enzyme converts ATP into cyclic AMP. Somewhat later the calcium ions reach and bind to calmodulin molecules in the cytoplasm and thus activate phosphodiesterase. The sequential activation of the two enzymes may be responsible for the transient increase in intracellular cyclic AMP that is commonly observed when certain tissues are stimulated.

James D. Porter of the University of Cincinnati College of Medicine recently showed that the two enzymes differ in their sensitivity to calcium ions. This could mean that during the initial phase of the calcium influx rather low levels of the ion suffice to activate adenylate cyclase. As the calcium level rises the ion does two things: it activates phosphodiesterase and it also somehow inhibits adenylate cyclase, thereby quickly reducing cyclic AMP to its preactivation steady-state level.

The finding that calmodulin regulates a second enzyme in the brain still could not explain the high concentration of calmodulin there (far in excess of what is needed to regulate adenylate cyclase and phosphodiesterase), to say nothing of the presence of calmodulin in tissues lacking either of the enzymes. Soon the search was on for other functions. The list is still growing, and most of us in the field do not expect the pace of discovery to slacken for several years.

In 1973 Guy H. Bond of the Medical



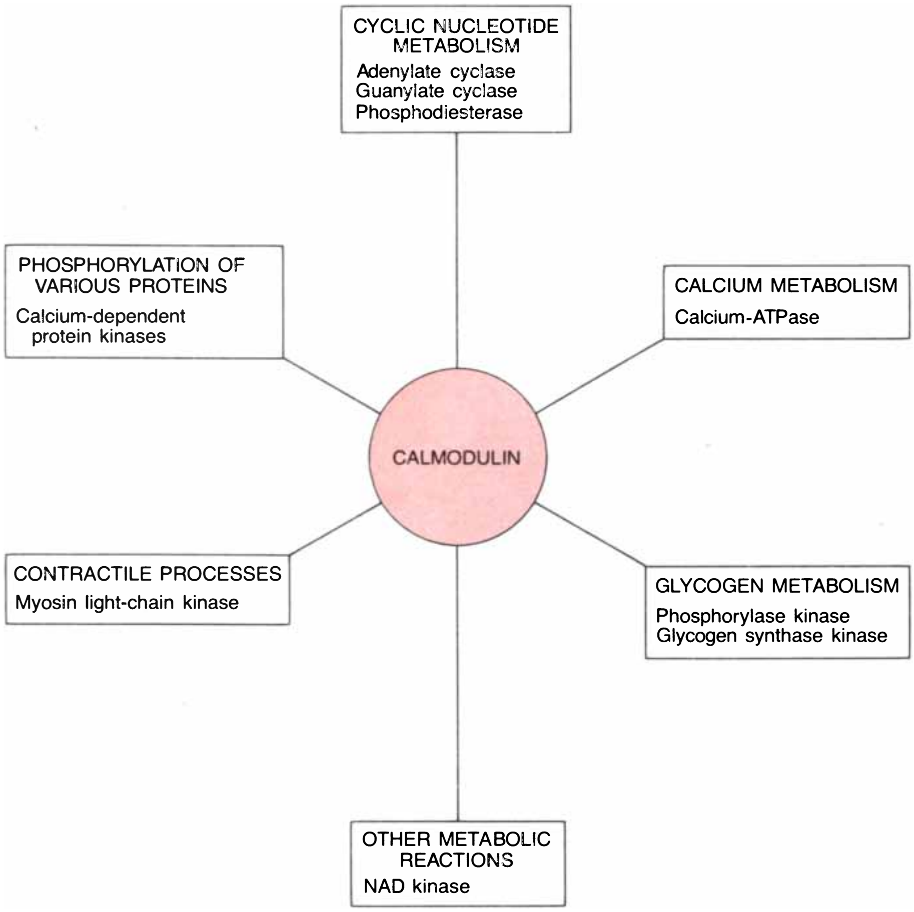

**GLYCOGEN METABOLISM in skeletal muscle is regulated by calcium ions, which bind to a calmodulin molecule (*hatching*) that is a subunit of the enzyme phosphorylase kinase. The activated kinase (*color*) adds phosphate groups to the inactive enzyme phosphorylase *b*, converting it into the active form, phosphorylase *a*. The activated phosphorylase initiates the breakdown of glycogen, which is a polymer of glucose, making glucose molecules available to supply the energy required for muscle contraction. Concurrently the polymerization of glucose to form glycogen is stopped. Calcium ions bind to calmodulin, which activates the enzyme glycogen synthase kinase. The kinase adds phosphate groups to glycogen synthase. In this case the phosphorylation inactivates the enzyme, which would otherwise initiate the polymerization of glucose and thereby make it unavailable as a source of energy for muscle contraction.**

College of Virginia found that in human red blood cells there is a factor that stimulates calcium-ATPase, the enzymatic pump in the cell membrane that moves calcium out of the cell. Frank F. Vincenzi of the University of Washington School of Medicine and John T. Penniston of the Mayo Clinic identified the pump activator as calmodulin, which has since been found to stimulate calcium-ATPase in other cells as well. Calmodulin, then, has a double set of functions: it not only transmits calcium's message to receptor enzymes but also modulates the intracellular concentration of the ion (something I had in mind when I coined the name calmodulin). Having sensed the arrival of calcium and thereby been activated to stimulate an enzyme, calmodulin proceeds to turn on the pump that rids the cell of unneeded calcium.

As I have mentioned, the first calcium receptor to be identified was troponin *c* in striated muscle. The arrival of a nerve impulse at the muscle releases calcium from storage in the part of the muscle cell called the sarcoplasmic reticulum. The ion binds to troponin *c* and alters the shape of the troponin molecule, initiating a series of interactions among muscle proteins that eventually catalyzes the hydrolysis of ATP to release energy for muscle contraction. In the contraction of smooth muscle and of filaments in nonmuscle cells, however, calcium's effect is mediated by calmodulin rather than troponin. David J. Hartshorne of the University of Arizona and others have shown that calmodulin, having bound calcium ions, activates a protein kinase: an enzyme that phosphorylates (adds a phosphate group to) another enzyme or some other protein. The kinase activated by calmodulin in smooth muscle phosphorylates a regulatory component of the protein myosin, thereby initiating interactions that presumably are similar to those controlled by troponin *c* in striated muscle.

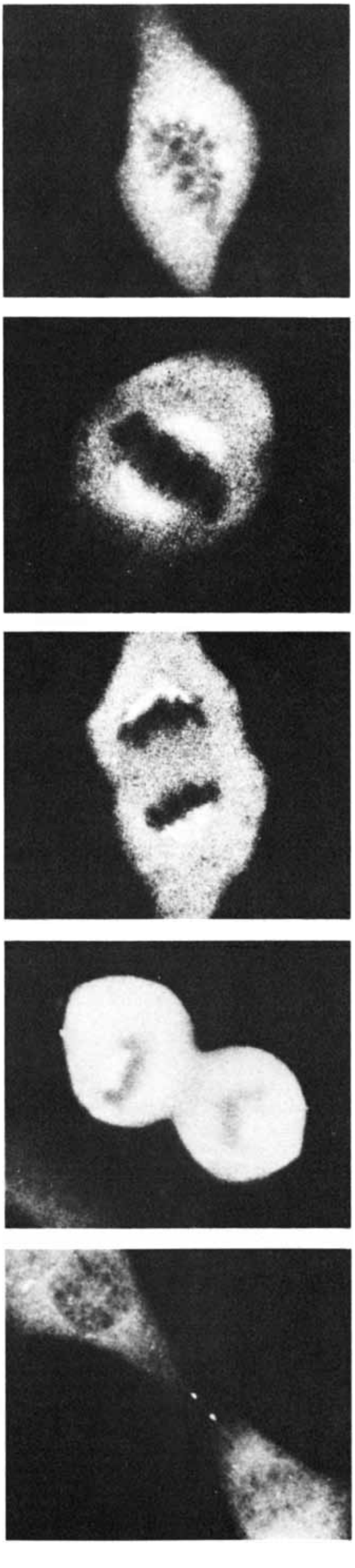
Muscle contraction, which consumes energy, is metabolically coordinated with a sequence of events that supplies energy. Glycogen, the storage form of glucose, is the main source of quickly available energy, and large reserves of glycogen are close at hand in muscle cells. Because glycogen is a polymer of glucose (a long chain of linked glucose molecules), a necessary step toward supplying energy is the depolymerization of glycogen. When striated muscle is stimulated, calcium ions released from the sarcoplasmic reticulum trigger not only contraction but also the depolymerization of glycogen. They do so by binding not to free calmodulin molecules but to a calmodulin unit that is itself a subunit of the enzyme being activated: a kinase whose structure has been worked out by Philip Cohen of the University of Dundee. The activated kinase adds phosphate groups to the enzyme phosphorylase, converting it from the inactive *b* form into the active *a* form. Phosphorylase *a* initiates a breakdown of glycogen, making glucose available to generate ATP, the universal energy transducer in cells.

There is more to the story. Muscle cells have an enzyme for storing glucose as well as one for bringing it out of storage. The storage enzyme, glycogen synthase, initiates a reaction that links glucose units to form glycogen, and its polymerizing activity needs to be turned off when glucose is required by the muscle cell. In this case phosphorylation turns the enzyme off rather than on. Thomas R. Soderling of the Vanderbilt University School of Medicine recently showed that calmodulin activates glycogen synthase kinase, an enzyme that phosphorylates and thereby inactivates glycogen synthase. Thus calcium and calmodulin coordinate the suppression of glycogen synthesis with the stimulation of glycogen breakdown. Both actions are linked by calcium (by way of calmodulin) to the contractile event they serve.




**VARIOUS FUNCTIONS OF CALMODULIN are summarized, together with some of the enzymes known to serve each function. Adenylate cyclase and guanylate cyclase catalyze the synthesis of cyclic AMP and cyclic GMP respectively; phosphodiesterase breaks down both cyclic molecules. Calcium-ATPase is the molecular pump that expels calcium ions from the cell. Phosphorylase kinase and glycogen synthase kinase respectively control the enzymes that initiate the breakdown and the synthesis of glycogen. NAD kinase synthesizes NADP from the cofactor NAD. Myosin light-chain kinase controls contraction in smooth-muscle and nonmuscle cells. Finally, there appear to be other calcium-dependent protein kinases that phosphorylate various proteins, including enzymes involved in the synthesis of neurotransmitters.**

The phosphorylation of a protein by a kinase is often a key step in a regulatory process. There is evidence that in addition to the three protein kinases implicated in smooth-muscle contraction and in glycogen metabolism there may be other calmodulin-dependent kinases. Paul Greengard of the Yale University School of Medicine has detected one such kinase that seems to phosphorylate a number of different proteins depending on the tissue in which it is active. Robert DeLorenzo of Yale has found in the membrane of nerve-cell terminals a calmodulin-dependent kinase that phosphorylates the protein tubulin. The phosphorylation changes the physical and chemical properties of the tubulin, which then aggregates to form filamentous structures. (The filaments do not



**CALMODULIN IS CONCENTRATED in the mitotic spindle, the apparatus that separates a cell's two sets of chromosomes during cell division. A dividing mouse fibroblast (a connective-tissue cell) is seen in photomicrographs made by John R. Dedman, B. R. Brinkley and Anthony R. Means at the Baylor College of Medicine. The cell was treated with a sheep antibody to calmodulin. A second antibody was added, this one directed against the sheep antibody and labeled with a fluorescent dye. The bright glow indicates that calmodulin is present throughout cell but is concentrated in spindle. Micrographs show successive phases of cell division (*from top*): prophase, metaphase, anaphase, telophase and finally the separation of daughter cells.**

seem to be microtubules, the cablelike structures usually formed by tubulin.) DeLorenzo thinks the filaments may interact with the membrane to facilitate the release of norepinephrine. Norepinephrine is a neurotransmitter, one of the substances that relay a nerve impulse from one nerve cell to another one or to a muscle cell.

Calmodulin may also have a role in the synthesis of neurotransmitters. The catecholamine transmitters dopamine, norepinephrine and epinephrine are synthesized from the amino acid tyrosine in several steps, the first of which is catalyzed by the enzyme tyrosine 3-monooxygenase. Hitoshi Fujisawa of the Asahikawa Medical College has shown that the enzyme's activity depends on a phosphorylation by a calmodulin-dependent kinase. Tryptophan 5-monooxygenase, which catalyzes the conversion of the amino acid tryptophan into the neurotransmitter serotonin, is similarly controlled by a calmodulin-dependent phosphorylation.

The cofactor called nicotinamide adenine dinucleotide phosphate (NADP) is required for the synthesis of many cellular constituents, including steroids, nucleotides (the subunits of nucleic acids) and fatty acids. NADP can be synthesized from nicotinamide adenine dinucleotide (NAD) by a specific kinase, which Milton J. Cormier of the University of Georgia has found is dependent on calmodulin. NAD and NADP are required for many cellular activities during the early development of the embryo. Calcium is known to be important for fertilization. David Epel of the Hopkins Marine Station of Stanford University and Robert W. Wallace and I have shown that an increase in the internal calcium level soon after fertilization of the sea-urchin egg leads to a transient activation of the NAD kinase by calmodulin. The result is increased synthesis of NADP, which could contribute to metabolic processes responsible for the successive cell divisions that produce the early embryo.

Cyclic guanosine monophosphate (cyclic GMP) is a nucleotide similar to cyclic AMP; like the latter molecule it is thought to serve as an intracellular messenger, although its functions are not well understood. Recently Yoshio Watanabe of the University of Tsukuba has reported that in the protozoan *Tetrahymena* calmodulin activates guanylate cyclase, the enzyme that catalyzes the synthesis of cyclic GMP. Whereas calmodulin from any species or tissue usually functions in other tissues or species, the protozoan guanylate cyclase seems to be activated only by its own calmodulin. Mammalian guanylate cyclase is known to require calcium for maximum activity, but whether the ion's effect on the mammalian enzyme is mediated by calmodulin remains to be shown.

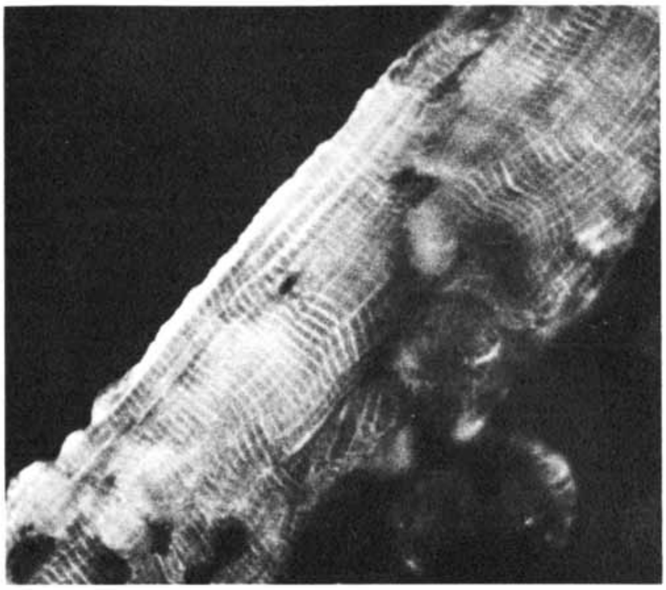
I have discussed some of the processes in which calmodulin regulates a specific enzyme whose function is known. The protein undoubtedly has other activities for which there is so far only suggestive evidence. It has been found in many tissues where its mode of action is not yet known. In other cases proteins have been identified that clearly bind to calmodulin but whose own function is not yet established.

One such protein is calcineurin, which is present, along with calmodulin, in many regions of the mouse brain. Calcineurin has two subunits, one of them about the size of calmodulin and the other larger. Claude B. Klee of the National Cancer Institute has found that the small subunit, like calmodulin, binds four calcium ions and that the large subunit binds to calmodulin. Calcineurin's function is not fully understood, but preliminary results from Cohen's laboratory and mine suggest it is a protein phosphatase: an enzyme that catalyzes the removal of a phosphate group from proteins. It is concentrated at postsynaptic densities, near where the receptors for neurotransmitters are situated, and my graduate student E. Ann Tallant has shown that its level increases with the formation of synapses between nerve cells during development. These findings suggest that calcineurin may have some role in neurotransmission.

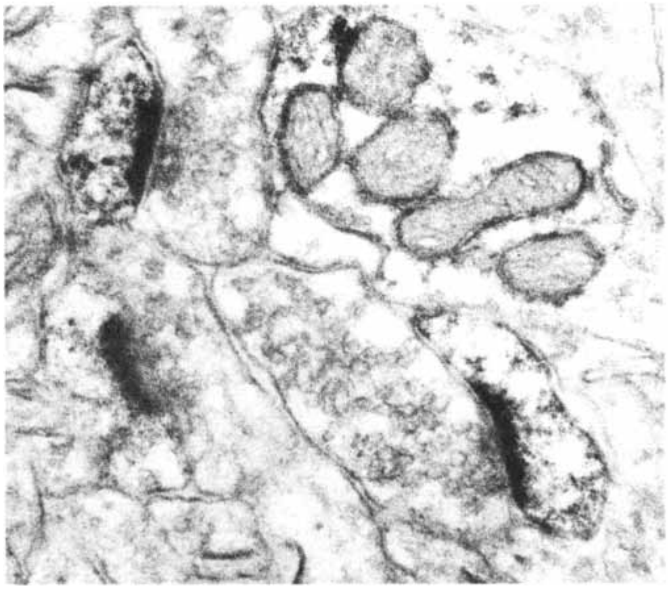
The metabolism and functions of calcium and cyclic AMP are intertwined. I have explained how in the brain the synthesis and the degradation of cyclic AMP are both regulated by calcium ions by way of calmodulin; the same may be true in some other tissues. Cyclic AMP promotes the uptake of calcium into organelles such as the sarcoplasmic reticulum, a process tending to turn off a calcium-initiated action. There are instances where calcium and cyclic AMP act in opposition. For example, in smooth muscle they seem to have opposite effects on myosin light-chain kinase, the enzyme that initiates contraction. Calcium-activated calmodulin stimulates the enzyme; Robert S.

Adelstein of the National Heart, Lung, and Blood Institute has preliminary evidence suggesting that cyclic AMP makes the enzyme refractory to such stimulation, lessening calcium's effect. In still other cases calcium has a function in one tissue that is performed in another tissue by cyclic AMP. The stimulation of phosphorylase to degrade glycogen is generally accomplished by calcium and calmodulin in muscle and by cyclic AMP in the liver.

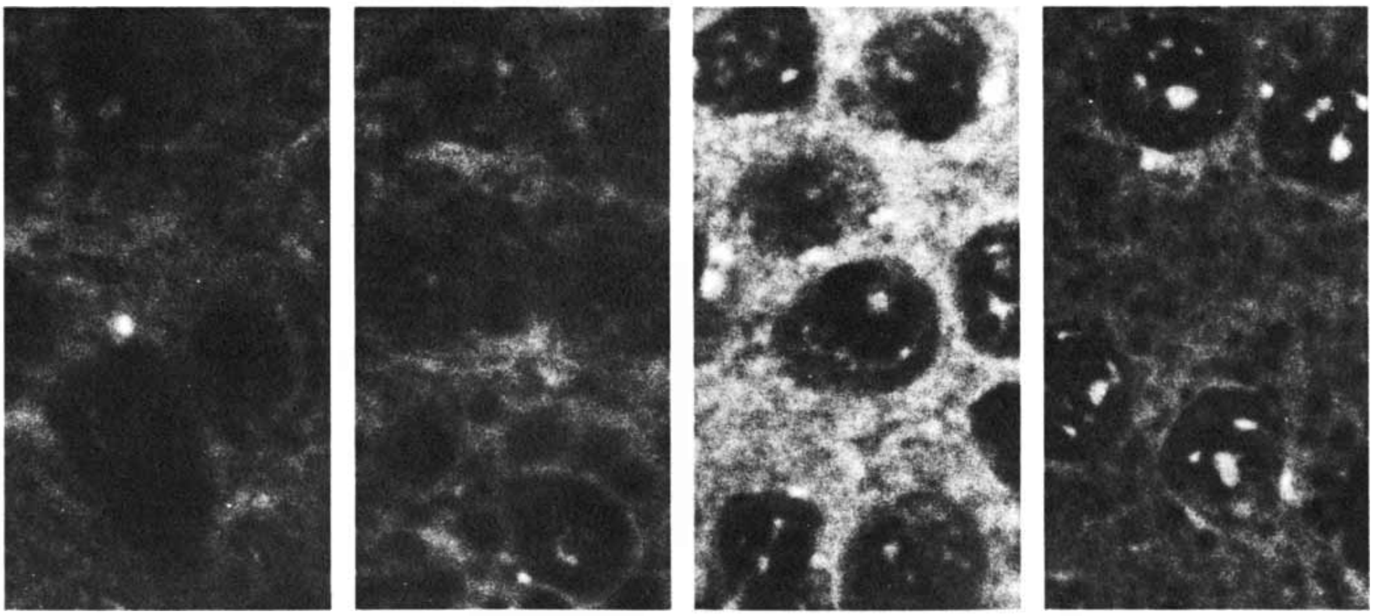
The calcium-calmodulin system and the cyclic-AMP system have different attributes, which may explain why the systems are active under different circumstances. Calmodulin is present in every cell and calcium is plentiful in the extracellular fluid; neither substance needs to be synthesized to initiate an action, and so the calcium system is inherently fast-acting. Cyclic AMP, on the other hand, is essentially synthesized de novo by adenylate cyclase when the lat-



**IN SKELETAL MUSCLE** calmodulin is apparently associated with one band (thought to be the one designated *I*) in the repeated pattern of transverse striations characteristic of such muscle cells. Calmodulin is also found in a longitudinal structure, probably the sarcoplasmic reticulum. In this micrograph, made by Jeffrey F. Harper and Alton L. Steiner of the University of Texas Health Science Center at Houston in collaboration with the author's group, a segment of the gastrocnemius muscle in the leg of a rat is enlarged 800 diameters. Again the location of antibody that binds to calmodulin is revealed by the binding of a second antibody labeled with a fluorescent dye.



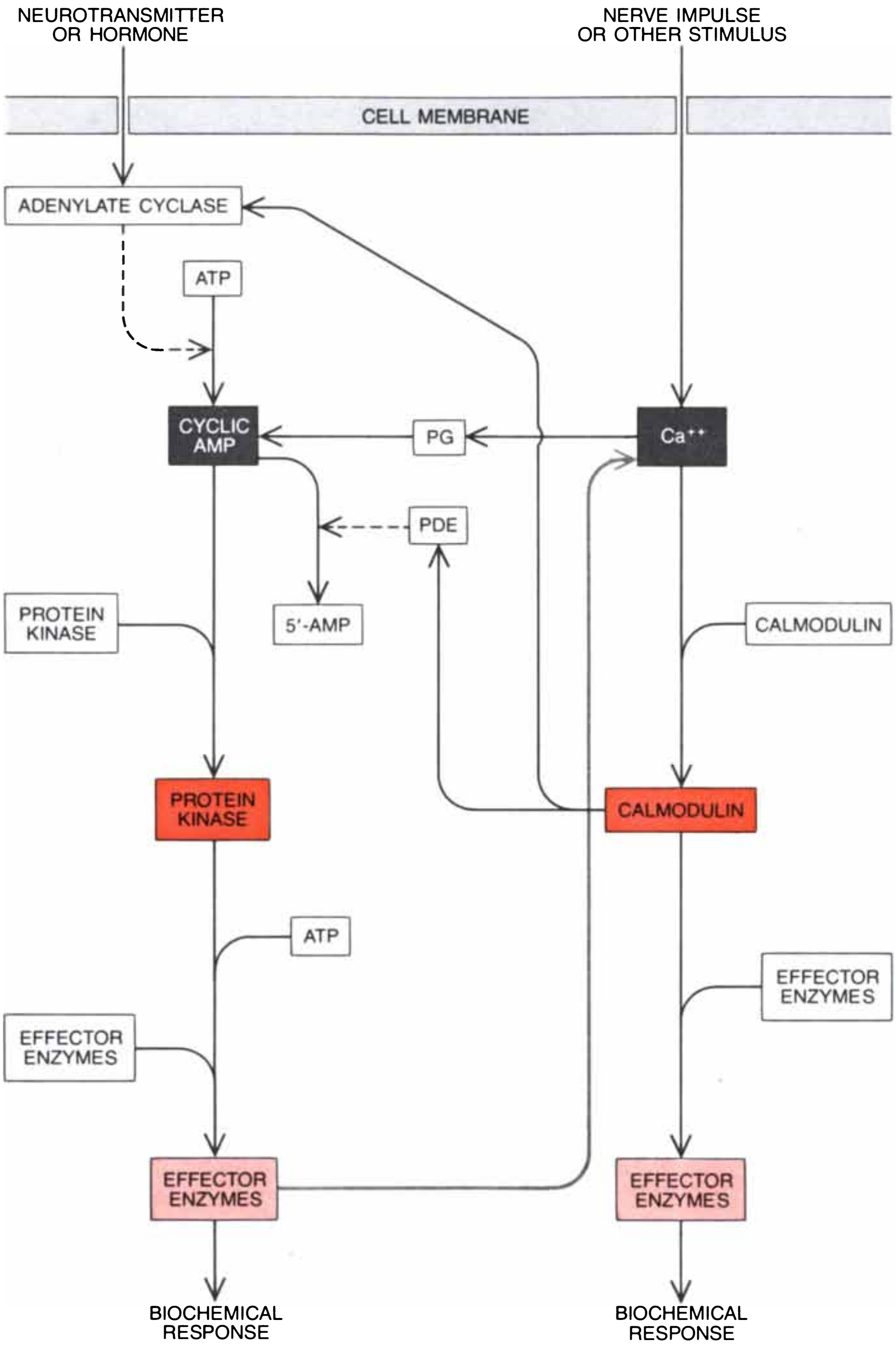
**IN THE BRAIN** calmodulin seems to be found near the postsynaptic region of a nerve cell, where impulses transmitted from another cell are received. In this electron micrograph, which was made by John G. Wood of the Emory University School of Medicine in collaboration with the author's group, a thin section from the region of the rat brain called the basal ganglia is enlarged some 70,000 diameters. The section was incubated first with a rabbit antibody to calmodulin and then with an antibody, labeled with the enzyme peroxidase, to the rabbit antibody. A stain precipitated by the peroxidase shows that calmodulin has become concentrated at the so-called postsynaptic densities.



**HORMONAL STIMULATION** of the rat adrenal cortex seems to cause calmodulin to become associated with the nuclei of cortical cells. The photomicrographs, made by Harper and Steiner with the author's group, show cortical tissue from adrenal glands removed before the injection of corticotropin (ACTH), which stimulates the cortex to synthesize steroid hormones, and at intervals after the injection. In unstimulated cells (*left*) some calmodulin is revealed by immunofluorescence in the cytoplasm. After half an hour (*second from left*) the nuclei show some fluorescence. After an hour (*third from left*) the concentration of calmodulin seems to have increased. After 11 hours (*right*) most of the calmodulin is associated with nuclei, suggesting that calmodulin has a role in some nuclear function of ACTH.

ter is stimulated, and that takes a certain amount of time.

The most important receptor for cyclic AMP in eukaryotic cells is a protein kinase that is activated by cyclic AMP and can thereupon phosphorylate a number of different substrate proteins. The calcium system, in contrast, has at least three different calmodulin-dependent kinases that phosphorylate specific substrates, and it probably also has a number of other calmodulin-dependent kinases that phosphorylate various enzymes. Moreover, calcium, through calmodulin, regulates many other enzymes in addition to protein kinases. Because calmodulin is flexible it can interact with a variety of effector enzymes, making the calcium ion much more versatile than cyclic AMP.




**INTERRELATION of two intracellular-messenger systems is shown by this diagram. The cyclic-AMP system is at the left, the calcium-calmodulin system at the right. The arrival of an intercellular message activates adenylate cyclase, which synthesizes cyclic AMP; this intracellular messenger activates a protein kinase, which phosphorylates some effector protein, either stimulating or inhibiting its biological activity. Similarly, calmodulin, when it is activated by calcium ions, activates an effector protein to initiate a biological response. Calcium stimulates the activity of adenylate cyclase and also that of phosphodiesterase (*PDE*), which breaks down cyclic AMP. Calcium modulates the metabolism of the hormonelike fatty acid prostaglandin (*PG*), which in turn influences adenylate cyclase metabolism. Cyclic AMP may have an effect on the availability in the cell of calcium ions, presumably by activating a phosphorylation.**

In mammals the important cellular messengers are hormones, cyclic AMP (and other cyclic nucleotides) and calcium. For the most part hormones make it possible for cells to communicate with one another; cyclic AMP and calcium carry messages from one part of a cell to another. The three messengers have complementary roles with respect to time as well as distance. For hormones the response time and the duration of action range from minutes to hours and even weeks. For cyclic AMP the range is from seconds to minutes and for calcium it is probably in the millisecond range. Hormones are often designated the "first messengers" and cyclic AMP a "second messenger." Considering the extent to which the metabolism and functions of the calcium ion and cyclic AMP are interconnected and therefore overlap in time and space, it is probably more accurate to refer to both of them simply as cellular messengers or regulators, rather than attempting to order them as being second or third messengers in any particular process.

The discovery of calmodulin has greatly clarified how calcium acts at the molecular level. In elucidating the various roles of calmodulin an agent that counteracts the activity of the protein has been particularly helpful. Benjamin Weiss of the Medical College of Pennsylvania reported in the mid-1970's that trifluoperazine, an antipsychotic agent, inhibits the activity of calmodulin by preventing the protein's interaction with receptor enzymes. The drug has aided immensely in the identification of some of the biological functions of calmodulin. Moreover, it has stimulated efforts to design and identify drugs that affect either the activities of calcium mediated by calmodulin or other aspects of calcium's action and metabolism.

How is it that calmodulin, the ever-present participant in calcium-initiated cellular processes, remained hidden from investigators through several decades during which those processes were under intensive study? The answer lies primarily in calmodulin's ubiquity and abundance. Unlike cyclic AMP, for example, calmodulin seems not to be the limiting factor in a cellular activity; since it is always present, both in cells and in cell extracts, it is never missed. Its role was disclosed only by its inadvertent removal from phosphodiesterase during the routine purification of the enzyme. The loss of enzyme activity in the course of purification is usually just an annoyance. In the case of phosphodiesterase I was fortunate in that the discrepancy between the results of two assay procedures pointed to the nature of the problem: the removal of an activator, which turned out to be calmodulin.

# Anyone who'd call Wagoneer a station wagon has obviously never driven one.

Any similarity between the Jeep Wagoneer Limited and a conventional two-wheel drive wagon is purely coincidental.

Because no conventional full-size wagon gives you all the security, luxury and economy of Wagoneer.

For security, it's the Wagoneer's sure-footed four-wheel drive, performing deftly under all kinds of pressure...snow buried highways, icy roadways, torrential downpours. Or just digging in off-road to take you far from the maddening crowds.

For luxury, it's an interior filled with supple leathers, thick carpeting, and a wealth of comfortable appointments you would expect to find only in the plushest automobiles.

And for economy, it's better EPA estimated mileage than *any* full-size two-wheel drive wagon.*

Luxury, economy *and* the security of four-wheel drive. With all this, and surprisingly affordable prices, why would anyone drive anything else?

The Jeep Wagoneer Limited. We call it the Ultimate Wagon. You'll call it *beautiful*.

18 EST* EPA MPG 25 HWY EST

**Jeep® Wagoneer Limited. The Ultimate Wagon.**

AT AMERICAN MOTORS

*Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Actual highway mileage lower.
Jeep Corporation, a subsidiary of American Motors Corporation.

# SCIENCE AND THE CITIZEN

## *On the Road*

In the past three decades the U.S. has spent a total of $2.6 billion on various programs intended to protect the country's population and industry against a nuclear attack by the U.S.S.R. The peak year for such expenditures was 1962, when the Kennedy Administration's accelerated shelter-building program pushed the annual outlay for civil defense to more than $200 million. By the mid-1960's that program had come to be widely perceived as ineffectual, and it was abandoned. Since then civil-defense spending has been maintained at a much lower level, presumably in recognition of the futility of efforts to lessen the destructiveness of nuclear war.

President Reagan proposes to change all that. For the current fiscal year the civil-defense budget is $133 million. For next year the Administration has requested more than $252 million for civil defense, almost doubling the current rate of spending. Moreover, the proposed increment represents only the first stage of a seven-year program designed to "provide for survival of a substantial portion of the population in the event of a nuclear attack preceded by strategic warning." According to the official budget projections released by the Administration, the president's "enhanced" civil-defense program will cost a total of $4.2 billion by the end of 1989. According to an unofficial estimate by staff members of the Office of Management and Budget, the ultimate cost of the program could exceed $10 billion.

The emphasis of the new National Civil Defense Program is on evacuation of the cities rather than on shelter building. A press release describing the program states: "Primary reliance will be placed upon relocating the population of U.S. metropolitan and other potential high-risk areas to surrounding areas of lower risk during a period of international crisis, taking advantage of extensive U.S. transportation resources." People who reach the designated rural relocation areas, it is said, "will be least affected by the blast and thermal effects of the nuclear explosions, leaving radiation as the major hazard for the dispersed population."

The responsibility for overseeing the new program has been assigned to the Federal Emergency Management Agency (FEMA), the successor to the Defense Civil Preparedness Agency. In the view of Louis O. Giuffrida, the director of FEMA, "the outlook for Civil Defense is good." Studies cited by his agency are said to indicate that a "civil defense program emphasizing crisis relocation might save—in a large-scale attack preceded by strategic warning—up to twice as many Americans as the 40 percent expected to survive under present civil defense."

Critics of the new program are not as sanguine. They contend that the present generation of civil-defense planners, like their predecessors in the 1950's and the early 1960's, grossly underestimate the potential destructiveness of nuclear war. The critics have also questioned both the practicality and the effectiveness of the evacuation plan. It is not clear that urban populations could be moved in the time available or supported in the relocation areas. Furthermore, the aims of the program could easily be thwarted: an attacker could simply retarget the nuclear warheads to fall on the relocation areas. Worse than that, a potential opponent might perceive that a civil-defense program has some potential effectiveness and build enough additional nuclear weapons to overwhelm it, thereby increasing the potential destructiveness of the nuclear attack. Evacuation itself might be seen as a threatening act, preparatory to war, and hence might invite a preemptive attack.

Senator Alan Cranston of California has called the proposed civil-defense program "a total waste of money" and "a cruel delusion if it leads anyone to believe nuclear war is survivable." Cranston asks: "Is it possible that the Reagan Administration—or at least some people within the Administration—are trying to prepare the American public for not just the acceptance of nuclear war but perhaps its desirability?"

The strongest rebuttal to the new plan may take the form of noncompliance. State, county and local officials in several parts of the country have declared their intention not to take part in preparing the crisis-relocation plans.

## *The Mostest*

The population of the People's Republic of China is without question the largest of any nation, constituting roughly a fifth of the population of the world. The size of the Chinese population, however, is not known with precision. Estimates made by demographers differ by about 100 million, equal to almost half of the U.S. population. Some of the uncertainty will soon be resolved. The third census of China since the founding of the People's Republic will begin on July 1.

Several factors have contributed to the scarcity of reliable information on China's population. They include the pressing needs of economic development, the size of the population and political perturbations that have entailed dramatic changes in the official attitude toward demographic matters. Much of the interest of demographers in the census is focused on the effectiveness of recent government efforts to reduce the birth rate.

The first census of China after the Communist Party took power was made in 1953; the population was then about 582 million. The second census, in 1964, recorded about 691 million people. Among the lowest estimates of the current population is one made by workers at the United Nations: 980 million people. Workers at the U.S. Bureau of the Census estimate that the population is now 1,073 million. Estimates by demographers at the World Bank and at universities in China and elsewhere lie between these extremes.

The birth rate now prevailing in the People's Republic is of as much interest as the size of the population. After several sharp fluctuations in policy the reduction of fertility was made an official goal in 1971. In 1979 the effort was intensified, and since then the government has attempted to persuade every Chinese couple to have at most one child. The campaign appears to have had considerable success. The annual number of births per 1,000 population is estimated to have decreased from 34 in 1970 to 18 in 1979.

If the estimated birth rate is accurate, the reduction is the fastest ever recorded in a large developing country. Indeed, it is possible the birth rate is falling fast enough to cause economic hardship. A possible consequence of a rapid decrease in fertility is a subsequent increase in the ratio of retired people to working people. The ratio rises when the large group born just before the decrease in the birth rate reaches retirement age. It appears that China's political leaders are now debating the optimum pace of birth-rate reduction.

The speed with which the birth rate falls will have a substantial effect on the size of the future population. On the basis of various estimates of the current population and birth rate a group of Chinese demographers has estimated that the population in the year 2000 will be 1,125 million; a group at the Bureau of the Census has estimated that it will be 1,380 million.

In the census that is about to start more questions will be asked of each respondent than were asked in earlier censuses. Every person will be questioned as to name, age, sex, ethnic group, occupation and education, and a sample of the population will be asked their place of birth, type of housing and marital status. Certain questions to be asked at the option of local administrators bear directly on the success of the family-planning program. They concern the number of surviving children and the number, birth order and sex of all children born in the previous year.

The results of past Chinese censuses have been tabulated by abacus. This time the government will have available more elaborate information-processing equipment. It has bought eight electronic digital computers, and 21 more computers have been provided by the UN. One computer will be installed in each province to tabulate census returns beginning in September. The results are to be published in 1984. The scope of the census is suggested by the fact that there are more than six million enumerators employed in counting the people of China.

### *Prion*

Scrapie is a fatal degenerative disease of the brain in sheep and goats. (It causes intense itching, so that the sick animals tend to scrape their wool off; hence the name.) The disease has long been known to be transmissible, but its agent has never been seen in an electron micrograph or characterized chemically. Because the agent is clearly not a bacterium it has generally been assumed to be a virus. When it is inoculated into an animal, the agent replicates, but only slowly, and it takes many weeks for the symptoms to appear. The disease is therefore commonly blamed on an "unconventional slow virus."

The scrapie agent is unconventional indeed, according to a report published in *Science* by Stanley B. Prusiner of the School of Medicine of the University of California at San Francisco; it seems not even to be a virus. For one thing, it is at least 100 times smaller than any known virus. Another property of the agent is still more extraordinary. Most viruses consist of a protein coat encapsulating several genes' worth of a nucleic acid (either DNA or RNA). The scrapie agent seems to consist only of protein; as yet there is no sign of any nucleic acid. Because the scrapie agent is proteinaceous and infectious, Prusiner calls it a prion (pronounced pree-on).

Prusiner's conclusions are based on a partial purification of the scrapie agent. Purification has been complicated because the agent is so small, because it cannot be grown in cultured cells and because it is nonantigenic: it does not give rise to antibodies, which could combine with the agent and thus reveal its presence. The only way to detect the agent is by means of tedious animal assays. At each stage of purification the agent must be injected into the brain of experimental animals, where it replicates. The length of time to the onset of illness and then to death is a measure of the concentration of the agent.

Prusiner and a number of his colleagues have been at the task of purification for several years. They begin with brain tissue from infected hamsters. They digest the tissue with enzymes, treat the digested material with detergents and separate the agent from most of the remaining cellular debris by the technique called gel electrophoresis. The most highly infectious preparation is still not pure scrapie agent, but it is pure enough for tests of the nature of the agent to be conducted.

The investigators subjected the partially purified agent to six different treatments that inactivate proteins: breakdown by enzymes that digest proteins, chemical modification and denaturation (unfolding of the protein chain) by specific detergents, by specific ions, by phenol and by urea. In every case the treatment destroyed the agent's ability to infect hamsters. Thus protein is required for infection. That does not mean the agent cannot be a virus, but it does rule out several other possibilities. It had been suggested, for example, that the agent might be a viroid, one of a class of infectious agents recently discovered in plants. Viroids are short strands of RNA without a protein coat, and so treatments that inactivate proteins would not inactivate a viroid.

Prusiner and his colleagues went on to test for the presence of nucleic acids. They subjected the scrapie agent to six kinds of treatment calculated to attack DNA or RNA or both, including digestion with specific enzymes and exposure to doses of ultraviolet radiation far greater than those required to inactivate the nucleic acids of a virus. In every case the agent remained infectious; it was inactivated by ultraviolet radiation only when the dose was large enough to affect a protein. Some viruses are resistant to inactivation by one or two of the anti-nucleic-acid treatments; none is resistant to all of them. The scrapie agent must therefore have a structure different from that of any known virus.

The failure to find evidence of nucleic acid is not proof that none is present. It is possible that some DNA or RNA is buried deep within a highly protective protein shell. It is also possible that the agent incorporates a hitherto unknown kind of nucleic acid that resists inactivation. Those possibilities are virtually excluded, however, by the small size of the scrapie agent, which seems to have a molecular weight of no more than 50,000. This is less than a hundredth the molecular weight of the smallest viruses known; it is about a third as large as a viroid, until now the smallest infectious agent known; it is smaller than a single molecule of the protein albumin. If the scrapie agent were constructed like a virus, its internal volume would be large enough to house only 12 nucleotides (subunits of DNA and RNA), which could code for no more than four amino acids (subunits of proteins). If the agent does harbor any nucleic acid, it has too little to direct the synthesis of its own coat protein.

If the scrapie agent has so little genetic material or (as seems more likely) none at all, how does it reproduce itself? If it does have a bit of nucleic acid, the short stretch of nucleotides might possibly act as a regulatory agent. What might it regulate? Perhaps there are genes in animal cells that under certain circumstances can direct the synthesis of a prion; perhaps the infecting prion's nucleic acid somehow turns those genes on or regulates their expression by altering the splicing of the messenger RNA that intervenes between a gene and a protein.

What if it develops that there is indeed no nucleic acid at all in a prion? Then the possibility that the prion is an infectious protein will have to be considered. The suggestion is "clearly heretical," Prusiner concedes. Yet it is conceivable that the prion's protein acts as a regulator that binds to the "promoter" region of cellular genes and thereby activates the genes. Most heretical of all, the prion protein might somehow code for its own replication. Presumably it could do so only by itself directing the synthesis of DNA or RNA coding for it ("reverse translation") or by directly specifying its own synthesis. Both reverse translation and protein-directed protein synthesis would violate the "central dogma" of molecular biology, which holds that genetic information flows from nucleic acids to protein, not from protein to nucleic acids or from protein to protein.

## *High-Tech Detectors*

Among the problems that confront physicists studying the collisions of elementary particles are identifying particles in the debris and measuring their exceedingly short lifetimes. Two main approaches have evolved. In an electronic "counter" the passage of a particle gives rise to ions or to quanta of electromagnetic radiation, which are then detected. In a bubble chamber a particle moving through a liquid medium leaves behind a trail of small bubbles, which are photographed. Improvements to devices of both kinds are now being tested. They are based on semiconductor technology for the counters and on holography for the bubble chambers.

Solid-state detectors based on semiconductor technology offer much higher sensitivity and resolution than detectors based on the movement of particles through a liquid or a gas. The sensitivity results from the ease with which electrons are freed in a semiconductive material such as silicon. For a given deposit of energy 10 times as many electrons are released in silicon as are released in a gas. High spatial resolution is possible because methods of fabrication developed for making microelectronic devices allow the individual elements of the detector to be very small and densely packed.

One high-resolution silicon counter was made by workers from the European Organization for Nuclear Research (CERN), the Max Planck Institute for Physics in Munich and the Technical University of Munich. Some 1,200 diode strips are mounted on a silicon wafer 300 micrometers thick. Each strip is 36 millimeters long and is separated from adjacent strips by 20 micrometers, creating a sensitive area of 24 by 36 millimeters. An electrically charged particle traversing the silicon dislodges electrons in the crystal lattice, providing a signal that is subsequently amplified and recorded.

In a test at CERN three silicon counters installed in the beam of a large proton synchrotron provided a spatial resolution of 10 micrometers and made it possible to distinguish particles separated by as little as 100 micrometers. More recently silicon detectors were employed at CERN to measure the lifetime of the *D* meson, a rare particle that carries the property called charm. The lifetime is $9.5 \times 10^{-13}$ second, or about a millionth of a millionth of a second.

A workshop on silicon detectors was held last fall at the Fermi National Accelerator Laboratory. The detectors are fairly expensive and are able to scrutinize only a small area. It therefore seems likely they will supplement rather than supplant liquid and gas detectors.

Workers in several laboratories are investigating the potential of laser holography for improving the resolution of measurements made with bubble chambers. Ordinarily the paths of particles or their decay products in a bubble chamber are recorded in two-dimensional photographs; in the course of an experiment it is not unusual to make several hundred thousand exposures. Holography gives a three-dimensional image and thereby increases the depth of field, that is, the volume that is in focus. A resolution of from five to 10 micrometers could be achieved with a holographic system. Furthermore, the information recorded in several two-dimensional photographs made from various points of view could be captured in a single hologram.

## *Help Wanted (Engineers)*

Schools of engineering are presumably gratified to find that the employment prospects of their graduates are currently quite good. Ironically, however, the strong demand for engineers is now perceived as a threat to the quality of engineering education. The problem is a notable illustration of the complexity inherent in the workings of simple economic laws.

Greater demand can be expected to raise the salaries and other compensation offered to engineers and thereby to attract more people to the profession. This much has happened: since 1975 enrollment in undergraduate engineering programs has increased by more than 50 percent. It might seem the supply would soon catch up with the demand and equilibrium would be attained. What complicates the situation is that the faculty of engineering schools respond to the same economic forces as their students. Universities are therefore competing with industry in the hiring of engineers, and it appears the universities are losing the competition. A survey made by the American Council on Education has found that more than 1,600 engineering faculty positions are now unfilled, for a vacancy rate of about 10 percent. In response to the shortage of instructors a number of universities have put a limit on enrollment in engineering, a measure that could ultimately lead to further increases in demand and make it still harder to recruit and keep faculty.

In 1980 the median starting salary for an engineer with a bachelor's degree was just under $20,000. A student who continued his education for four more years, earning a doctoral degree, and then took a position as an assistant professor of engineering could expect to earn slightly less. Evidently many students have recognized the economic disincentive to further education and academic employment. Although the number of engineering bachelor's degrees granted has been increasing for some years, the number of doctoral degrees has declined sharply.

In April a conference on the imbalances in engineering education was organized by Paul E. Gray of the Massachusetts Institute of Technology and chaired by Edward E. David, Jr., of the Exxon Research and Engineering Company. Statements issued during the conference suggest that the only practical way to make academic salaries competitive with industrial ones is to pay engineering faculty more than instructors of equal rank in other disciplines. The regents of the University of California have recently approved such a plan, establishing differentials in salary similar to those that have been adopted in some schools of medicine and law. Nonmonetary issues, and in particular the improvement of laboratory facilities and apparatus, are also considered urgent.

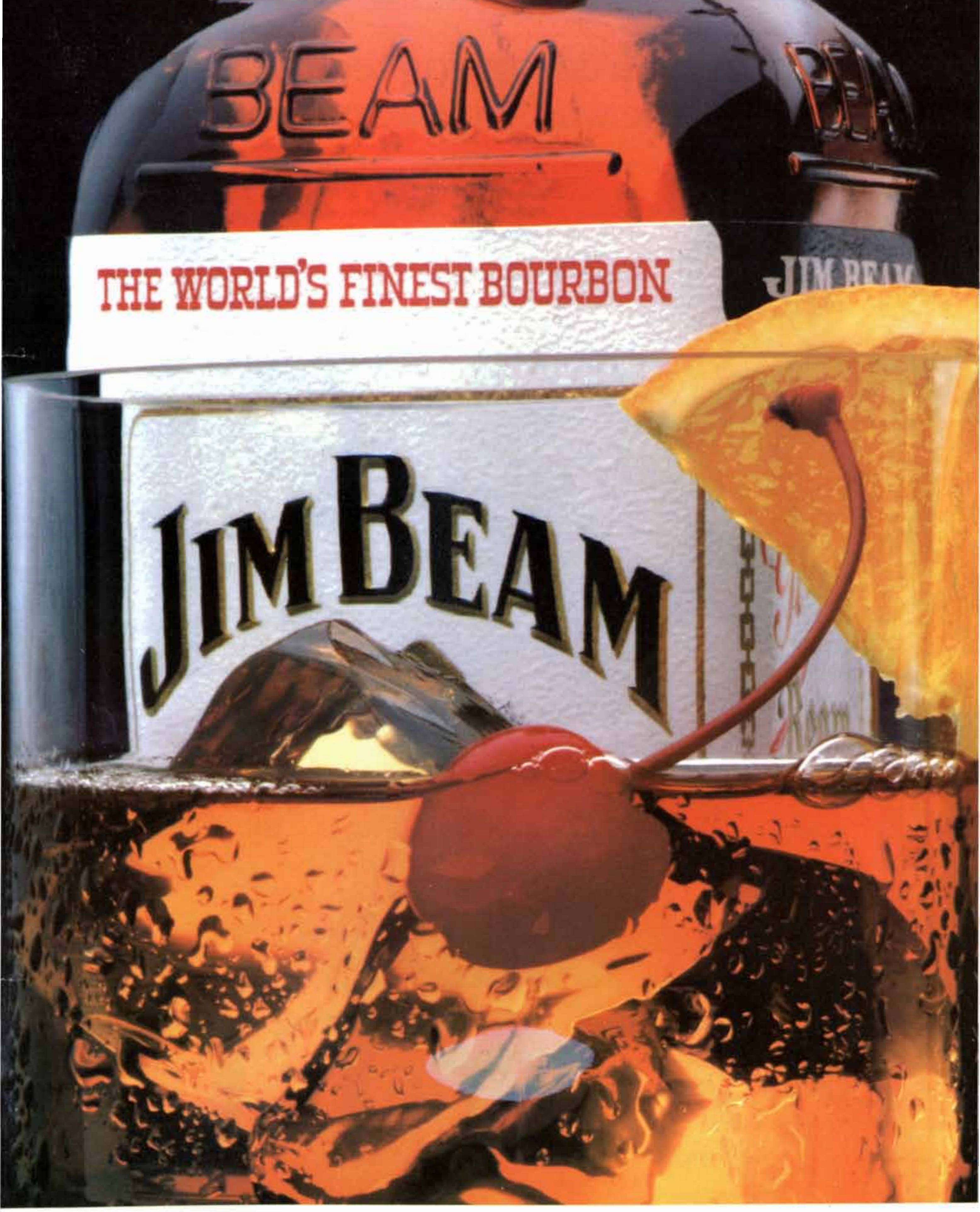
Participants in the conference emphasized that because the problem is one of human resources it should not be considered solely in terms of supply and demand. Engineering schools were said to be attracting not only more students but also better ones, and the quality of education provided them is at issue. David commented: "With the shortage of engineering faculty and the deteriorating academic environment for engineering, these young men and women will not receive the education that they want, that they deserve and that the times require."

## *Resonant Proteins*

Techniques based on nuclear magnetic resonance (NMR) have lately been developed for mapping the three-dimensional distribution of tissues in the body (see "NMR Imaging in Medicine," by Ian L. Pykett; SCIENTIFIC AMERICAN, May). NMR methods of a quite different kind are now being applied to the analysis of three-dimensional structures on a much smaller scale, namely those of protein molecules. Kurt Wüthrich and his colleagues at the Swiss Federal Institute of Technology in Zurich have refined an NMR technique that allows each hydrogen atom in a molecule to be correlated with a peak in the NMR signal. The correlations are so accurate and efficient that the several hundred hydrogen atoms in a protein molecule can be mapped in the time formerly needed to plot the positions of only a few atoms. So far the three-dimensional structures of six small and medium-size proteins have been elucidated. The work is reported in *Journal of Molecular Biology*.

NMR methods have long been employed in biochemistry, but up to now X-ray-diffraction studies have yielded the most comprehensive data on three-dimensional molecular structures. One serious limitation of X-ray diffraction is that the protein being analyzed must first be stabilized in a crystal. The conformation of the protein in the living organism, however, can be profoundly different from its conformation in a crystal. Few proteins have a single, rigid form; most of them can bend or twist, and indeed such changes of shape have important biological functions.

NMR methods do not require a crystallized specimen; they can be applied to a protein in solution. Until recently, however, NMR techniques have been limited by the lack of a reliable way to resolve the magnetic signals and assign them to individual protons, or hydrogen nuclei, in a protein. In an NMR study the sample is placed in a constant magnetic field and a second field that rotates in the plane at right angles to the first field is superimposed for a short time. Each proton in the sample has a magnetic moment, which can be reversed by the rotating field. When the rotating field is turned off, the protons revert to their original orientation and emit an electromagnetic signal, which can be detected. The exact energy needed to reverse the magnetic moment of a given proton de-



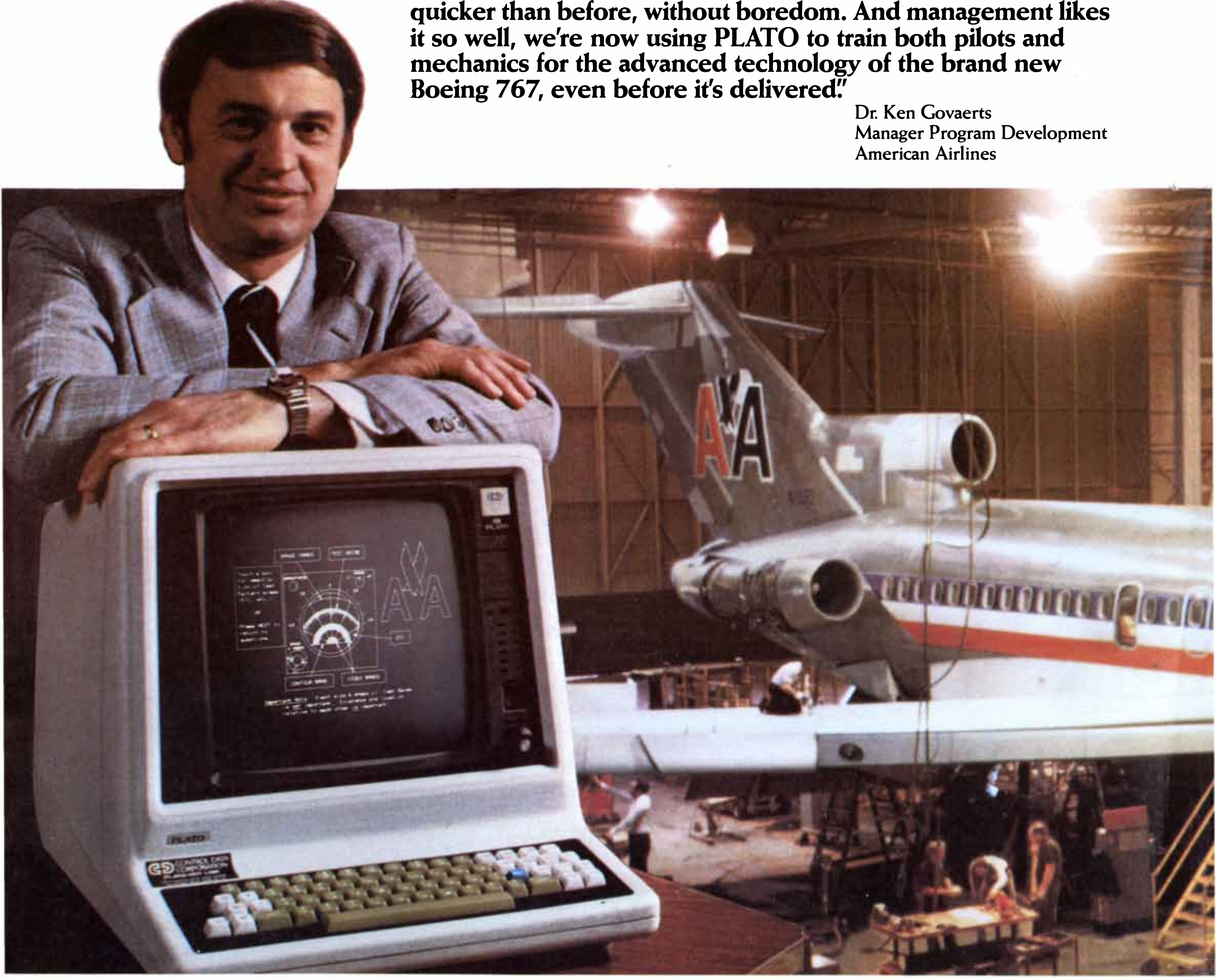

Taste is all it takes to switch to Jim Beam.

KENTUCKY STRAIGHT BOURBON WHISKEY. 80 PROOF. DISTILLED AND BOTTLED BY JAMES B. BEAM DISTILLING CO., CLERMONT, BEAM, KY.

# Keep pace with advancing technology by training the way American Airlines does.

"PLATO computer-based training is remarkably effective. Our mechanics like it because they learn what they need to know quicker than before, without boredom. And management likes it so well, we're now using PLATO to train both pilots and mechanics for the advanced technology of the brand new Boeing 767, even before it's delivered."

Dr. Ken Govaerts
Manager Program Development
American Airlines



# PLATO

## CHANGING HOW THE WORLD LEARNS

pends on the configuration of the other nuclei and the electrons in its vicinity. In a protein molecule each proton generally has a different chemical environment, and so each proton emits at a different point in the spectrum of applied energies. In principle the spectrum can be correlated with the position of each proton in the molecule. In practice the method has led to only rough, qualitative assessments of molecular structure.

Several methods developed by Wüthrich in collaboration with R. R. Ernst, also of the Swiss Federal Institute, now make it possible to associate each spectral peak with a proton position. Two of the techniques identify pairs of protons connected to each other by fewer than four atomic bonds; such pairs must lie in a single amino acid of a protein molecule. In a third technique pairs of protons in close proximity to each other (but not necessarily connected by interatomic bonds) can similarly be identified. By this means the distance between protons in different amino acids can be estimated. From the sequence of amino acids on the backbone of the protein molecule, which is known independently, the spatial conformation of the protein can be determined.

Wüthrich and his colleagues have now mapped the protons in glucagon (a hormone that regulates certain functions of the liver), pancreatic trypsin inhibitor (which regulates the activity of a digestive enzyme), three proteins from the venom of snakes and busi II (a protein that controls the production of semen in bulls).

## *299,792,458 Meters per Second*

The special theory of relativity states that the speed of light in a vacuum is a constant of nature. Measurements of the speed made in any frame of reference yield the same value. The speed of light may soon become a constant of another kind: a defined quantity rather than a measured one, with a value fixed by international convention. The value likely to be adopted is 299,792,458 meters per second.

The change in status would result from a proposed redefinition of the meter. Originally the meter was set equal to one ten-millionth of the distance from the North Pole to the Equator, measured on the meridian passing through Paris; somewhat later it was defined by reference to a platinum-iridium bar; since 1960 the standard meter has been a certain multiple of the wavelength of light emitted by krypton gas. In the new definition the standard unit of length would no longer be a fundamental one; instead it would be derived from the standard unit of time, namely the second. According to a proposal now being considered by a committee of the International Bureau of Weights and Measures, the meter would be defined as "the distance traveled in a time interval of 1/299,792,458 of a second by plane electromagnetic waves in a vacuum." In other words, the meter would be 1/299,792,458 of a light-second.

The meter and the second are already closely coupled. The best available method of establishing the speed of light is to compare the wavelength of electromagnetic radiation with its frequency, and thus to compare a standard of length with a standard of time. Under the proposed redefinition such experiments would still be possible, but their results would have a different interpretation. The speed of light would no longer be subject to revision; any refinement in the accuracy of measurement would alter not the velocity of light but the length of the meter.

## *Gambler's Ruin*

A few years ago Herman Chernoff, a professor of mathematics at the Massachusetts Institute of Technology whose specialty is the theory of probability, became interested in the Numbers Game, a daily lottery run by the state of Massachusetts. In the lottery a bettor picks a four-digit number (from 0000 to 9999) and wins if it matches a number chosen at random that day. The state keeps 40 percent of the total amount bet and divides the remaining 60 percent among the winners. A simple analysis of these rules might suggest that any long series of bets would lose 40 percent of the amount staked. Chernoff noted, however, that a number chosen by few of the bettors brings a greater reward if it wins because the payoff is shared by few. Chernoff also noted that some numbers were less popular than others. For example, the payoffs for four-digit numbers beginning with 0 or 9 were almost always comparatively large in the first 100 days of the game.

Chernoff reports on his experiences with the lottery in *The Mathematical Intelligencer*. As his interest became known he "was visited by a number of students who had tried to beat the game. In each case they had failed." The reasons were evident. Suppose a gambler selects 1,000 unpopular numbers and bets on all of them for 300 days. Since he is betting on a tenth of the 10,000 numbers, the probability that he will win on any given day is .1 and the expected number of wins over 300 days is 30. There is a probability of about 15 percent, however, that the actual number of wins will be less than 25 and a "nontrivial probability of about 2.5 percent" that it will be less than 20. The prudent gambler must therefore choose numbers so unpopular that their payoffs will more than compensate not only for the 40 percent that goes to the state but also for the chance that there will be fewer wins than expected. This may well mean he cannot safely bet on 1,000 numbers. Yet if he bets on fewer numbers, his probability of winning decreases and it becomes more likely that his finite bankroll will be exhausted before any potential gain can be realized. The phenomenon is known in the theory of probability as gambler's ruin.

Suppose too the gambler selects numbers "by virtue of their history of large payoffs." Such a history reflects two circumstances: the numbers have not been bet on heavily and they have been "lucky," that is, they have won. The same numbers may continue to be unpopular, but "there is no reason to expect that they will continue to be lucky." Thus Chernoff's visitors "had selected numbers that looked good and had anticipated future payoffs similar to those in the past.... This led to overoptimistic estimates and contributed to disappointing performance." The phenomenon is known as regression to the mean.

Chernoff himself was intrigued by a further aspect of the game. The gambler can bet on the first three or the last three digits of the daily number and receive 14 percent of what a four-digit winner gets. (According to the probabilities, an equitable payoff would be only 10 percent.) If everyone adopts this strategy, however, the advantage is lost; the best situation for a gambler is to bet on three-digit combinations while the rest of Massachusetts is betting on four-digit numbers.

Apparently the bettors caught on fast. For one thing, the payoff tended to decline, which suggests to Chernoff that "the proportion of money being bet on three-digit numbers rose from zero to about 50 percent in the first 720 games and has remained there since." The variability of the payoff from day to day also tended to decline. Evidently "a larger proportion of the bettors has recognized the value of betting on less popular numbers and has managed to identify such numbers."

Nevertheless, Chernoff analyzed two betting systems (although only on paper). In the first one 33 unpopular three-digit numbers were backed for 210 days. The system yielded "a 34 percent return on investment." In the second system 11 unpopular numbers were backed. The system lost 15 percent. "The main reason for the difference," Chernoff notes, "was luck. The first system had 10 hits where 6.93 were expected while the second had eight hits while 8.19 were expected." With so few good numbers to back he found that neither system could be relied on to yield a profit by the end of a year.

The one-year limit is crucial. The income-tax laws allow a gambler to deduct his losses from his winnings within a calendar year. They do not allow him to deduct his losses in one year from his winnings in the following year. "Thus," Chernoff writes, "income-tax problems add to the problems of the gambler."

# The Quantum Mechanics of Materials

*Structural properties of solids can be predicted by a theory in which electrons act like light beams in a box of partially silvered mirrors. The design of materials from first principles may now be in prospect*

by Marvin L. Cohen, Volker Heine and James C. Phillips

The properties of matter are ultimately determined by the laws that govern the interactions of electrons and atomic nuclei. These particles are the constituents of all ordinary matter. Their interactions are completely described by the laws of quantum mechanics, which govern the familiar forces of electrostatic repulsion and attraction and impose certain constraints on the motions of the particles. Until recently, however, the idea that the properties of aggregates of particles could be predicted from such fundamental laws was little more than a philosophical acknowledgment that the laws are indeed fundamental. Except for a few simple systems of electrons and atomic nuclei, such as the hydrogen atom, the properties of aggregates have seldom been derived from first principles. As a result the applications of the fundamental laws of physics to disciplines that describe more complex states of matter, such as chemistry, biochemistry, metallurgy and ceramics, have been quite limited.

In the past 25 years quantum theorists who study solid materials have begun to close the gap between basic principles and practical applications. They have developed new quantum-mechanical methods for predicting the configuration of the valence electrons in an atom or an aggregate of many atoms, within the range of energy excitations in which the atoms form interatomic bonds. The valence electrons of an atom are the outermost, least tightly bound electrons, which form chemical bonds with other atoms. A theory specifying the configuration of the valence electrons therefore has much to say about the bulk properties of matter that depend on the nature of the interatomic bonds. Among these properties are electromagnetic ones (such as electrical conductivity and optical reflectivity), chemical ones (such as the capacity to adsorb various substances), thermal ones (such as specific heat) and mechanical ones (such as hardness, malleability and elasticity).

In the hydrogen atom the configuration of the single electron (which is necessarily a valence electron) can be calculated from the potential energy of the electron. In more complex atoms it is also possible to give a mathematical expression for each valence electron from which its configuration can be determined; such an expression is a kind of generalized potential energy. Quantum theorists have found it mathematically impractical, however, to work with expressions for the true potential energy of the valence electrons in atoms that have a total of more than two electrons. The true potential energy of a valence electron must take into account its interactions with the more tightly bound core electrons of the atom; these interactions are in turn governed by the true potential energy of each core electron. Now a way around this impasse has been found. It turns out there is no need to consider the true potential energy of all the electrons if what is needed most is a good approximation to the configuration of the valence electrons; the interactions of the core and the valence electrons have virtually no effect outside the core. Hence the new method regards the core electrons and the atomic nucleus as if they constituted a single particle without internal structure. The method is called the pseudopotential theory.

The number of valence electrons in an atom is the organizing principle of the periodic table of the elements. Chemists have recognized for some time that the spatial configurations assumed by valence electrons at various energy levels determine the nature of the chemical bond and with it the properties of a molecule or a solid lattice of atoms. Unfortunately distortions in the paths of the valence electrons induced by interactions with neighboring valence electrons and with atomic cores can be quite large. For this reason a general quantum-mechanical prediction of the properties of a substance in terms of the additive properties of separate chemical bonds is not yet feasible for molecules, including even relatively small ones.

There is one realm where prediction is now practical: in crystalline solids. The regularity of the lattice into which the atoms are organized in a crystal makes it possible to calculate the properties of a macroscopic solid, made up of perhaps $10^{23}$ atoms. In other words, many properties of an elemental solid such as lead or a simple binary solid such as gallium arsenide can now be deduced from energy considerations alone: from the identity of the constituent atoms and from the external conditions of temperature, pressure and magnetic field. The improvement in theoretical understanding has broad practical implications for the design of materials whose properties, such as superconductivity, semiconductivity or optical reflectivity, can be precisely specified in advance.

## Early Computational Strategy

It is not hard to appreciate why the laws of quantum mechanics are difficult to apply to the descriptions of matter offered by chemistry, biology and other disciplines in which materials are studied. In part the difficulties arise because the electronic structure of a system of $10^{23}$ particles, even a system with great structural regularity, is enormously complex. Early quantum theorists nonetheless thought successful predictions of the properties of solids would depend on the quantum-mechanical description of the isolated atoms that make up the solid. A corollary to this idea was the supposition that the valence-electron configuration in a solid could be determined, to a first approximation, by superposing the valence-electron configurations of isolated atoms. Once the superposed configuration had been determined corrections could be made to account for the distortion of the electron paths by neighboring atoms.

The strategy that was envisioned is

analogous to predicting the motion of a meteoroid in the solar system. The path of the meteoroid is completely determined by Newton's law of gravitation. It would be a simple matter to find the path the meteoroid would take if the only relevant gravitational force were the one between the meteoroid and the sun. The motion of the meteoroid is slightly perturbed by interactions with the planets, however, so that its trajectory is influenced by several forces, which are changing constantly as both the meteoroid and the planets move.

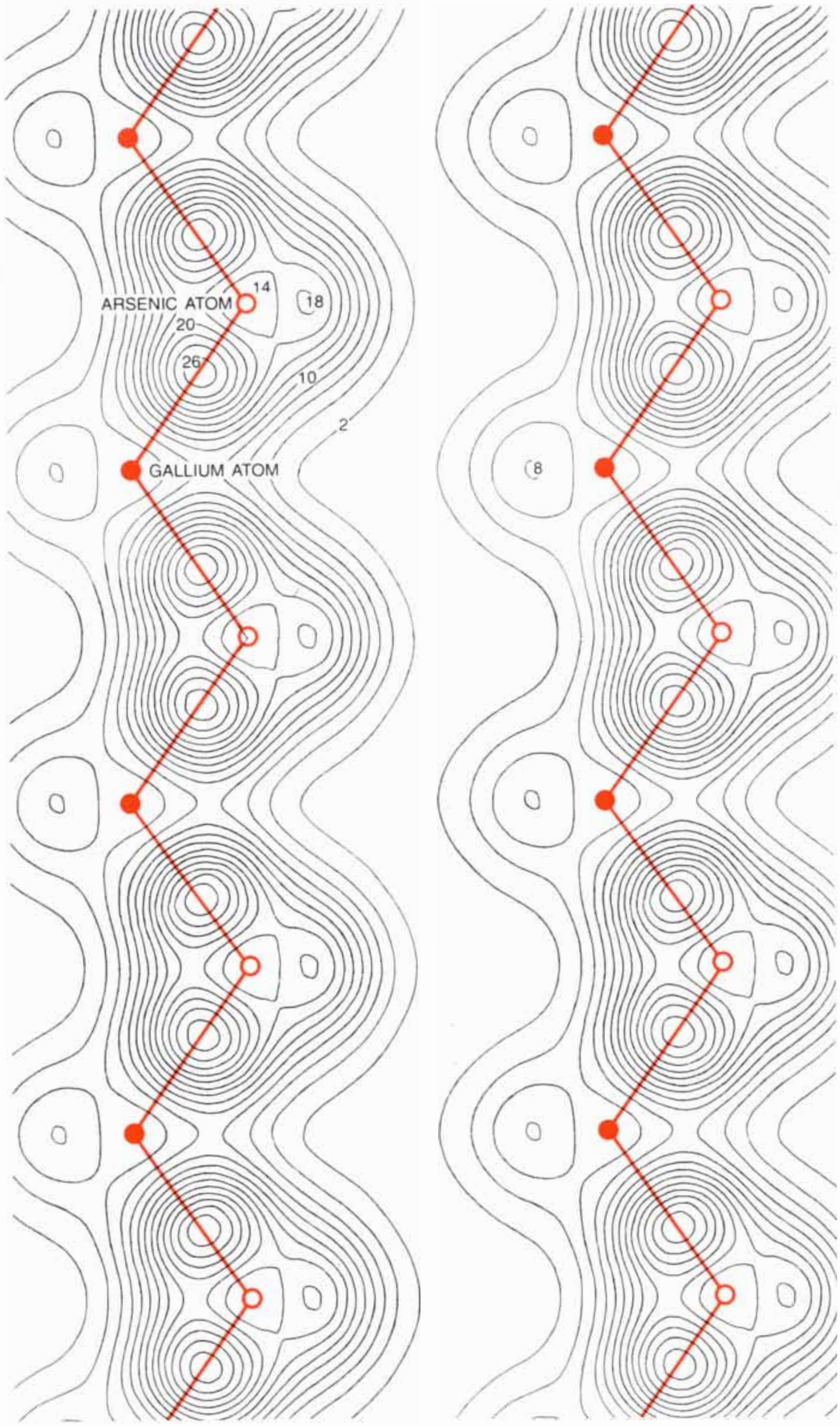
The standard method of calculating the path of such a meteoroid is first to compute the path as if only the sun's gravity were acting and then to determine over short periods how the path is perturbed by the current configuration of the planets. The perturbed position of the meteoroid can then be considered a point on a new orbit around the sun, and the perturbation from the new orbit can be computed in turn. By means of such an iterated calculation the true path of the meteoroid can be approximated to any desired accuracy; the approximation converges to the true path. The method works well, however, only if the first estimate of the orbit is a reasonably good one. This requirement reflects the physical intuition that the sun's gravity predominates over all other effects on the orbit.

In the 1930's many investigators were optimistic that such a computational strategy would lead quickly to the application of the quantum theory to larger aggregates of matter. The theory had already been successful in explaining the sharp lines in the spectrum of radiation emitted by an atomic gas; the lines are formed when electrons make transitions among discrete energy states. The invention of wave mechanics by Erwin Schrödinger and Werner Heisenberg made it possible to predict the properties of simple atoms, such as hydrogen, and of the simplest molecules, such as the diatomic hydrogen molecule ($H_2$).

## Electron Waves

Soon after these successes Erich Hückel of the University of Marburg and Robert S. Mulliken of the University of Chicago found that the method of superposing electron configurations calculated for isolated atoms gives an acceptable approximation of the configuration in certain molecules. For example, they applied the method to diatomic molecules heavier than hydrogen and to hydrocarbons of high molecular weight. Many workers began to think that successful predictions of the properties of solids from the quantum theory would soon be forthcoming. As understanding deepened, however, such hopes were replaced by widespread pessimism.

In quantum mechanics the motion of an electron is defined by an orbital, just

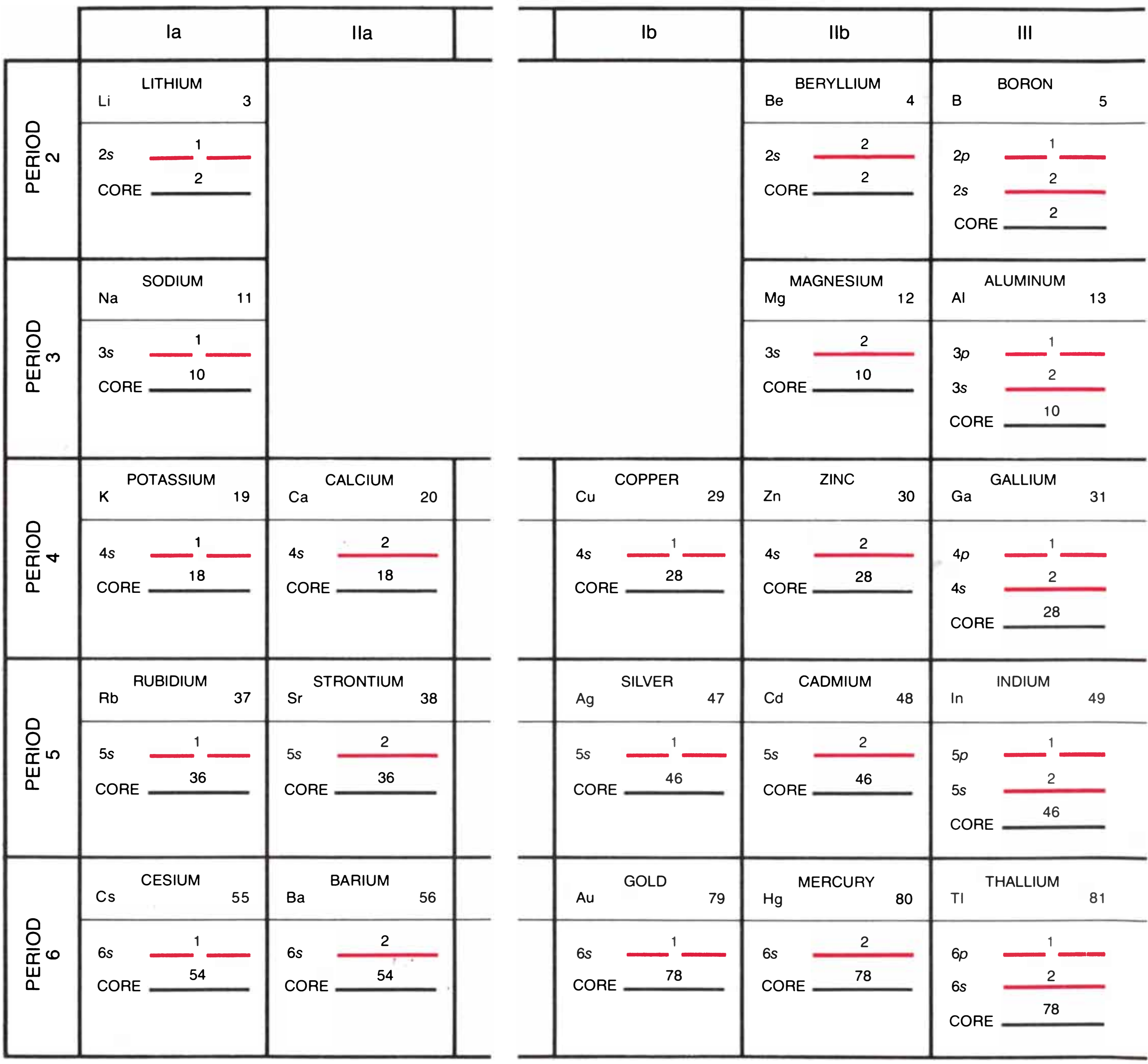



**CHARGE-DENSITY MAP of the valence electrons in a cross section of crystalline gallium arsenide shows the distribution of the electrons with unprecedented accuracy. The valence electrons are the outermost electrons in an atom, which participate in chemical bonding. The curves are contours of equal electronic charge density, in arbitrary units; along a contour the probability of detecting an electron remains constant. Each gallium atom in the crystal is at the center of a tetrahedron having arsenic atoms at its vertexes, and each arsenic atom is at the center of a tetrahedron having gallium atoms at its vertexes. (The three-dimensional structure is shown in the bottom illustration on page 98.) The cross section passes through the centers of some of the atoms; other atoms to which they are bonded are outside the plane of the map. The interatomic bonds that lie in the plane of the map are represented by straight lines. The bonds are characterized by a buildup of charge density. The computer-generated map is the result of a complex calculation that has only recently been made possible by the pseudopotential theory, a method for applying the principles of quantum mechanics to valence electrons in solids. A detailed understanding of such solids as gallium arsenide is of technological as well as scientific importance: the material is widely employed in the manufacture of electronic devices.**

as the path of a meteoroid is given by an orbit. An orbital is not a path, however, but rather a probability distribution. In predicting the properties of organized aggregates of electrons and atomic nuclei one can at best determine the probability that electrons occupy various regions of space. The probability can be thought of as the relative density of electronic charge.

In calculating the probability distribution or charge density for a given electron it is convenient to consider the electron to be a superposition of waves rather than a particle. The charge density in any region is the square of the magnitude of a complex number whose value varies from point to point in space. (A complex number has both a real part and an imaginary part, or a multiple of the square root of $-1$. If a complex number is plotted as a point in a plane, so that the value of the real part is labeled along the $x$ axis and the value of the imaginary part is labeled along the $y$ axis, the magnitude of the complex number is the distance from the origin to the point in the plane.) The equation that generates the complex number has the same form as the equation that gives the amplitude of a wave. Because the amplitude of the electron wave depends on position in space the mathematical expression that gives the amplitude of the electron wave throughout space is called the electron wave function.

The wave function is an abstract mathematical expression that cannot be measured directly. Certain properties such as the charge density, however, can be derived from the wave function and studied empirically. For example, wherever two waves contributing to an electron wave function intersect in space, their amplitudes add algebraical-




**PARTIAL PERIODIC TABLE of the elements gives the structure of the core and the valence electrons for the elements having the simplest repetitive valence properties. Partially filled energy levels for the valence electrons are shown as broken colored lines; filled valence-electron energy levels are shown as solid colored lines and core-electron energy levels are shown as solid black lines. The number of electrons in each energy level is given above each line; the symbols to the left of each line are the traditional names of the atomic orbitals, or en-**

ly and the electronic charge density at that point is the square of the resulting sum of amplitudes. The algebraic addition of wave amplitudes is called interference. If two wave functions reinforce each other, the result is constructive interference; if they tend to cancel, the result is destructive interference.

### Standing Wave Patterns

An electron wave is scattered, or reflected in all directions, when it encounters an atomic nucleus, just as a water wave is scattered when it strikes a rock that breaks the surface. The scattered wave interferes with the incident wave both in the direction of travel of the incident wave and in other directions. The interference creates a pattern of nodes where the waves cancel, which alternate with other regions where the waves reinforce. In some cases the result is a standing wave much like the one set up on a plucked string. If the electron wave encounters two or more atomic nuclei, the wave is scattered by each nucleus and the interference pattern that results is given by the algebraic sum of the amplitudes of the incident wave and all the scattered waves at each point in space.

In the general case, with many atomic nuclei arranged randomly, the interference pattern of the electron wave may resemble the chaotic surface of a stormy sea. Certain simple geometric arrangements of atoms, however, can give rise to simple interference patterns. For instance, when a plane electron wave, in which the troughs and crests are parallel lines, is scattered by a plane of atoms, the interference pattern is made up of two plane waves: one wave is transmitted through the plane of atoms in the initial direction of wave travel and the other is reflected from the plane of atoms just as a light ray is reflected from a mirror. In all other directions the scattered waves interfere destructively and the amplitude is zero. In a crystal lattice, where the regular packing of atomic nuclei generates planes of atoms in several directions, an electron wave is transmitted and reflected as if it were a beam of light in a box filled with many half-silvered mirrors. At the boundary of the lattice, from which the electronic charge cannot escape, the plane of atoms acts like a fully silvered mirror, reflecting the entire wave back into the crystal.

The electron wave function is the steady-state, standing interference pattern that results from all scattering, reflection and transmission of the electron wave in a given atomic environment. The wave function also depends on the energy of the electron: roughly speaking, the higher the energy of the electron, the greater the number of wave oscillations per unit time. When a second electron is present in a given region of space, it too generates a pattern of standing waves.

Each standing-wave pattern determines how the charge density of a single electron varies throughout a region of space. The total charge density in the region is the sum of the charge densities of all the electrons. The primary goal of quantum-mechanical analysis is to find, for a given arrangement of atomic nuclei, the standing-wave pattern of minimum energy that results from the interaction of the charge density of each electron in a structure with the atomic nuclei and with the charge density of every other electron. Such a description of a system of electrons and nuclei is given by a solution to the mathematical expression called the Schrödinger equation.

### Early Computational Difficulties

When a solution to the wave equation is sought for a physical system incorporating more than one atom or more than one electron, the interference of the electron waves usually makes it impossible to express the wave functions in simple mathematical terms. The amplitudes given by the simple method of superposing the wave functions of isolated atoms turn out in most cases to be a

| IV | V | VI | VII |
|---|---|---|---|
| CARBON<br>C 6 | NITROGEN<br>N 7 | OXYGEN<br>O 8 | FLUORINE<br>F 9 |
| 2p 2<br>2s 2<br>CORE 2 | 2p 3<br>2s 2<br>CORE 2 | 2p 4<br>2s 2<br>CORE 2 | 2p 5<br>2s 2<br>CORE 2 |
| SILICON<br>Si 14 | PHOSPHORUS<br>P 15 | SULFUR<br>S 16 | CHLORINE<br>Cl 17 |
| 3p 2<br>3s 2<br>CORE 10 | 3p 3<br>3s 2<br>CORE 10 | 3p 4<br>3s 2<br>CORE 10 | 3p 5<br>3s 2<br>CORE 10 |
| GERMANIUM<br>Ge 32 | ARSENIC<br>As 33 | SELENIUM<br>Se 34 | BROMINE<br>Br 35 |
| 4p 2<br>4s 2<br>CORE 28 | 4p 3<br>4s 2<br>CORE 28 | 4p 4<br>4s 2<br>CORE 28 | 4p 5<br>4s 2<br>CORE 28 |
| TIN<br>Sn 50 | ANTIMONY<br>Sb 51 | TELLURIUM<br>Te 52 | IODINE<br>I 53 |
| 5p 2<br>5s 2<br>CORE 46 | 5p 3<br>5s 2<br>CORE 46 | 5p 4<br>5s 2<br>CORE 46 | 5p 5<br>5s 2<br>CORE 46 |
| LEAD<br>Pb 82 | BISMUTH<br>Bi 83 | POLONIUM<br>Po 84 | ASTATINE<br>At 85 |
| 6p 2<br>6s 2<br>CORE 78 | 6p 3<br>6s 2<br>CORE 78 | 6p 4<br>6s 2<br>CORE 78 | 6p 5<br>6s 2<br>CORE 78 |

**ergy levels. The number in the upper right of each cell is the atomic number of the element, and the abbreviation in the upper left is its chemical symbol. Because the elements in a given column of the table have similar valence-electron configurations they also have similar chemical properties. The similarity guided the development of the pseudopotential theory.**

# Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

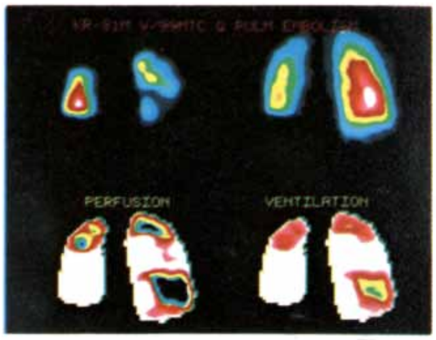
**Comprehensive.** Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

**Complimentary CME.** At no extra cost, we offer a CME program in which you can earn 32 top credits* per year.

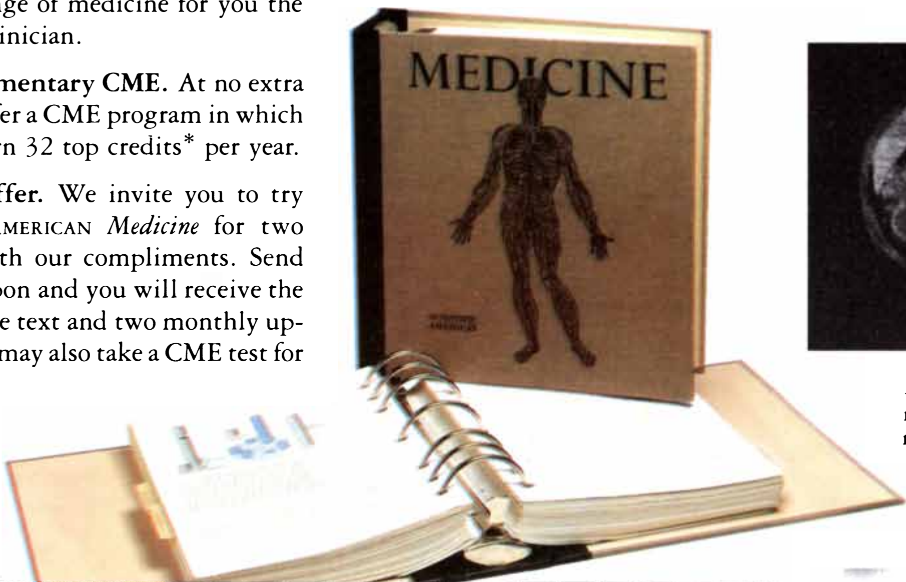
**Trial Offer.** We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $220 for the first 12 months (current renewal is $170); otherwise, return the books; there is no charge.

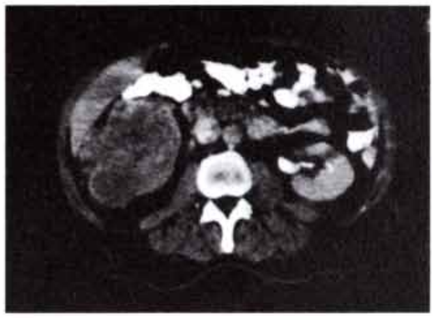
Please mail the coupon today so that we can put medical progress to work for you.

You may also order your trial copy by calling toll-free 1-800-345-8112 (in Pennsylvania, 1-800-662-2444).

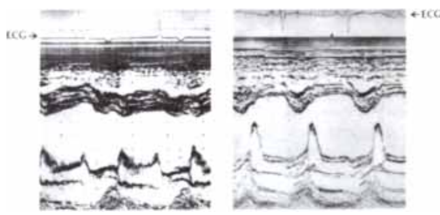



**Computerized scintigraphy reveals pulmonary thromboembolism.**



**Abdominal computed tomogram reveals large renal carcinoma replacing part of right kidney.**




**Echocardiograms from patients with aortic regurgitation.**

## SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $220 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $220*
☐ Bill me ☐ VISA ☐ MasterCard

Expiration Date__________ Account Number__________

*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name__________ MD Specialty__________

Address__________

City__________ State__________ Zip__________

Signature__________

Please allow six to eight weeks for delivery. All payments must be in U.S. dollars. CME is available outside the U.S. and its possessions for an additional $65. Surface routes used unless airmail delivery is prepaid.

P1

*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

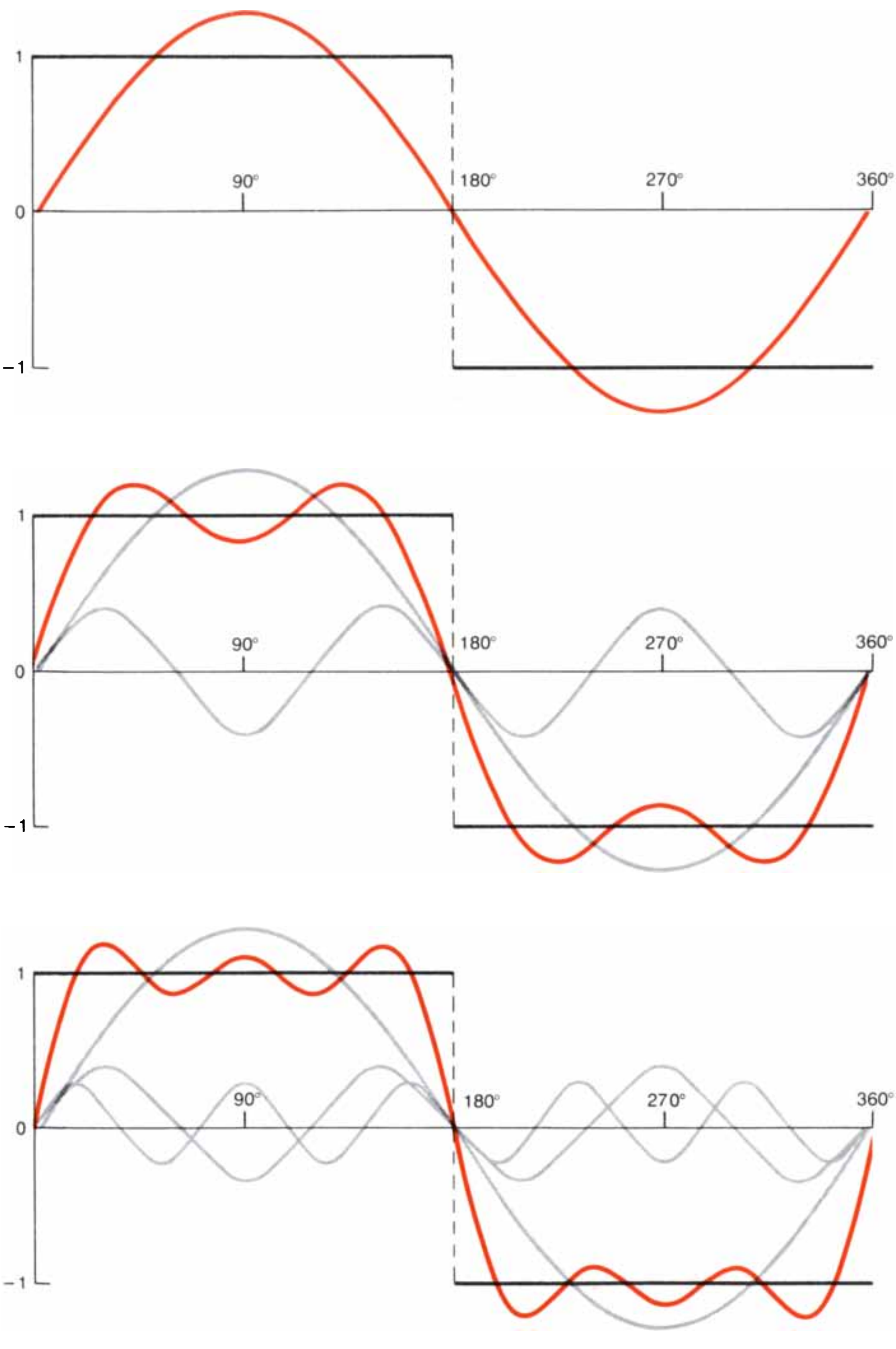
poor approximation. When investigators tried to employ the approximation as the first step in a perturbative calculation analogous to the one that gives the orbit of a meteoroid, they often found that in solids the iterations converge too slowly to be of more than qualitative value. Hence, with the exception of the work of Hückel and Mulliken on selected molecules, early attempts to understand the properties of matter by solving the wave equation fared poorly.

The decision to start with isolated atoms, in the belief that the electron wave functions in a solid or a molecule could be computed as mere perturbations, was only one of the impediments to progress. Another difficulty was the practice of considering all the electrons in each atom from the inside out. The first wave functions to be calculated were those for the innermost, lowest-energy electrons. Wave functions for higher-energy electrons, and ultimately for the valence electrons, were calculated by taking into account the interference effects of all the lower-energy electrons whose wave functions had already been computed. The wave functions of the higher-energy electrons, when they are calculated in this way, depend quite sensitively on the values of the previously computed wave functions. Thus any errors or small approximations in the description of the inner electrons can give rise to enormous variations in the predicted configuration of the valence electrons. For example, in the lead atom which has 82 electrons, the risk of compounding errors of approximation in such a calculation is considerable.

One of the most persevering of the early investigators was Douglas R. Hartree of the University of Cambridge, who used a desk calculator to determine numerical values of the combined electron wave functions for isolated atoms of the first 29 elements in the periodic table. As his results accumulated Hartree found that the evaluation of the wave function at each point in space was the lesser part of his task. The more important part was the display of numerical results in a way that would highlight their physical significance. Even when such laborious calculations are sufficiently accurate and error-free, the data relevant to interatomic bonding must be distinguished from the far greater quantity of data about the atomic core that has no bearing on the kind of chemical bond that is formed. The problem of what might be called numerical noise grows to nightmarish proportions in calculations done with a modern electronic computer. Thus, paradoxically, the computer makes ever greater demands on our understanding of physical processes as its capabilities increase.

Three sources of computational difficulty can therefore be identified in early approaches to the quantum mechanics of solids: the preoccupation with the isolated atom, the practice of computing the electron wave functions from the inside out and the problem of discerning physically important processes above the numerical noise. These problems were dramatized by the patent inability of early quantum-mechanical calculations to account for several important findings of structural chemistry.

One failing of the early methods was in predicting the length of chemical bonds, or the distance between two atomic nuclei in a solid or a molecule. Linus Pauling of the California Institute of Technology and other structural chemists showed that although bond lengths are roughly constant from one substance to another, significant differ-

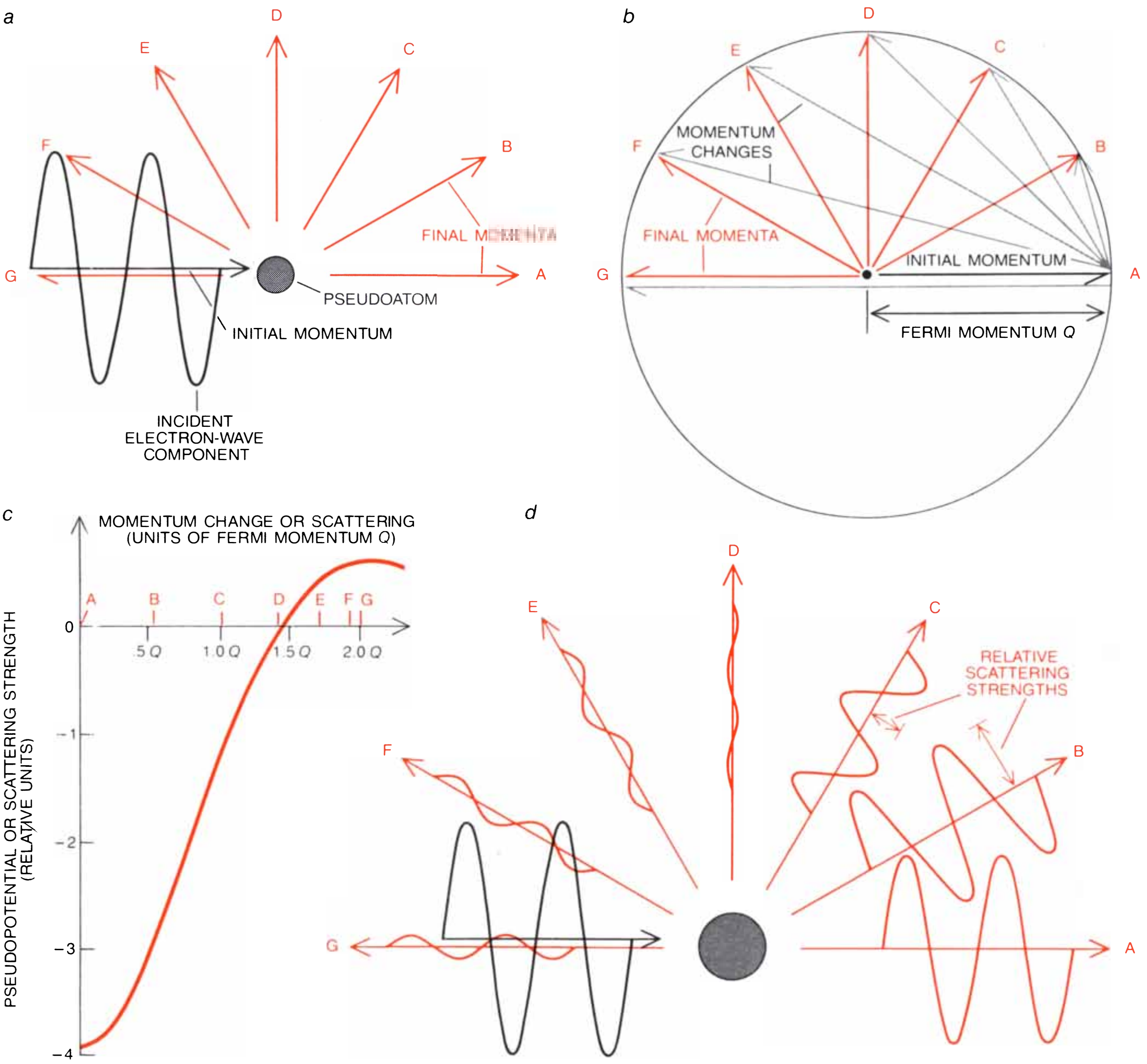



**FOURIER ANALYSIS can approximate the shape of a waveform to any desired accuracy by superposing sine and cosine waves of certain amplitudes and frequencies. In the pseudopotential theory the electron waves are represented by their Fourier components (the sine and cosine waves) because of the simplicity of manipulating Fourier-analyzed waves with a digital computer. In the diagrams the square wave to be approximated is drawn in black; the gray sine waves in the lower two diagrams are the Fourier component waves from which the approximation is constructed. The curves in color are obtained by adding the amplitudes of all the Fourier components at each point; the three diagrams show how the first three stages of the Fourier approximation converge to the shape of the square wave. (The colored wave in the diagram at the top is the single Fourier component of the first-order approximation.) In practice as few as four or as many as 1,000 sine and cosine waves have been employed to describe an electron wave.**

ences in chemical structure may be associated with quite small differences in the bond length.

The length of the bond between two carbon atoms in diamond, a pure-carbon material, is the same as it is in ethane ($CH_3$–$CH_3$); in both substances each carbon atom is bonded to four other atoms, one at each corner of a tetrahedron. In ethylene ($CH_2$=$CH_2$), however, the carbon-carbon bond is measurably shorter, presumably because each carbon atom is bonded to only three other atoms. (There is a so-called double bond between the two carbon atoms in ethylene.) Moreover, the carbon-carbon bond in graphite, another pure-carbon material, has still another length, even though, as in ethylene, each carbon atom in graphite is bonded to three nearest-neighbor atoms. All these differences result from the mutual interplay of the electron wave functions and from their tendency to assume the least-energetic standing-wave pattern that is consistent with the laws of quantum mechanics.

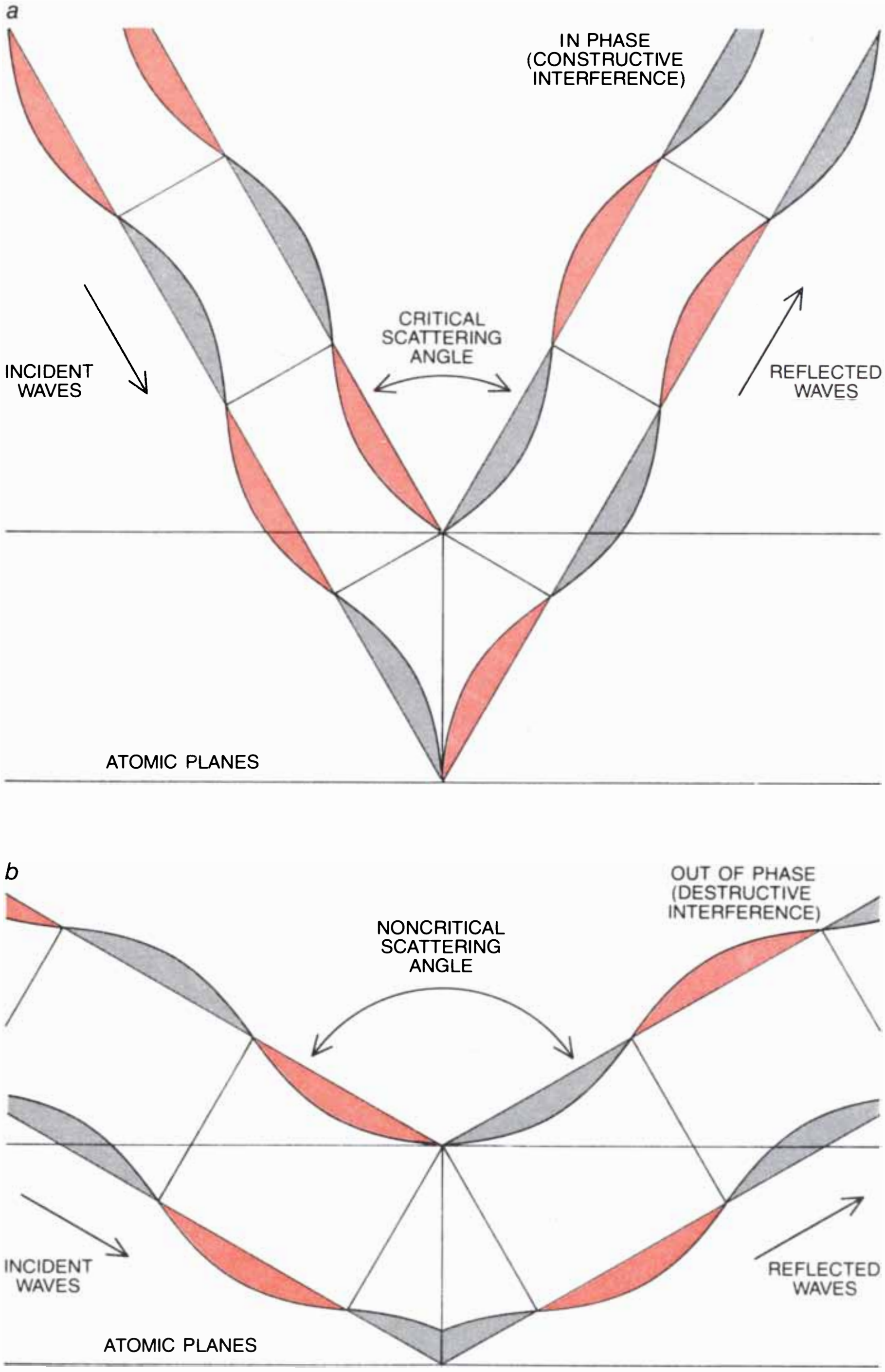
The structural differences among materials made up of the same constituent atoms, such as diamond and graphite, are called phase differences. They can correspond to slightly different total energies of the solid. Studies of the phase




**PSEUDOPOTENTIAL can be understood as the net scattering strength of an atom in a crystal. The pseudopotential represents the effectiveness of the atom in transforming the momentum of a single Fourier component of an electron wave into some specified final momentum. The momentum of a wave in quantum mechanics depends on its direction of propagation and on the number of oscillations of the wave per unit distance. At low energy the only changes in the momentum of an electron wave that are induced with any significant scattering strength are changes in the direction of the wave (*a*). Because of quantum-mechanical constraints on the allowed momentum states of an electron the initial and final momenta of the scattered waves are approximately equal in magnitude (although not in direction) to the quantity called the Fermi momentum of the solid. If the momenta of the incident wave (*black arrow*) and the scattered wave (*colored arrows*) are represented as vectors with a common origin (*b*), the change in momentum caused by scattering can be represented as the vector difference between the incident-momentum and the scattered-momentum vectors (*gray arrows*). The pseudopotential depends only on the change in momentum of the electron-wave component; thus it can be graphed as a function of the length of the gray arrows. The graph (*c*) shows the pseudopotential, or the scattering strength, of aluminum atoms in a solid. The scattering strength is the ratio of the amplitude of the scattered wave to the amplitude of the incident wave. Hence the relative amplitudes of the scattered electron-wave components in aluminum can be determined from the graph and represented as waves scattered in various directions (*d*).**

diagrams of many materials have shown that competing solid structures often differ in energy by only about .1 electron volt per valence electron. In comparison quantum-mechanical calculations ordinarily deal with energy differences on the order of a tenth of the binding energy of the electron in the hydrogen atom, or about one electron volt.

Although the differences in energy among competing solid phases are small, they usually correspond to differences in interatomic structure. The structural differences determine some of the most important properties of the solid, such as whether it is an electrically conductive metal, a semiconductor or an insulator.




**BRAGG DIFFRACTION is the scattering of waves by the parallel planes of atoms in a crystal. When an electron wave intersects a single plane of atoms, part of the wave is reflected and part of it is transmitted. A wave impinging on a set of parallel planes of atoms can therefore be reflected along more than one path. When the direction and wavelength of an incident wave and the spacing between the atomic planes are such that the path lengths differ by an integral number of wavelengths, the waves traveling along the various paths emerge from the planes in phase and interfere constructively (*a*). For other incident angles the interference is destructive and the amplitudes of the emerging waves cancel (*b*). Thus the scattering of electron waves by a crystal is simpler than the scattering of waves by a single atom, which scatters waves in all directions. The pseudopotential needs to be evaluated only for the critical scattering angles.**

The calculations of the early quantum theorists were often unable to discriminate among the conducting, semiconducting and insulating phases of solids. They were barely accurate enough to predict when a group of isolated atoms in a gas would begin to favor the solid state. Any two solutions to the wave equation whose total energies differed by less than half an electron volt per valence electron could not be distinguished by early quantum-mechanical methods. The failure of such methods to resolve energy differences that correspond to the most important phase transitions in solids made the prospects for a quantum theory of solids seem bleak indeed.

## Computational Resolving Power

How can quantum calculations be made to distinguish among the competing structures that determine chemical properties? The answer is to arrange the calculations so that they are most sensitive to the range of energy differences and chemical interactions one wants to investigate. In doing so one may have to abandon the comprehensive vantage that can be gained through an exhaustive, brute-force calculation. The choice is between a comprehensive overview in which fine details cannot be resolved and a closeup examination in which large-scale features are deemphasized.

One of the best ways to improve the resolving power of quantum calculations for the energy differences of interest in the study of materials is to allow the computational methods to be guided by the periodic table. Early quantum theorists, when solving the wave equation for isolated atoms, considered each element individually. According to the periodic table, however, for the lighter atoms the chemical properties of an element with atomic number $A$ are similar to the properties of the element with atomic number $A + 8$. (For the heavier atoms there are similar but more complicated periodicities.) The reason for the similarities is that the number of valence electrons in two such atoms is the same, and the chemical properties depend almost entirely on the number of valence electrons. The inner electrons are so tightly bound to the nucleus that they can be considered part of a "frozen" atomic core. Moreover, the negatively charged electrons in the atomic core neutralize or screen part of the positive charge of the nucleus. If quantum calculations are to be maximally sensitive to the energy levels of interest in chemistry, such facts should be built into the computational method from the outset.

In 1939 W. Conyers Herring, then of Princeton University, showed how the effects of the atomic core could be minimized. Herring observed that the wave function of a valence electron must have one more node, or zero value, than the wave function of the outermost core electron. The extra node lies near the region where the wave function of the outermost core electron has its maximum value. In other words, the amplitude of the valence-electron wave function is approximately zero on the surface of the core, and so the square of the amplitude, or the probability that the valence electron will be found in the region, is also nearly zero. Hence the valence electron tends to avoid the surface of the atomic core.

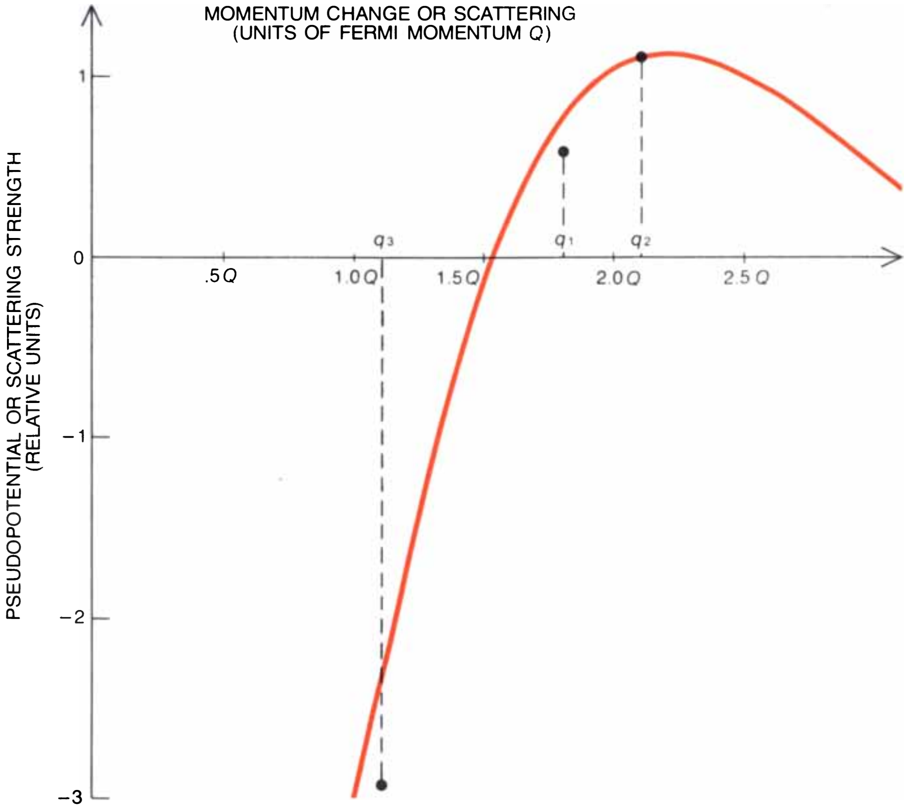
### Energy within the Atomic Core

What about differences in the energy states of the valence electron that might be caused by differences in the wave function of the valence electron within the core region? In answer to this question Herring made a crucial contribution. From his numerical calculations he observed that the extra oscillation of the valence-electron wave function associated with its extra node implies that the valence electron has additional kinetic energy when it passes through the core region. He found that the extra kinetic energy (which is always positive) is almost exactly canceled by the large negative potential energy the electron acquires by approaching the nucleus as it enters the core. It follows that the size of the core makes little difference to the energy of the valence electron outside the core, because its total energy within the core is always quite close to zero.

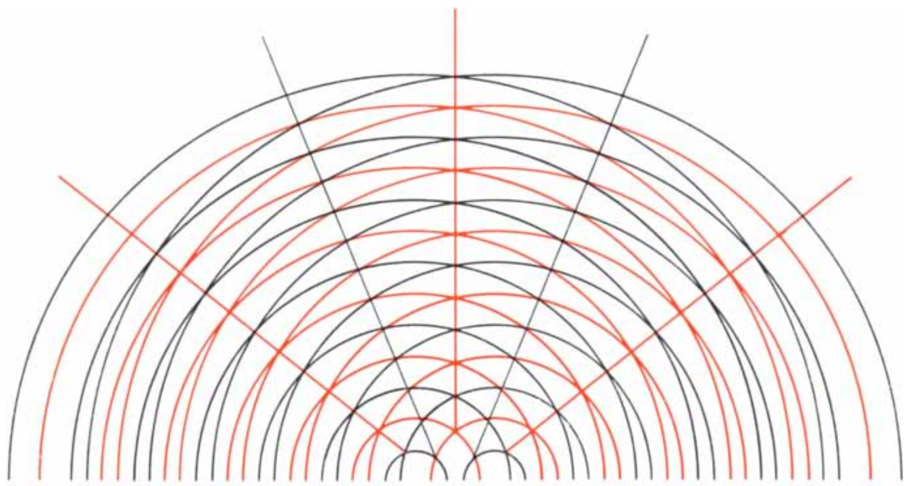
In 1939 Herring succeeded in explaining why solid beryllium, which has a total of four electrons (including two valence electrons), coheres as a close-packed metal. Nine years later Herring and Frank Herman of Columbia University analyzed the structure of diamond by solving the wave equation; the calculation was done with an early electronic digital computer. On the other hand, in later applications Herring's method failed to resolve the measured differences in electronic energy that distinguish carbon from silicon and germanium, which are heavier but occupy the same column of the periodic table.

With the benefit of hindsight it is possible to trace such difficulties to the method of solving the wave equation for an isolated atom from the innermost electrons outward. For heavy atoms the approach obscures the simplicity inherent in the periodic table. From 1957 to 1966, however, an international group of solid-state theorists developed new methods that incorporated Herring's ideas about the energy cancellation of the valence electrons but avoided many conceptual and computational difficulties. These methods constitute the pseudopotential theory. The group included G. Franco Bassani of the University of Rome, Morrel H. Cohen of the University of Chicago, Walter A. Harrison of Stanford University, John M. Ziman of the University of Bristol, ourselves and many others.

Instead of defending their methods on purely theoretical grounds the new generation of investigators focused their efforts on comparing the predictions of their theory with experiment over as wide a range as possible. Such comparisons are often so difficult that initially the group hoped only for modest success. To their surprise they encountered a situation quite rare in science: the theory proceeded from success to success




**SCATTERING STRENGTH of isolated silicon atoms in a gas (*colored curve*) is compared with their scattering strength when they are arranged in a crystalline solid (*black dots*). The changes in momentum $q_1$, $q_2$ and $q_3$ are the changes brought about by Bragg diffraction of the electron waves by three sets of parallel planes of atoms in the crystal. The differences between the scattering strength in the gas and the scattering strength in the solid are caused mainly by the rearrangement of the valence charge when the atoms are bonded together in the solid.**



**SOLUTIONS TO THE WAVE EQUATION that describes the motion of electrons in a solid can be standing waves, or wave patterns that persist over time. Any computation of standing-wave patterns must take account of the effects of wave interference. For a two-dimensional surface on which there is no wave reflection at any boundaries identical circular waves propagated from two point sources generate the interference pattern shown. Where wave crests coincide with crests (*colored curves*) and troughs coincide with troughs (*black curves*) the interference is constructive (*colored lines*). Where crests coincide with troughs the interference is destructive (*black lines*). The bulk properties of a solid can be predicted from the standing wave.**

almost without effort. The new methods solved long-standing problems that older theories had been unable to treat and predicted certain features of the quantum structure of solids well before they were measured experimentally. The experimental confirmations of the theory played a decisive role in settling otherwise unresolvable differences in methodology. By 1970 the pseudopotential theory had gained almost universal acceptance, and today it stands virtually alone as the theory of the quantum structure of solids.

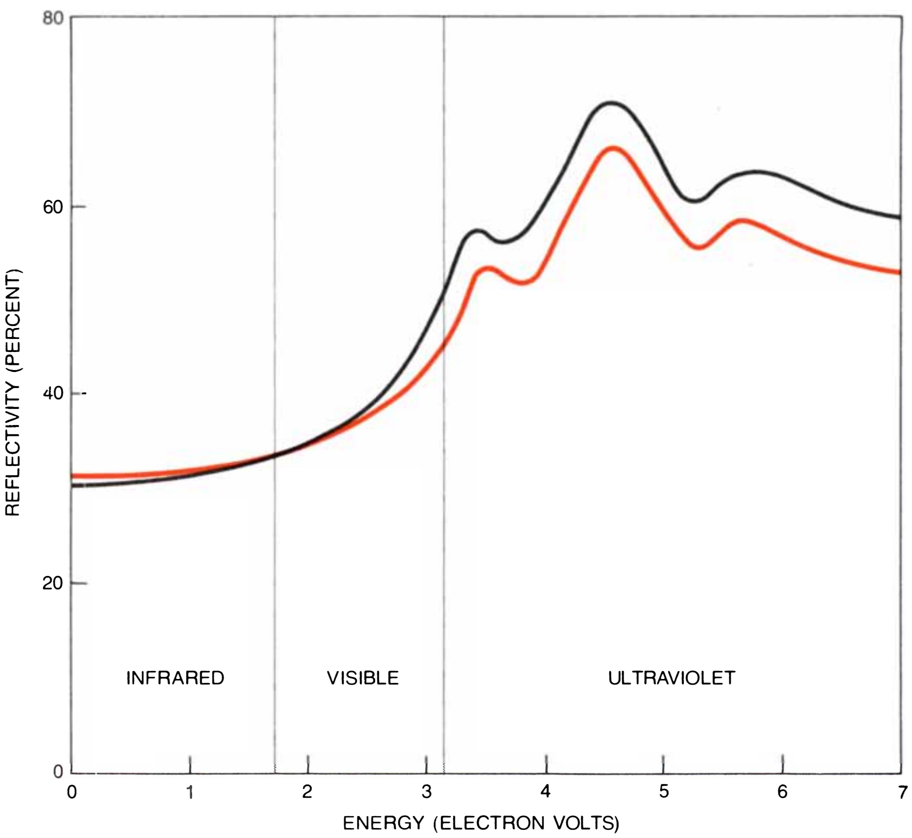
## The Pseudopotential

The new group were among the first scientists to design their calculations to take optimal advantage of the full power of the electronic computer. Their computational strategy sought to combine and modify the wave functions of the valence electrons directly, while systematically and consistently discarding the atomic cores. They realized that Herring's conclusions about energy cancellation in the core region of the atom are better applied by assuming them at the outset than by applying them sequentially to the calculation of the core-electron wave functions. The new strategy considers a valence electron to be moving under the influence of an "atom" called a pseudoatom, which carries a positive charge that is small compared with the nuclear charge. The charge of the pseudoatom is the net charge remaining on the nucleus after the negative charge of the core electrons has been canceled.

It is mathematically convenient, particularly when working with a digital computer, to determine the electron wave function by considering how the momentum of the electron wave changes when it is scattered by a pseudoatom. In classical physics the momentum of a particle is the product of its mass and its velocity; the idea that an analogous concept can be defined for a wave may seem strange. Nevertheless, in wave mechanics the momentum of a wave is given by the direction of the wave's motion and by the wave number, or the number of oscillations per unit distance.

In practice the scattering of the electron is described in terms of the scattering of each of its individual waves or Fourier components. According to a theorem first proved by the French mathematician Jean Baptiste Joseph Fourier, the shape of almost any periodic function can be generated by superposing simple sine and cosine waves of various amplitudes and frequencies and allowing them to interfere with one another [*see illustration on page 88*]. Conversely, a complicated waveform can be resolved into sine and cosine components. Mathematical operations can be carried out on the components and the new components can be reassembled to synthesize a new wave. The shape of the latter wave may be difficult to express mathematically by any other means.

Unless a given Fourier component is scattered entirely in a single direction, the conservation of charge requires that any scattered wave have a smaller amplitude than the incident wave. The ratio of the amplitude of the scattered wave to the amplitude of the incident wave is called the scattering strength. The quantity now employed by quantum theorists to solve the wave equation is a measure of the scattering strength of one atomic core with respect to a single Fourier component of one valence-electron wave. The quantity is called the pseudopotential of the electron.




**OPTICAL REFLECTIVITY is a measure of the response of a material to incident radiation. It is the proportion of the energy carried by radiation of a given wavelength that is reflected by the material. In the diagram the experimental values of the reflectivity of crystalline silicon (*black curve*) are plotted over a broad band of incident wavelengths or energies. The theoretical reflectivity of silicon, as predicted by the pseudopotential theory, is plotted on the same set of axes (*colored curve*). Although the agreement between the two graphs is striking, the resolution is not fine enough to convey dozens of detailed features of reflectivity and other optical properties that were first suggested by such calculations and later discovered experimentally.**

## The Fermi Momentum

At first it might seem that the scattering strength, and hence the pseudopotential, should depend on the momentum of both the incident electron wave and the scattered one. If the waves were classical (non-quantum-mechanical) waves, or if the energy of the electrons were high, this would be the case. The quantum nature of the system and the relatively low level of electronic excitation in solids, however, greatly reduces the amount of information needed to characterize the scattering of valence electrons by pseudoatoms.

According to the uncertainty principle of Heisenberg, the position and the momentum of a particle cannot be determined simultaneously with absolute precision. The more accurately the position of the particle is known, the more uncertainty there must be about its momentum, and vice versa. When valence electrons are confined to a finite volume, as they are in a solid, it is possible to estimate their average density, and so each electron is confined, on the average, to a volume of known size. From the uncertainty principle, therefore, one can estimate the uncertainty about the momentum of any electron in the solid: the momentum must lie between zero and a maximum momentum called the Fermi momentum, after Enrico Fermi.

Electron waves are also subject to a quantum-mechanical constraint called the Pauli exclusion principle, after Wolfgang Pauli. The principle states

# 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 6,500 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/50,000 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.*

*Seat belts save lives—buckle up.*

**Ford Extended Service Plan** Ford

that only two electrons in any system can have precisely the same momentum. Moreover, the momentum of an electron is quantized: it can take on only certain discrete values. Hence when electrons are confined to a given volume their momenta must be distributed among all the possible discrete momentum states from the lowest or ground state up to the state corresponding to the Fermi momentum. In general it is not possible to know which of the momentum states is occupied by a particular electron. If the electron's state were known precisely, there would be maximum uncertainty about its position; it could be anywhere.

Without knowing the states of individual electrons, however, it is still possible to show that the scattering of an electron wave by a pseudoatom is a relatively rare event. Suppose the valence electrons were confined to a box the size of the solid, but the pseudoatoms were removed from the box. The momenta of the electrons would be distributed among all the momentum states between the ground state and the Fermi momentum, but the electron waves would not be scattered by any atom.

Suppose now the pseudoatoms are reintroduced into the box. The electron waves are scattered by the pseudoatoms and the standing wave pattern adjusts to minimize the energy of the system in the new configuration. The scattering of an electron wave, however, changes its direction and so changes the momentum of the electron. Since all the momentum states between the ground state and the Fermi momentum are fully occupied, the only way the momentum of the electron wave can change is to shift from an occupied state below the Fermi momentum to an unoccupied state above the Fermi momentum.

## Quantized Momentum Change

In a solid the magnitude of an electron's momentum can be changed only slightly by scattering: outside the core of the pseudoatom the direction of the electron wave can be greatly altered, but the rate of oscillation per unit distance is little changed. Thus the only momentum changes that are important to scattering outside the core are the relatively small ones associated with changes in direction, namely those that excite the electron wave from a momentum state slightly below the Fermi momentum to a state slightly above the Fermi momentum. Indeed, it turns out that the scattering strength for a single pseudoatom need not consider at all the initial and final directions of the wave. Instead the scattering strength is primarily dependent on the angular difference between the two directions. The dependence on only one quantity rather than two can greatly reduce the complexity of the computations.

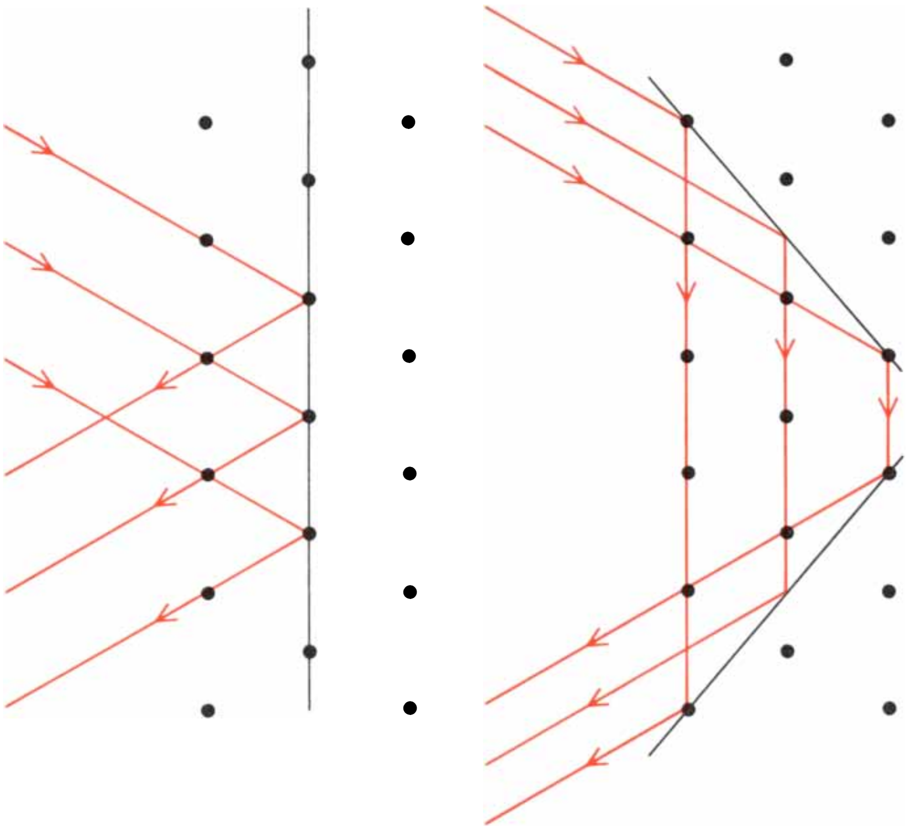
Both the initial and the final momenta of the electron wave can be represented as vectors, that is, quantities with both a magnitude and a direction. Each vector has a magnitude nearly equal to the Fermi momentum, and the difference in their directions can vary between zero and 180 degrees. The change in the momentum of the electron wave when it is scattered can also be represented by a vector. Its magnitude varies from zero to two times the magnitude of the Fermi-momentum vector [*see illustration on page 89*]. Hence the scattering strength outside the pseudoatom must be determined only for momentum changes between zero and twice the Fermi momentum. Once again there is a great gain in computational simplicity. The variation in the scattering strength with the change in momentum is a convenient mathematical means of representing the shape of the pseudoatom in space.

What is the relation between the change in momentum of the electron wave and the scattering strength of the pseudoatom inside the atomic core? Because of the Pauli principle the valence electron cannot occupy any of the energy states filled by core electrons. Accordingly there are nodes in the wave function of a valence electron wherever the charge density of the core electrons is at a maximum. Thus the valence-electron wave function oscillates more rapidly in the core region, which implies that its momentum is increased. In the core there are large changes in the momentum of the electron wave.

The energy of the valence electron in the core, on the other hand, is approximately zero. This fact can be reflected in the pseudopotential by smoothly extrapolating the energy to zero for values of the change in the momentum that correspond to the interior of the pseudoatom. Physically this procedure is equivalent to assuming that however the incident electron wave is altered in the core region, the amplitude of the altered wave outside the core is zero: the valence electron wave is little scattered by its interaction with the atomic core.

## Scattering in Crystals

The pseudopotential is particularly appropriate for expressing the structure of simple crystalline solids, such as elemental metals and semiconductors. In-



**REGULAR ARRANGEMENTS of atoms in a crystal scatter electron waves as if the waves were reflected from planes passing through the atoms. In certain atomic configurations the combined scattering from two or more planes can reinforce the scattering from a single plane. Among elements whose constituent atoms are arranged in tetrahedrons, such as silicon, the reinforcement of the kind of scattering diagrammed at the left ($q_1$ scattering) by the kind of scattering diagrammed at the right ($q_3$ scattering) can lead to strong covalent bonding between the atoms. Such bonds prevent the electrons in silicon from drifting throughout the material, as they do in an electrically conducting metal. Hence silicon is a semiconductor and not a metal.**

Niagara Falls, Ontario/Canada

# *Feel the thunder!*

Come wonder at Niagara from underneath the falls. And we've only just begun. For just beyond, we've sand dunes and blue lagoons and patchwork fields and orchards, and farther still, white water rapids, green rolling ranges, miles of pineclad granite isles, and a rail excursion through glorious gorges or all the way north to Arctic tides. Out of this world! Call TOLL FREE 1-800-828-8585 or from New York State 1-800-462-8404 or write: Ontario Travel, Queen's Park, Toronto M7A 2E5, Ontario/Canada.

ONTARIO yours to discover!


# THE $149.95 PERSONAL COMPUTER.

**Introducing the Sinclair ZX81.**

If you're ever going to buy a personal computer, now is the time to do it.

The Sinclair ZX81 is the most powerful, yet easy-to-use computer ever offered for anywhere near the price: only $149.95* completely assembled.

Don't let the price fool you. The ZX81 has just about everything you could ask for in a personal computer.

**A breakthrough in personal computers.**

The ZX81 is a major advance over the original Sinclair ZX80 – the first personal computer to break the price barrier at $200.

In fact, the ZX81's 8K extended BASIC offers features found only on computers costing two or three times as much.

Just look at what you get:

- Continuous display, including moving graphics
- Multi-dimensional string and numerical arrays
- Mathematical and scientific functions accurate to 8 decimal places
- Unique one-touch entry of key words like PRINT, RUN and LIST
- Automatic syntax error detection and easy editing
- Randomize function useful for both games and serious applications
- Built-in interface for ZX printer
- 1K of memory expandable to 16K
- 164-page programming guide and operating manual

The ZX81 is also very convenient to use. It hooks up to any television set to produce a clear 32-column by 24-line display. It comes with a comprehensive 164-page programming guide and operating manual designed for both beginners and experienced computer users. And you can use a regular cassette recorder to store and recall programs by name.

**Order at no risk.****

We'll give you 10 days to try out the ZX81. If you're not completely satisfied, just return it to Sinclair Research and we'll give you a full refund.

And if you have a problem with your ZX81, send it to Sinclair Research within 90 days and we'll repair or replace it at no charge.

**Introducing the ZX81 kit.**

If you really want to save money, and you enjoy building electronic kits, you can order the ZX81 in kit form for the incredible price of just $99.95.* It's the same, full-featured computer, only you put it together yourself. We'll send complete, easy-to-follow instructions on how you can assemble your ZX81 in just a few hours. All you have to supply is the soldering iron.

**A leader in microelectronics.**

The ZX81 represents the latest technology in microelectronics. More than 10,000 are sold every week. In fact, the ZX81 is the fastest selling personal computer in the world.

We urge you to place your order for the ZX81 today.

**To order.**

To order, simply call toll free. Or use the coupon below. Remember, you can try it for 10 days at no risk.** The sooner you order, the sooner you can start enjoying your own computer.

**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. Phones open 24 hours a day, 7 days a week. Have your MasterCard or VISA ready.

These numbers are for orders only. If you just want information, please write: Sinclair Research Ltd., 2 Sinclair Plaza, Nashua, NH 03061.

*Plus shipping and handling. Price includes connectors for TV and cassette, AC adaptor, and FREE manual.
**Does not apply to ZX81 kits.

**NEW SOFTWARE:** Sinclair has published pre-recorded programs on cassettes for your ZX81. We're constantly coming out with new programs, so we'll send you our latest software catalog with your computer.

**ZX PRINTER:** The Sinclair ZX Printer will work with your ZX81. It will be available in the near future and will cost less than $100.

**16K MEMORY MODULE:** Like any powerful, full fledged computer, the ZX81 is expandable. Sinclair's 16K memory module plugs right onto the back of your ZX81. Cost is $99.95, plus shipping and handling.

# sinclair

## To order call toll free: 800-543-3000

| Ad Code 06SA | Price* | Qty. | Amount |
|---|---|---|---|
| ZX81 | $149.95 | | |
| ZX81 Kit | 99.95 | | |
| 16K Memory Module | 99.95 | | |
| Shipping and Handling | 4.95 | | $4.95 |
| | | TOTAL | |

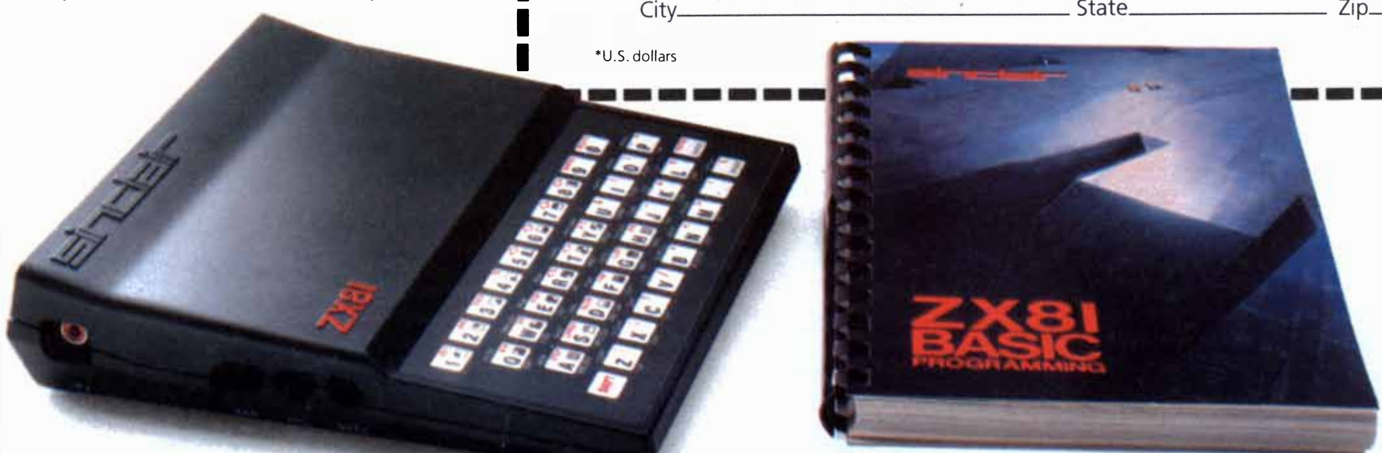
**MAIL TO:** Sinclair Research Ltd., One Sinclair Plaza, Nashua, NH 03061.

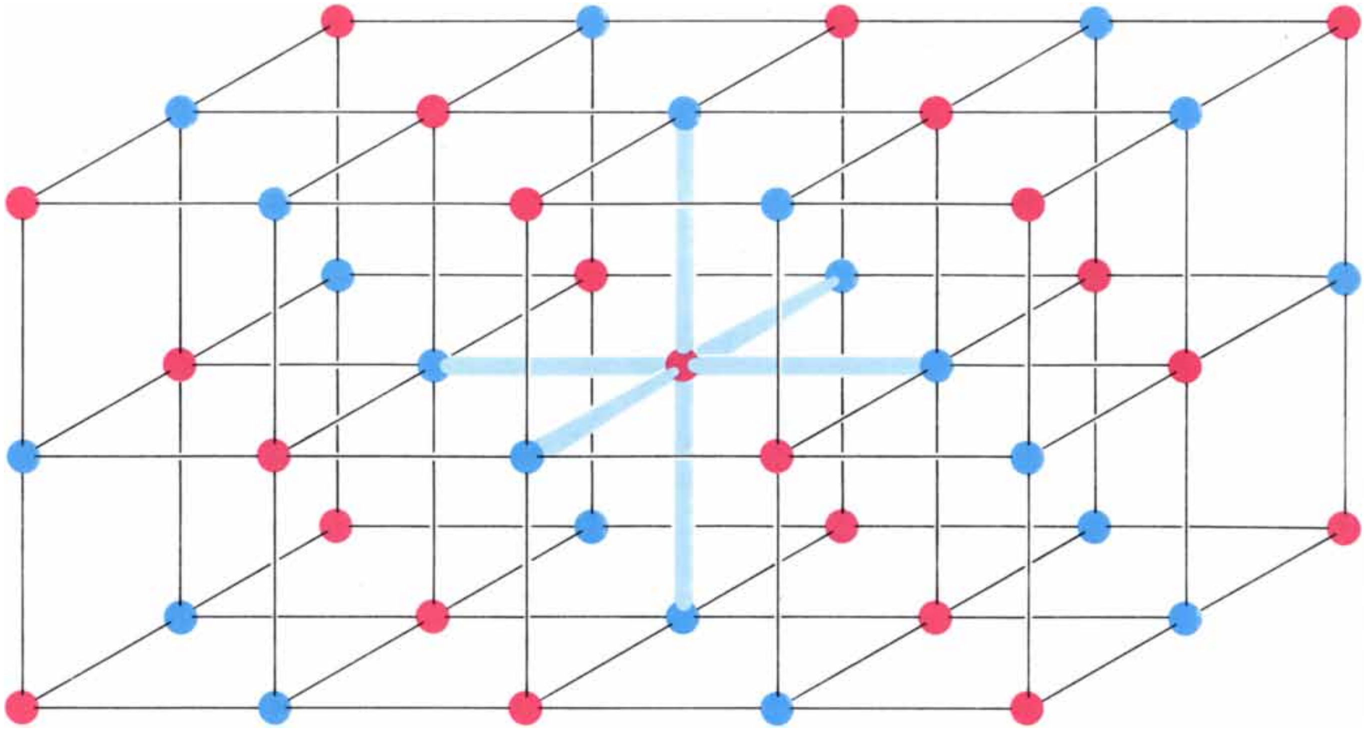
Name____________________

Address____________________

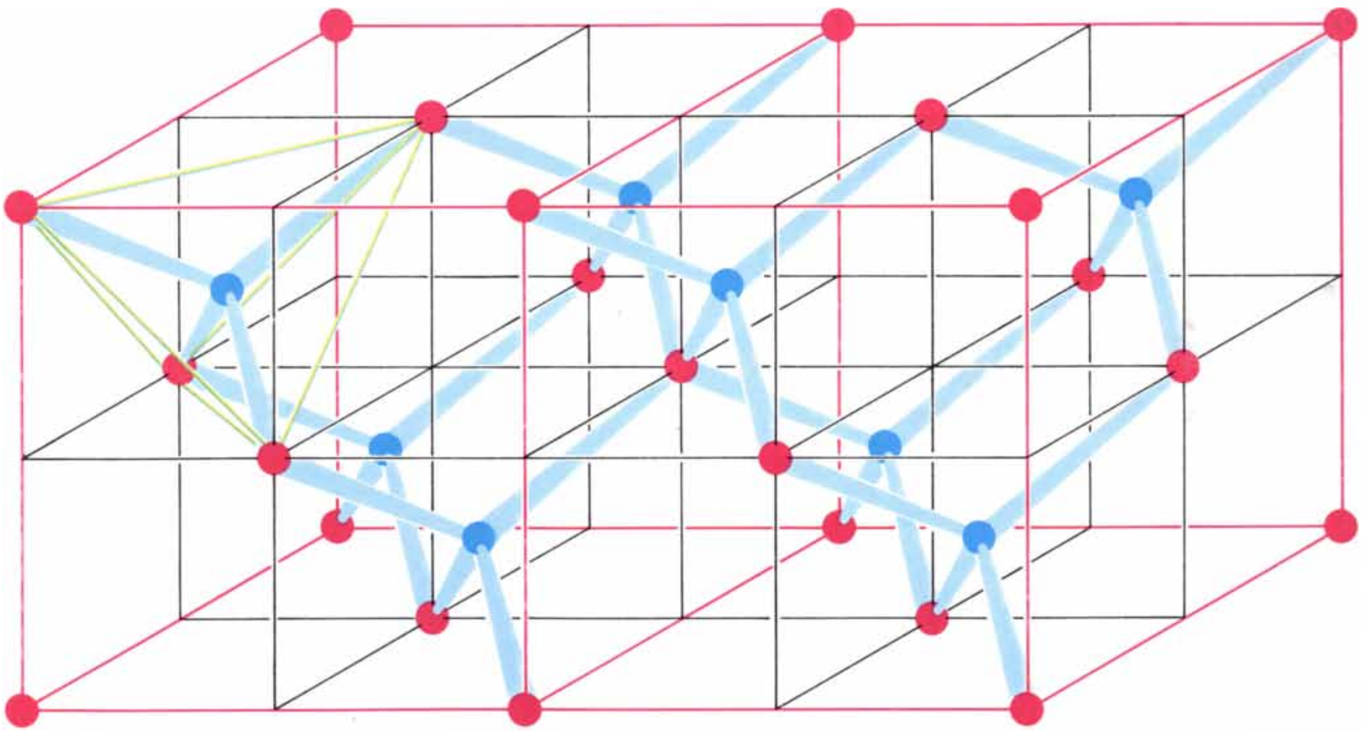
City__________ State__________ Zip__________

*U.S. dollars





**SIXFOLD COORDINATION** is the lattice configuration assumed by binary compounds such as sodium chloride (table salt) in which atomic bonding is caused by the transfer of electronic charge from one atom to another. The charge transfer results in attraction between two atoms with opposite charges. Each atom in the crystal is bonded to six nearest neighbors. The bonds are called ionic bonds.



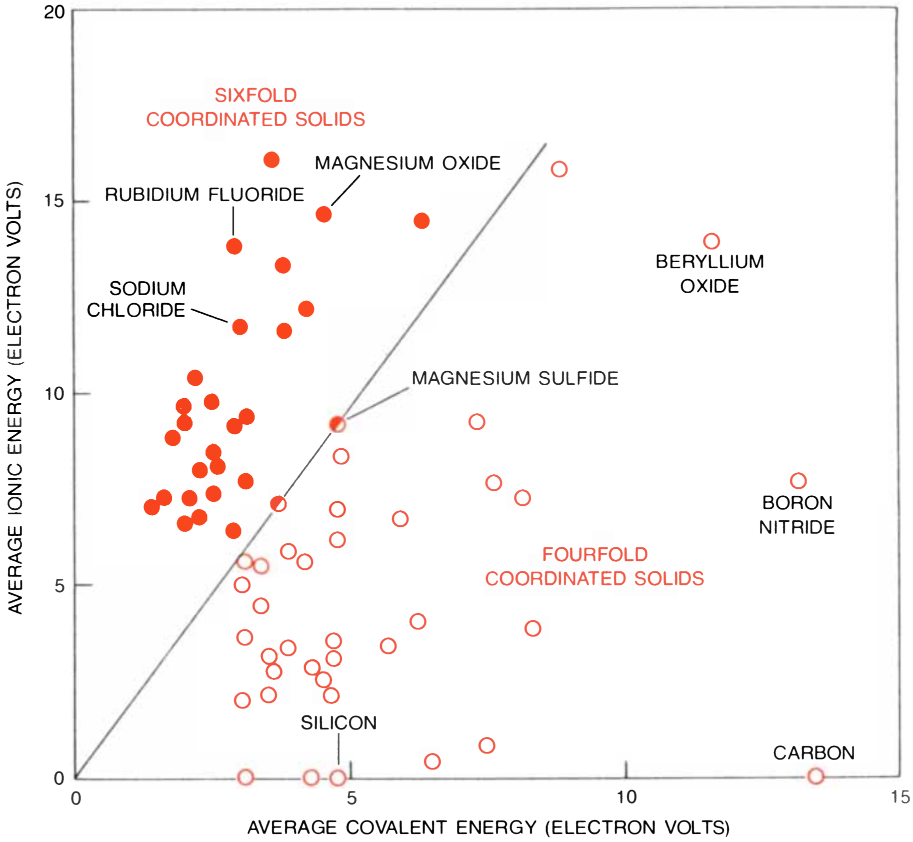
**FOURFOLD COORDINATION,** or tetrahedral coordination, is the lattice configuration of binary compounds such as gallium arsenide that form covalent bonds. In covalent bonding the transfer of electronic charge between two atoms is not sufficient to form ionic bonds, and the charge builds up instead in the space between the atoms. Each blue atom in the crystal (representing one of the elements in the compound) lies at the center of a tetrahedron and is bonded to four red atoms (representing the second element in the compound) at the vertexes of the tetrahedron. One such tetrahedron is outlined in green. Similarly, each red atom lies at the center of a tetrahedron on whose vertexes are four blue atoms. The structure can be defined by a lattice of cubes; the red atoms are found at the centers of the faces and at the vertexes of the red cubes, whereas the blue atoms are found at the centers of half of the black cubes, into which the red cubes are divided.

deed, it is sometimes easier to determine the pseudopotential for a crystal than it is for an isolated atom. The scattering of electron waves in a crystal is extremely sensitive to the angle between the incident and the scattered waves. As we noted above, a plane of atoms scatters an incident wave as if the plane were a half-silvered mirror. In a crystal the orderly arrangement of the atoms generates parallel planes in certain orientations.

When a wave front meets a plane of atoms, part of the wave is reflected and part is transmitted to the next parallel plane. Part of the transmitted wave is then reflected from the second plane and is transmitted back through the first plane, where it interferes with the first reflected wave [*see illustration on page 90*]. If the wavelength, the spacing between the two planes and the angle between the planes and the incident wave are such that the difference in path length between the two waves is a whole number of wavelengths, the interference is constructive. When many planes are arranged in parallel, the incident-wave angles required for constructive interference are sharply defined; for all other angles the interference is destructive.

The phenomenon is called Bragg diffraction, after the British physicists Sir William Henry Bragg and his son Sir William Lawrence Bragg, who investigated the scattering of X rays by crystalline solids. The critical angles for Bragg diffraction correspond to critical values of the change in momentum when an electron wave is scattered. The critical scatterings, which depend only on the relative orientation of the various sets of parallel atomic planes, are the only scatterings of any importance for determining wave functions in a crystal. In a historical context such a simplification is remarkable; in aluminum, for example, only two values of the critical scattering are needed to describe the wave functions of some $10^{23}$ valence electrons. This is a long way from the many pages of numbers Hartree needed to publish his results on the electrons of a single aluminum atom.

The pseudopotentials, or scattering strengths, that correspond to the critical scatterings can be expressed as differences between electronic energies. The energies can be determined experimentally, just as they are in gases, by passing light through a thin slice of the crystal and observing the spectral bands, or colors, of the light that is transmitted. Similar information can be obtained by reflecting light from thick crystals. In this way the values of the pseudopotential are experimentally calibrated for all the critical scatterings for which the momentum change is less than or approximately equal to twice the Fermi momentum. The valence-electron wave functions for the crystal can then be calculated. The resulting wave functions are accurate enough to serve as a basis for perturbation studies that converge rapidly to solutions of the wave equation for many of the structural phases the solid can assume.




**OCTET SOLIDS, in which eight valence electrons are available for bonding from each pair of atoms, are plotted according to the covalent energy and the ionic energy that can be defined for each bond from the atomic pseudopotentials. Ionic solids with sixfold coordination are represented by colored dots; covalent solids that have fourfold tetrahedral coordination are represented by colored circles. Two solids that can assume either configuration are represented by half-colored circles. According to the pseudopotential theory, when the ratio of the ionic energy to the covalent energy of a bond is greater than the ratio defined by the slope of the black line, the configuration of the atoms in a solid changes from fourfold to sixfold coordinated. The graph shows that the theory predicts the atomic configurations of more than 60 octet solids.**

## Elemental Solids

In a metal such as aluminum there are two critical momentum changes caused by scattering; they are designated $q_1$ and $q_2$. The wave functions calculated from these scatterings show that the aluminum atoms minimize their energy by forming a tightly packed solid in which each atom has 12 nearest-neighbor atoms. In a semiconductor such as silicon, on the other hand, each atom lies at the center of a tetrahedron and has only four nearest neighbors. The lower density of the tetrahedral structure gives rise to a third important set of parallel atomic planes and so to a third critical scattering $q_3$. The pseudopotential corresponding to $q_3$ in silicon is relatively large and greatly enhances the total electron scattering. The resulting change in the electron wave functions tends to concentrate the electronic charge along the tetrahedral bonds. It is traditional to think of each atom in such a bond as sharing two valence electrons, and the bond is therefore called covalent.

A comparison of the relative magnitudes of the pseudopotential at $q_1$ and $q_3$ for a given material is a sensitive predictor of the bonding configuration the material will assume. The comparison can be more straightforward than the alternative explanations for the bonding differences offered by structural chemistry. In a single column of the periodic table the elements having four valence electrons form structures that are either tetrahedral, semiconductive solids or close-packed, metallic solids. Carbon, silicon and germanium are all tetrahedral and semiconducting; lead is a close-packed metal. Tin has two stable forms: it is a white metal at high temperature and a gray, tetrahedral semiconductor at low temperature. The existence of two crystalline phases of tin has been particularly difficult to explain in terms of structural chemistry. The problem can be resolved in simple terms in the language of the pseudopotential theory.

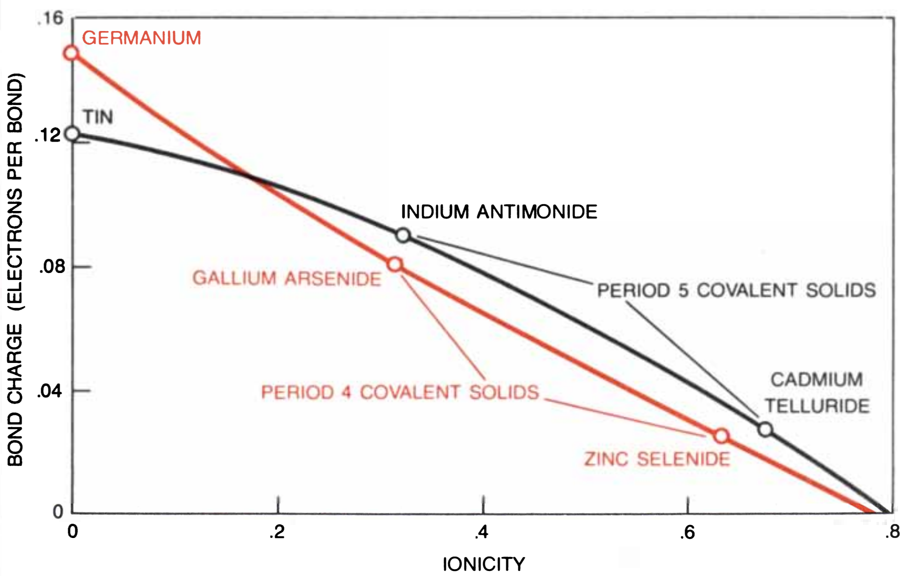
Most of the cohesion of the covalent bond in crystals having a tetrahedral structure arises from the strength of the critical scattering $q_1$. Because of the geometric relation between the parallel planes that cause $q_1$ scattering and the planes that cause $q_3$ scattering, however, two successive $q_3$ scatterings have the same effect as a single $q_1$ scattering [*see illustration on page 94*]. The strength of the bond therefore depends on the strength of the $q_1$ scattering plus the strength of the square of the $q_3$ scattering. If the effective overall scattering strength is large enough, the atoms of the material form covalent bonds.

In carbon, silicon and germanium the pseudopotentials that correspond to the $q_1$ scattering and to the square of the $q_3$ scattering are both positive, so that the total scattering strength is large enough to produce covalent bonding. In lead, however, the two components of the pseudopotential have opposite signs and tend to cancel. The scattering is not strong enough to form a covalent bond, and so lead is metallic. In tin the value of the pseudopotential corresponding to $q_1$ is approximately zero. Hence the scattering is so close to the exact critical value needed for covalent bonding that temperature differences can apparently determine the interatomic structure of the material.

The account given by pseudopotential theory for the structures of gray and white tin is typical of the explanatory simplicity of the theory. Pseudopotentials and elementary mathematical expressions involving them have proved to be both readily comprehensible and so successful quantitatively that the theory has become a new language for solid-state theorists. The range of material properties discussed in the pseudopotential language is by now so wide that it has probably become uniquely suitable for describing the quantum mechanics of electrons in materials.

## Binary Compounds

Pseudopotentials have now been constructed for most of the elements in the periodic table. They also readily account for bonding in alloys and binary solid compounds. In such a compound the atoms of the different elements attract valence electrons with different strengths, and so the valence electronic charge is shifted toward the atoms of one of the elements. A bond is then formed partly because of electrostatic attraction between the two oppositely charged atoms. The degree to which this kind of bonding takes place is called the ionicity of the bond. The remainder of the energy stored in the bond is covalent or metallic in character. It is caused, just




**IONICITY of a bond is a measure of the total energy of the bond that can be attributed to ionic bonding. The pseudopotential theory makes it possible to compute the bond charge, or the electronic charge in excess of a background atomic charge, that is present between two covalently bonded atoms. In the diagram the bond charge is plotted against the ionicity of the bond for six covalently bonded solids. For the elemental solids germanium and tin there is no charge transfer because there is no difference in the affinity of two identical atoms for valence electrons. Hence the ionicity is zero and the bond charge is maximum. When binary compounds are formed from elements to the left and right of germanium and tin in the periodic table, the ionicity of the bond increases and the total bond charge decreases. If the bond charge is extrapolated to zero for the covalent solids made up of elements from a single period of the periodic table, the corresponding value of the ionicity is approximately .79. The value is in good agreement with the critical value assumed by the ionicity along the black line in the illustration on the preceding page. Thus pseudopotential theory makes the intuitively satisfying prediction that the transition from covalent to ionic bonding takes place when bond charge becomes zero.**

as it is in a bond between two atoms of the same element, by the tendency of the charge density to be distributed in the way that minimizes the total energy of the solid.

The ionicity of a crystalline bond can be described by the differences in the pseudopotentials of the two atoms, evaluated at certain critical momentum changes caused by scattering from lattice planes. (As a check on the consistency of the argument, consider the circumstance in which the bond is formed between two atoms of the same element. In this case one can readily see that when the difference between the pseudopotentials of the two atoms is zero, the ionicity of the bond is also zero: no net shifting of the valence electrons can take place.) The strength of the covalent or metallic part of the bond can be described by the sum of the pseudopotentials of the two atoms.

Pseudopotential calculations of the electron wave functions in partially ionic and partially covalent solids show that the greatest valence-electron charge density lies along the bond between the two atoms, rather than close to either atom. One can therefore define a charge called the bond charge: it is the amount of charge in the bonding region in excess of the charges obtained by superposing the charge densities of the isolated atoms that make up the bond.

It turns out that whenever the bond charge predicted by the pseudopotential theory is greater than zero, a binary compound with an average of four valence electrons per atom forms covalent bonds having the tetrahedral structure of diamond or silicon. When the bond charge given by the pseudopotentials becomes zero, however, the tetrahedral structure no longer forms. What forms instead is an ionic lattice in which each atom has six nearest neighbors instead of four [*see illustrations on page 98*]. The structures of more than 60 binary compounds of the form $A^N B^{8-N}$ (the superscripts refer to the number of valence electrons in each atom) have been correctly predicted from the pseudopotential theory in terms of the covalent and ionic parts of the bond energy.

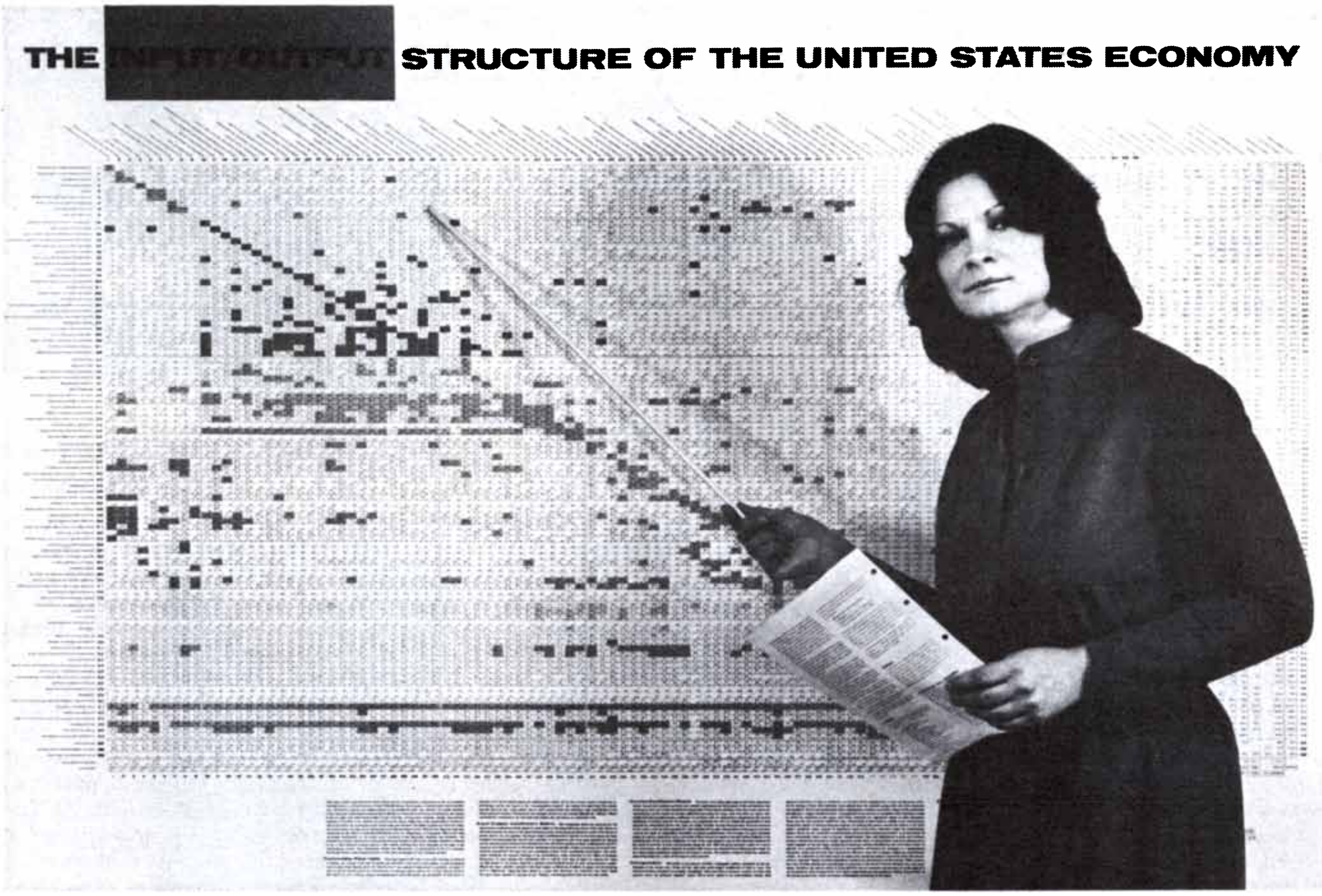
## Recent Theoretical Advances

Perhaps the most remarkable aspect of pseudopotential theory is its capacity for steady growth through increasing physical and mathematical sophistication. The cumulative development has enabled pseudopotential theorists to keep basic physical issues as well as technical details in mind in describing the quantum structure of materials. By emphasizing conceptual simplification the theory has stayed ahead of the rapid advances in numerical analysis associated with the development of the computer. We can do no more than mention a few of the theory's recent achievements.

It has been the most important theoretical tool in investigations of the properties of solid surfaces. At surfaces and interfaces the patterns of regular bonding are profoundly disturbed and the redistribution of the electron charge density that results may affect the bonds several atomic layers deep within the material. Joel A. Appelbaum and Donald R. Hamann of Bell Laboratories, James R. Chelikowsky, Steven G. Louie and Michael Schlüter of the University of California at Berkeley and Kosal Pandey of the Thomas J. Watson Research Center of the International Business Machines Corporation have described the structure of broken-bond states and other electronic states at the surfaces of materials. The properties of interfaces between metals and semiconductors and between two semiconductors have also been explained with the theory. The understanding of such interfaces is important to the design of many electronic devices.

During the last two years Ming Tang Yin and one of us (Cohen) at Berkeley and others have calculated the total energy of semiconducting materials with an accuracy of about .02 electron volt per atom for a variety of densities or volumes. The calculations have provided a more quantitatively detailed understanding of such phenomena as the structural instability of gray and white tin. When such calculations are extended to other materials, the investigator can explore similar phase changes under extreme conditions of temperature and pressure. The interatomic structure of silicon, for example, is tetrahedral at room temperature and pressure, but it assumes a denser metallic form at pressures comparable with those found in the earth's mantle. The calculated pressure at which the phase transition takes place agrees quite accurately with the experimentally determined value.

The pseudoatom approach has made it possible for theorists to discuss the properties of light or heavy elements and covalent, ionic or metallic compounds throughout the periodic table more accurately, more quickly and more confidently than they discussed any one light element 25 years ago. Fifty years ago atomic physics was considered the only area of science where quantum interactions could be well understood. Today the quantum mechanics of materials has become both conceptually and practically a simpler study than the study of the electronic structure of atoms having more than one possible valence-electron configuration. Through worldwide collaboration on pseudoatoms, alchemy, the original black art of materials science, is now well on its way to becoming one of the better developed parts of human knowledge.




# 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. SA/6

Make check payable to Scientific American. Please add sales tax for California, Illinois, Massachusetts, Michigan and New York.

# The Archaeology of Lascaux Cave

*Forty years' work has revealed much about how the great Paleolithic paintings of Lascaux Cave were created. It has also focused attention on hundreds of engravings that rival the paintings in their import*

by Arlette Leroi-Gourhan

The cave site of Lascaux, with its magnificent array of some 600 paintings from the Old Stone Age, was discovered more than 40 years ago. Situated in the Dordogne region of southwestern France, the cave has been closed to all but officially sanctioned visitors for the past 20 years. The closing was part of a conservation effort, fortunately successful, aimed at halting further deterioration of these Paleolithic art treasures. Until recently, however, few people other than specialists have been aware that the mighty paintings of Lascaux are only a part of the wealth of archaeological material discovered in the cave over the past four decades.

Among these discoveries are nearly 1,500 engravings on the cave's walls and ceilings. They were all painstakingly copied by Abbé André Glory, who until his untimely death in 1966 was by far the most indefatigable of the scientific investigators at Lascaux. Glory and others have also excavated the flint tools used to make the engravings, have found the stone lamps that provided light for the artists, have found the palettes that held the painters' pigments and most recently have studied the minerals that were ground into pigments. They have even been able to reconstruct the kind of scaffolding that allowed the painters and engravers to work on rock faces far out of normal reach. These findings and others now enable prehistorians to picture in some detail man's activities at Lascaux as the Old Stone Age entered its final stages some 17,000 years ago.

The discovery of the cave in September, 1940, early in World War II, has often been attributed to the actions of some village boys after a dog had fallen into a deep hole. The facts are somewhat different. A boy from the village of Montignac, a kilometer away, did come on a hole in the ground. It was newly made by the uprooting of a pine tree, but no dog had fallen into it. He and some friends began to test the depth of the hole by dropping stones into it.

Their curiosity further aroused, the boys enlarged the opening and were able to slide down a steep tunnel of wet clay. They landed a good deal farther underground than they had expected. What they saw by the glimmer of a weak flashlight left them amazed. After they had led the local schoolmaster to the site their discovery was reported in the school paper. Specialists in prehistoric studies were told of the find, and a small ramp was dug to give easier access to the cave. The first professional archaeologist to inspect the site was Abbé Henri Breuil.

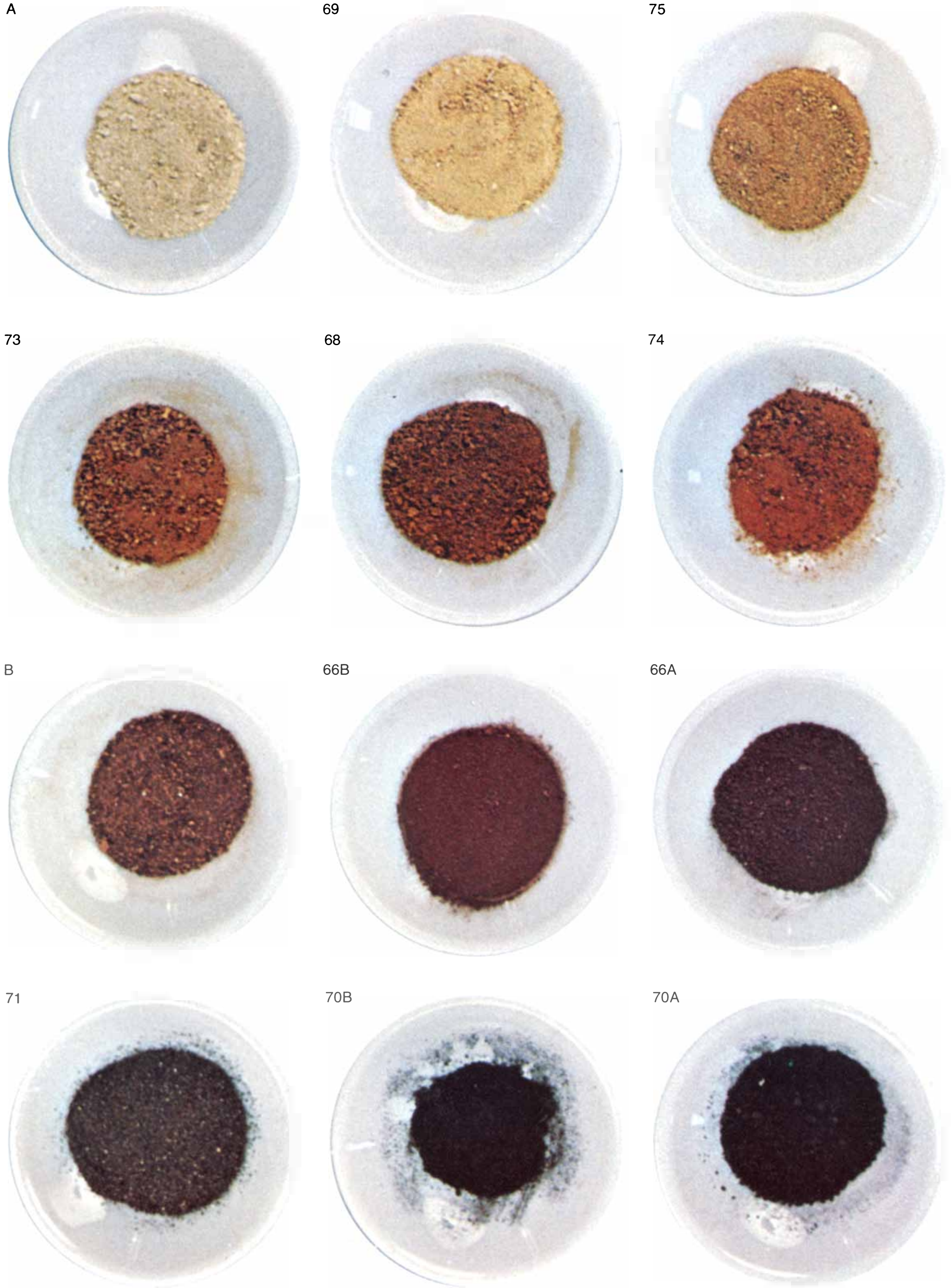
The lower end of the ramp opens into a great chamber some 20 meters long. Its ceiling is covered with immense paintings of bulls; some of the animal figures are more than five meters long. At the far end of this "Hall of the Bulls" open two passages. One passage is straight ahead and the other branches off to the right. The ceiling of the passage straight ahead (known as the Axial Alleyway) is also covered with paintings; they depict deer, horses, wild cattle, ibexes and a bison. Access to the branch to the right (known simply as the Passage) was difficult in the early years. Visitors had at first to go on all fours, but they were soon rewarded by another series of animal paintings. A second branching to the right leads to "the Apse," beyond which lies "the Well," a five-meter depression with a curious group of paintings depicting a man, a bird perched on a staff, a rhinoceros and a charging bison with a horse's head above it. Opening to the left of the Apse is "the Nave." Here and in the long, narrow passage farther along are several more animal paintings. The passage leads eventually through "the Cat's Hole" to "the Alleyway of the Felines," where engravings are found not only of felines but also of several other kinds of animal. This final excursion completes the visitor's tour of Lascaux.

Soon after the war ended the cave was fitted out to accommodate visitors; Lascaux was transformed into an underground museum. Then, after two decades of tourism, it became apparent that the dust, the dampness and the fungi that entered the cave along with the visitors were threatening to destroy this unique Paleolithic gallery. The closing of Lascaux to tourism in 1963 has saved the paintings. In what follows I shall briefly recount what the work of Abbé Glory and others over the past four decades has revealed about the contents of the cave.

How can one determine the age of a prehistoric painting? This question was perhaps the most difficult one that faced the scholars at Lascaux from the first. There is no known way of dating such a painting by itself; one must begin by determining the context of the artists' work. That means looking for traces the artists left behind. To find such traces one must dig, first in order to expose the levels of soil that contain artifacts contemporaneous with the artists' work. Next one must establish the sequence of the layers, since not all the layers from top to bottom will be present in all parts of the site. Even when the stratigraphy is known, however, one has the problem of determining which levels are contemporaneous with the paintings.

At Lascaux, fortunately, this problem did not arise. The first stratigraphic in-

**TWELVE PIGMENTS applied by the painters at Lascaux range from a pale yellow (*A*) to black (*70A*). All the pigments used at Lascaux, with the exception of those incorporating charcoal, consisted of powdered minerals. The application of a single mineral as a pigment, however, was rare; the minerals were more usually mixed together. The powder labeled *66A* is unusually pure, being a mixture of hematite (70 percent), clay (20 percent), quartz (5 percent) and other substances. Powder labeled *70B*, in contrast, is a black containing 40 percent calcium phosphate, 25 percent quartz and 15 percent manganese dioxide, the chief mineral for blacks.**


A
69
75
73
68
74
B
66B
66A
71
70B
70A

vestigations were conducted by Abbé Glory, who excavated 15 cross-section trenches in different parts of the cave. All the evidence of human activity he found proved to come from the same stratum. No further proof that this layer was contemporaneous with the paintings was needed other than the uncovering, along with the other objects in the layer, of the lamps, the stone palettes and the lumps of mineral coloring matter used by the Lascaux painters.

The archaeological remains unearthed in these excavations indicate that although the cave never served for living purposes, those who made the paintings and engravings worked there for many hours at a time, often eating their meals on the spot. It also seems likely that religious rites were held in the cave. In the Well beyond the Apse are the remains of meals, the finest stone lamp found at Lascaux, ornamental seashells and bone lance heads. Among the shells was a particularly beautiful one that had been pierced for stringing as an ornament. It shows grooves worn by the cord that suspended it and bears traces of red ocher. Since the nearest seashore is 200 kilometers away, the shells must have come from at least that distance. The Well seems to have served as a kind of sanctuary.

The stratigraphic investigations at Lascaux, together with analyses of pollen grains in the strata from contemporaneous plants, show that Paleolithic man used the cave for a relatively brief period, perhaps only a few hundred years. This conclusion is consistent with the fact that the stone and bone artifacts excavated in the cave all belong to a single late Upper Paleolithic culture well known in the Dordogne and elsewhere in France: the Lower Magdalenian. It is also in accord with studies of the many Lascaux paintings and the more numerous engravings demonstrating that their style is homogeneous. Hence everything points toward the utilization of the cave for a single short

**STAG'S HEAD, its eye and ear carefully outlined, was engraved on the south wall of the area called "the Apse" at Lascaux Cave. The animal's body was outlined but is not seen in this view. Its exaggerated antlers, also outlined in part, have in addition been painted in black.**

period. Carbon-14 dating of charcoal associated with the painters' lamps places that period at about 15,000 B.C.

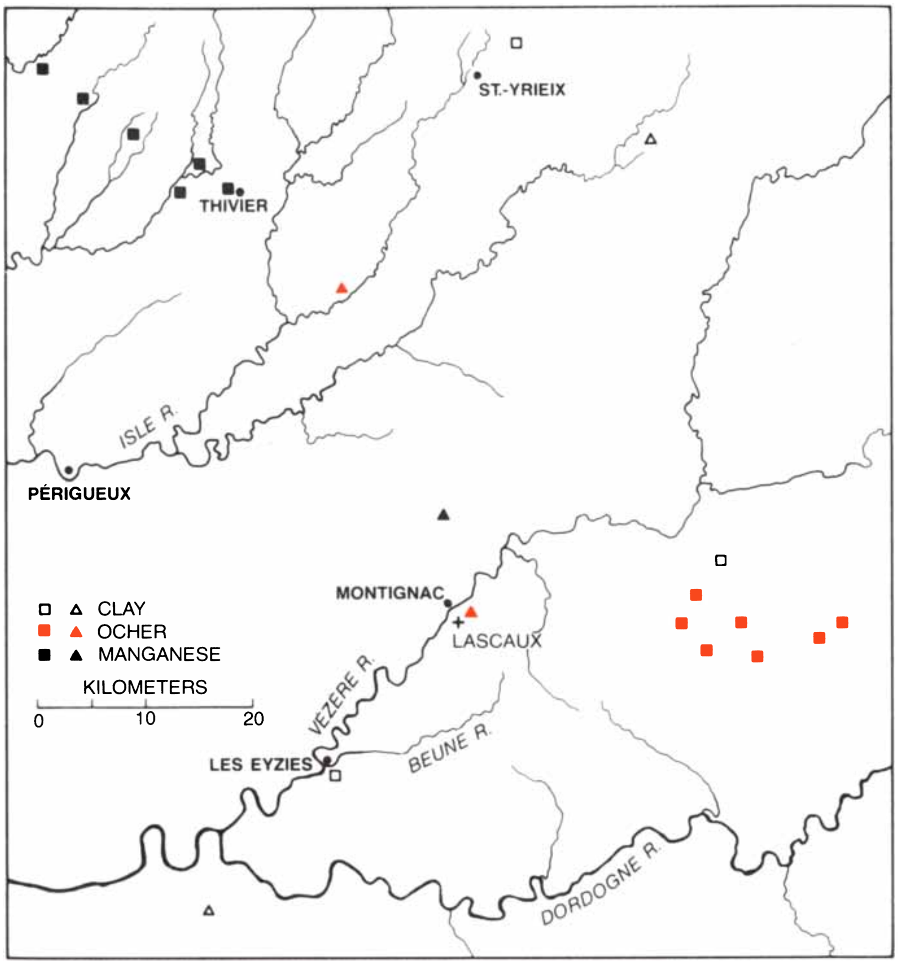
How did the artists do their work? Their first problem must have been the need for artificial light. In Magdalenian times the entrance to the cave was sizable and the roof stood about three meters above ground level. This means that the front of the Hall of the Bulls did receive some daylight. At the back of the hall, however, the natural light would not have been enough to work by, whereas in the Axial Alleyway and all the other galleries artificial light was an absolute necessity.

It is thus not surprising that stone lamps have been found by the score at Lascaux. Some 130 of the total have been carefully studied. The Magdalenian artists did not have to invest much effort in making the kind of lamp most often used; these lamps were simply stones selected from the rocky hillside around the cave mouth that had natural cup-shaped depressions in them. Indeed, some of the stones have depressions so shallow that one may wonder whether they served as lamps at all. The residue of soot, charcoal and ash found in the depressions, however, along with evidence of heating, leaves no doubt about their function.

To better understand how these simple lamps worked, Brigitte and Gilles Delluc of the Musée des Eyzies collected similar flat stones in the vicinity of the cave and tried some experiments. They found that suitable wicks could be made out of dried lichen, moss or twigs of juniper. (Analysis of the charcoal from seven of the lamps had identified juniper wood as one of the wick materials.) The fuel they found most suitable was tallow; in each lamp they put a 50-gram piece of this animal fat. It burned without smoking, gave as much light as an ordinary candle and the 50 grams lasted for about an hour. If one had wanted more illumination, several wicks could have been put in one lamp.

The handsome lamp that was found in the Well was smoothly carved out of red sandstone and decorated with linear incisions. It was found on the floor of the Well, just below the tip of the tail of the rhinoceros in the painting above it. The refinement of the lamp's workmanship and its presence in this putative sanctuary suggest that it had served some purpose other than mere illumination.

The artists' second problem became apparent as soon as they had decided which parts of the cave were suitable for painting. Paintability depended on the condition of the rock surface. For example, in the Axial Alleyway the surface that is sufficiently smooth for painting is three to four meters above the floor and therefore out of reach. The




**LASCAUX CAVE, near the village of Montignac on the Vézère River 40 kilometers east of Périgueux, lies in a region where deposits of clay, manganese oxide and red and yellow ocher are common (*squares*). The closest source of manganese mapped by geological surveys is at Thivier, 40 kilometers to the northwest of Lascaux, and the closest sources of clay, red ocher and yellow ocher are 20 kilometers to the east. Investigators who sought closer sources (*triangles*) found manganese within six kilometers of the cave and both ochers within 500 meters.**

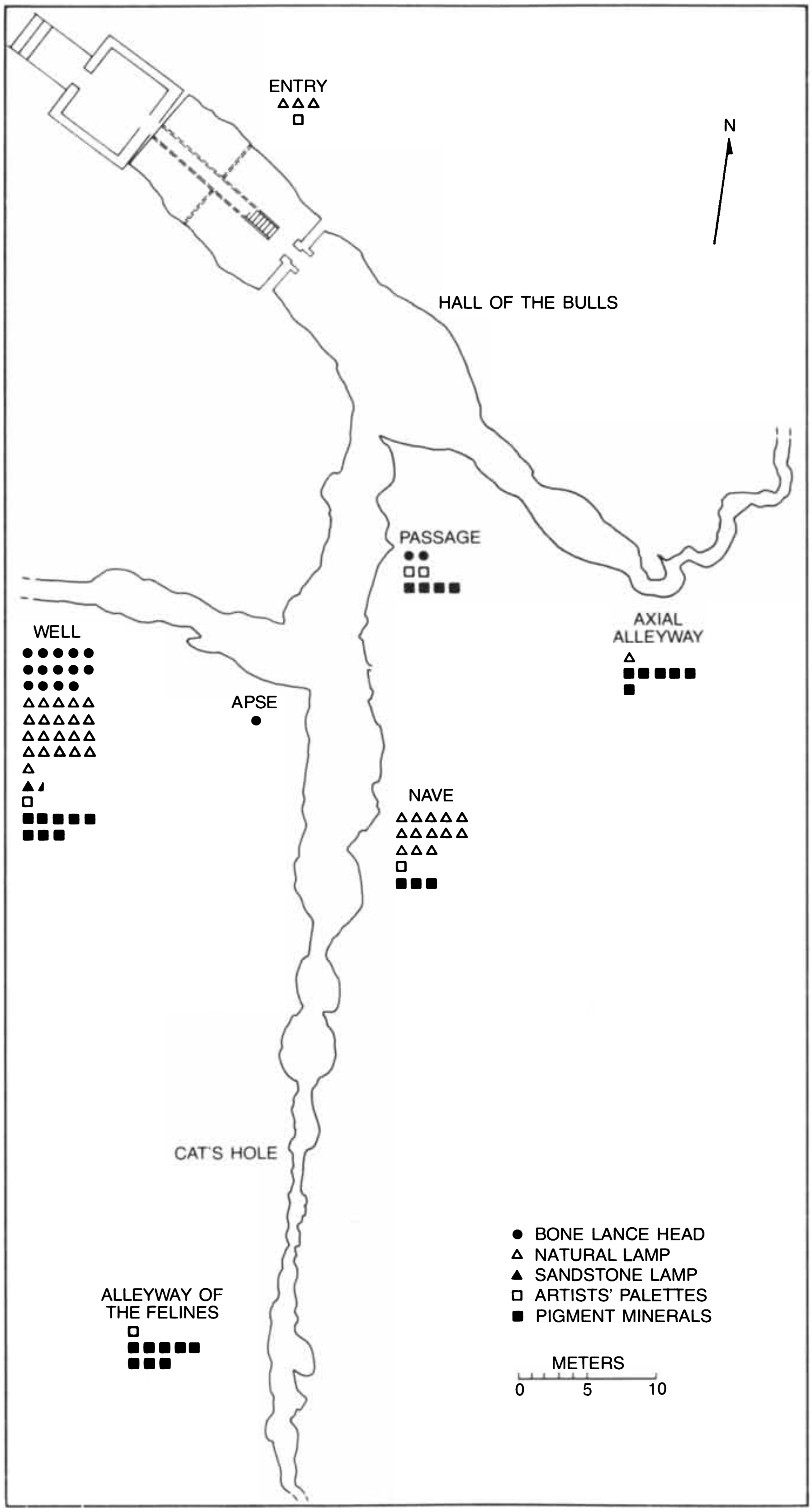
**BITS OF PAINT from Lascaux are seen in these scanning electron micrographs. At the left is a red ocher, enlarged 6,000 diameters; the platelike crystals are hematite ($Fe_2O_3$). At the right is a black manganese, enlarged 7,000 diameters; the needlelike crystals are pyrolusite ($MnO_2$). The study is being done by Pamela B. Vandiver at the Massachusetts Institute of Technology.**

same is true of many of the best surfaces for painting elsewhere in the cave. In the Hall of the Bulls one can imagine the Magdalenians bringing saplings in and propping them up to form a crude scaffolding for the painter to perch on. Traces of more ambitious scaffolding, however, have been found in the Axial Alleyway. From them it has been possible to reconstruct how the painter was given a level platform close to the ceiling of the passage.

The investigators found a series of recesses cut into the rock on both sides of the passage, all on the same level about two meters above the floor. The recesses had been packed with clay, and holes about 10 centimeters deep can be seen in the soft clay packing. The surface of the wall later became covered by a thin coat of the soluble mineral calcite, so that 17 of the indentations were perfectly preserved. Evidently stout tree branches, long enough to span the passage, were fitted into the recesses on each side and were cemented in place with clay. This array of joists could then support a flooring of additional branches that was high enough to give the painters access to the entire ceiling.

Once the painters were settled on their platform they must have been quite comfortable. Judging from the animal bones found on the floor below, they did not even bother to come down when they ate. Animal bones also indicate that a second area in the cave was a popular eating place: a large rock ledge in the Nave near a painting of a black cow.

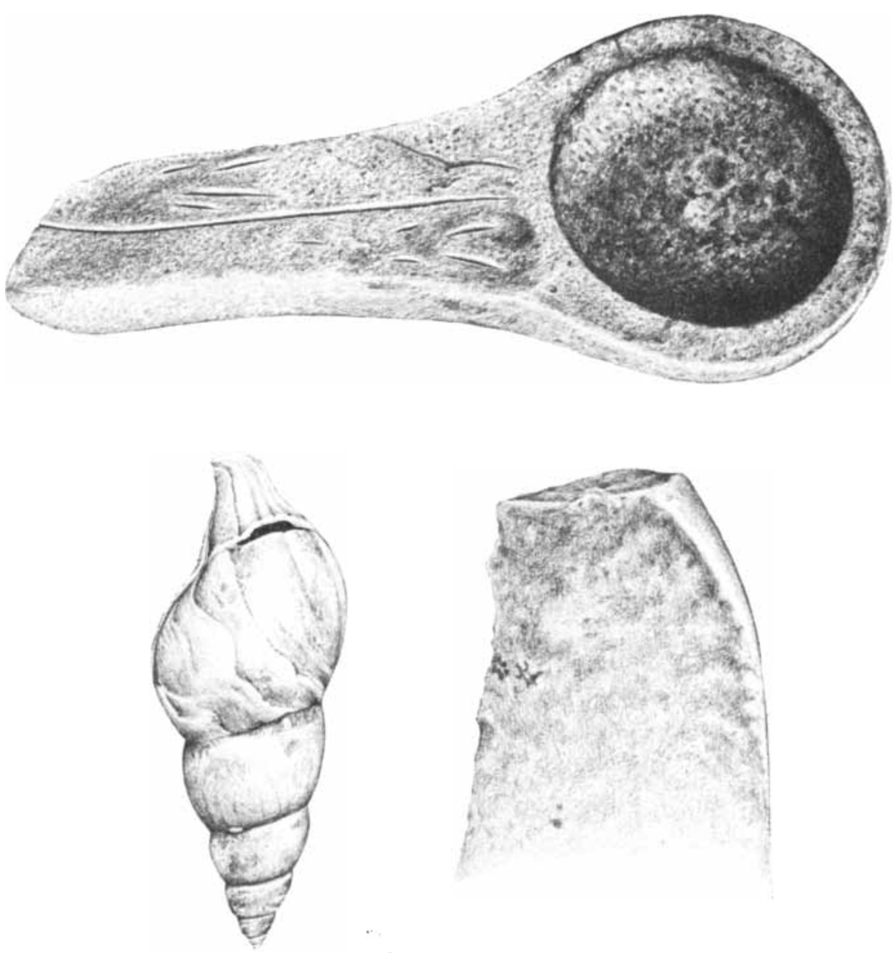
Analysis of bone in both deposits suggests that the artists of Lascaux were partial to the meat of young reindeer, but this may have been less a matter of gastronomic preference than one of what game was available. It is of interest in this connection to note that only one of the animal paintings at Lascaux depicts a reindeer. Facts such as this one help to emphasize the hazards facing those who seek to speculate about the "why" of Paleolithic painting.




**EIGHT PRINCIPAL LANDMARKS in the cave at Lascaux are named on this plan; where some key artifacts were found is also indicated. The richest source of artifacts was "the Well," 55 meters from the entrance to the cave at the left. The Well yielded 21 lamps of unworked stone, one carefully fashioned stone lamp (and a fragment of another), 14 of the 17 bone lance heads found at Lascaux, one of the six artist's palettes and eight of the 29 lumps of minerals the artists powdered to make their pigments. (One hundred and twenty-nine other lumps were found in the cave, but their location was not recorded.) Many stone lamps were also found in "the Nave." Additional minerals were found in the Nave, in "the Passage," in "the Axial Alleyway" and in the remotest gallery, "the Alleyway of the Felines," 100 meters from the entrance.**

The mineral pigments applied by the Magdalenian painters at Lascaux have been the subject of three related studies. First, the late Annette Laming-Emperaire of the École Pratique des Hautes Études in Paris and one of her students, Claude Couraud, have analyzed the raw materials used to produce powdered pigments and have determined where in this part of France the sources of these minerals are. Second, at the suggestion of Aimé Bocquet of the Center for the Documentation of Alpine Prehistory and under the direction of R. Bouchez of the Nuclear Research Center at Grenoble, a group of scholars undertook a technical analysis of the powdered pigments and of the Magdalenian methods of preparing them.

Third, Pamela B. Vandiver, a gradu-



**THREE ARTIFACTS among the many discovered at Lascaux are the carefully worked stone lamp found in the Well (*top*), a seashell stained with ocher and pierced for stringing as a personal ornament (*bottom left*), which was also found in the Well, and the worn, work-polished end of one of the 27 flint engraving tools that have been discovered in the cave (*bottom right*).**

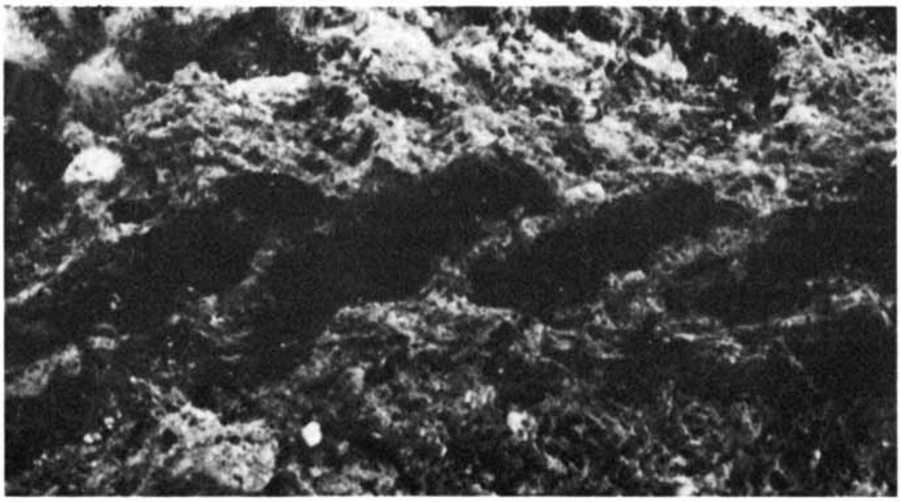
ate student at the Massachusetts Institute of Technology working under the direction of William D. Kingery, has used scanning electron microscopy to study chips of pigment from several of the Lascaux paintings. The chips were furnished by S. Delbourgo of the Musée du Louvre. Vandiver's findings are about to be published; some of the highlights will be summarized below.

In the first of these studies Laming-Emperaire and Couraud had at their disposal a total of 158 mineral fragments Abbé Glory had found at various places in the cave. Many of the fragments show scratch lines and other evidence of purposeful wear. To judge from the fragments, the colors the Magdalenians most prized were blacks (105 fragments), yellows (26), reds (24) and, as a poor fourth, whites (3). There was substantial variation within the blacks, yellows and reds. For example, the blacks ranged from very dark black to olive gray, the yellows from light yellow to reddish and brownish yellow and the reds from brownish and yellowish red to light red and very pale red. Abbé Glory's collection included some crude pigment-stained stone "mortars" and "pestles" one may assume the painters used in preparing their colors. It also included a number of naturally hollowed stones that still held small amounts of powdered pigment.

To determine the sources of the minerals Laming-Emperaire and Couraud first turned to government geological-survey data on the known deposits in the vicinity of Lascaux of manganese oxide (blacks), iron oxide (reds and yellows) and porcelain clay (white). The closest source of manganese, according to geological-survey maps, was 40 kilometers to the north-northwest of the cave and the closest source of clay was 20 kilometers to the east. Red and yellow ocher were available 20 kilometers to the east. Later field work by the two scholars turned up unmapped sources much closer to Lascaux: a source of manganese oxide five kilometers away and one of ocher only half a kilometer away.

Laming-Emperaire and Couraud next turned their attention to the painting methods of the Magdalenians. Experiments demonstrated that the pigments had been prepared by mixing the ground mineral powders with cave water, which has a naturally high calcium content

**REPLICA OF A LASCAUX LAMP, consisting of a fieldstone with a natural hollow, has juniper twigs as a wick and animal tallow as a fuel. Here it illuminates one painting in the cave.**



**PALEOLITHIC ROPE, seven millimeters in diameter, was "fossilized" in the Alleyway of the Felines. Calcite has replaced the plant fibers that had been twisted together to form the rope, but the 30-centimeter fragment preserves details of the twisting. Bone needles are common late Paleolithic artifacts, and it has been assumed they were used to pull some kind of "thread." It has not been clear, however, whether the thread was plant fiber or animal sinew or both. Discovery of the plant-fiber rope at Lascaux suggests that fiber was in common use by 15,000 B.C.**

that ensures good adhesion and great durability. Further experiments showed how the pigments were applied to the rock surface. Laming-Emperaire and Couraud made paintbrushes by macerating the tips of twigs and by binding bison hair in small bundles. They applied powdered pigments directly to damp stone surfaces and also used fragments of ocher as crayons for initially sketching out a painting.

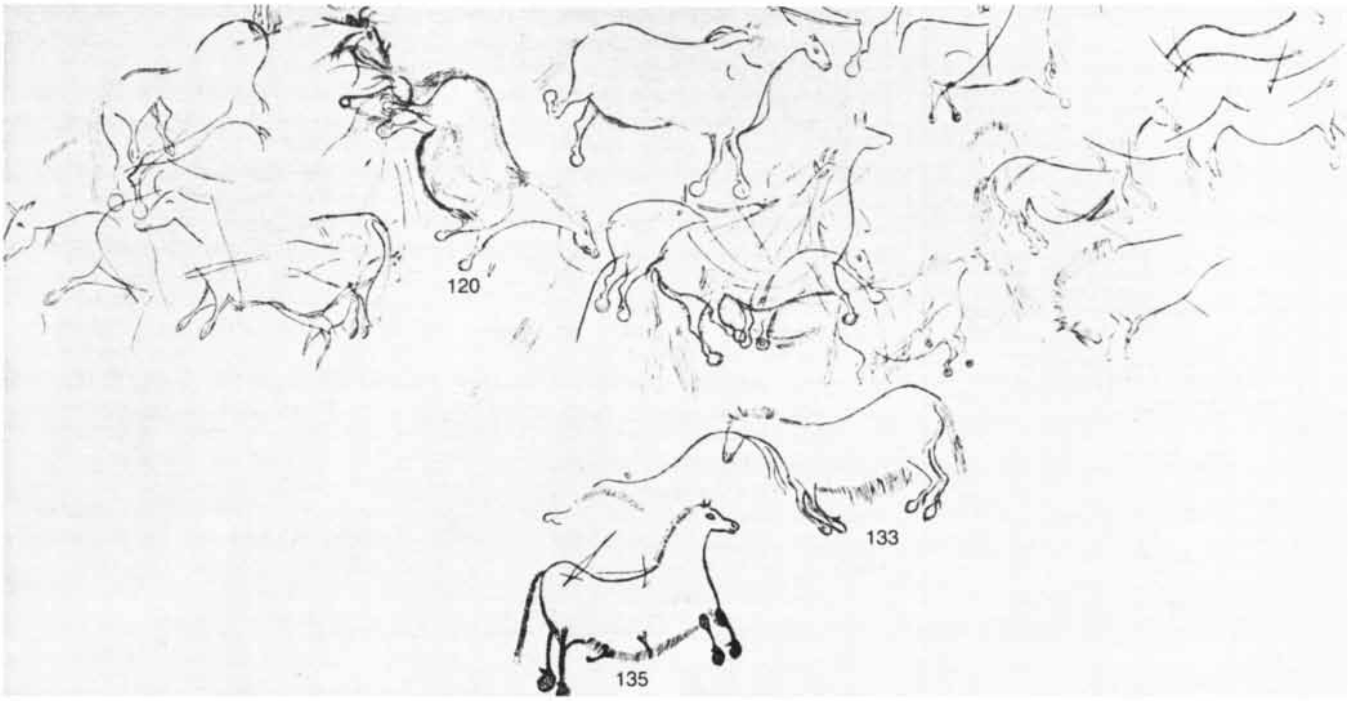
They soon found the best way to hold the paintbrushes, and they demonstrated the efficiency of transferring colors to the stone with a tamping pad. Finally they applied pigments on top of other pigments to determine the effects of such superpositions. In this way they were able to confirm what infrared photography had already suggested: the red cows in the Hall of the Bulls were painted before the great black bulls.

The group in Grenoble analyzed 10 samples of the powdered pigments recovered by Abbé Glory and a further sample of a black pigment removed from the surface of one mortar. For each sample they determined both the chemical elements present and the specific minerals the elements came from. Bits of pigment from a number of Upper Paleolithic paintings had been subjected to similar analysis in the past, but Lascaux is unique in yielding the powders from which the cave artists prepared their pigments. Hence it is not surprising that the work of the Grenoble group has brought out a number of new facts.

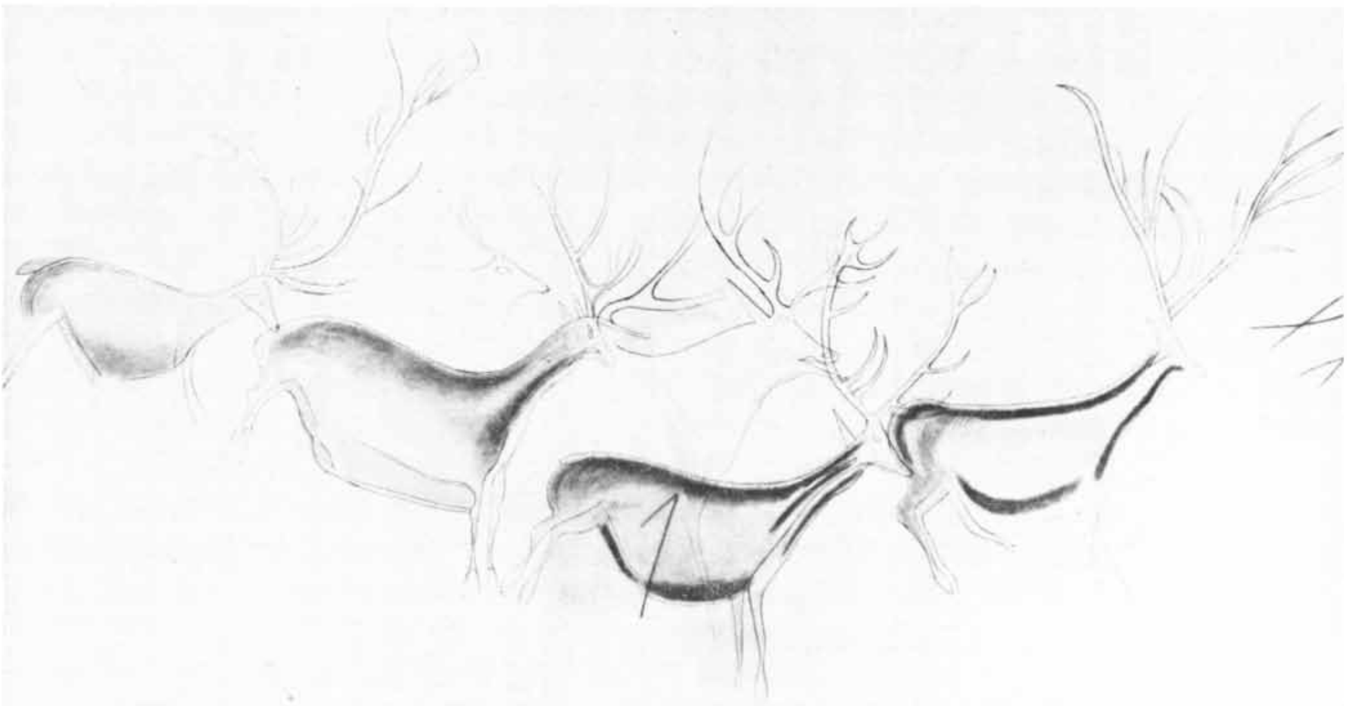
Foremost among them is that the Magdalenian painters not only mixed pigments to achieve the colors and other properties they sought but also processed pigments in unexpected ways. For example, a component of one pigment was calcium phosphate, a substance obtained by heating animal bone to a temperature of about 400 degrees Celsius. The calcium phosphate was then mixed with calcite and heated again, to about 1,000 degrees C., transforming the mixture into tetracalcitic phosphate ($4CaO \cdot P_2O_5$).

A white pigment that at first glance seems reddish because of surface contamination with red ocher proved to be not pure porcelain clay but a mixture of clay (10 percent), powdered quartz (20 percent) and powdered calcite (70 percent). Similarly, whereas manganese oxide was the primary substance of the dark pigments, the black pigment removed from the mortar proved to be a blend of charcoal (65 percent), clay rich in iron (25 percent) and smaller amounts of other minerals, including powdered quartz. Clay had also been added to the powdered red and yellow ochers, in proportions ranging from 20 to 40 percent.

Vandiver's analysis of pigment chips with the scanning electron microscope has yielded more detailed evidence on the preparation and application of pigments at Lascaux. She has also exploited energy-dispersive X-ray analysis, X-ray diffraction and optical petrography to evaluate the technology developed by the Lascaux artists. Such techniques reveal that the microstructure of the red pigments differs from that of the black ones. For example, the hematite crystals of a red ocher pigment are characteristically platy, and the pyrolusite crystals of a black manganese oxide pigment are fibrous or needlelike. An important part of Vandiver's program has been the establishing of a reference collection of natural minerals related to the minerals of the cave-painting pigments. The microstructure and composition of the Lascaux pigments are currently being compared with those of the natural sam-




**ENGRAVED ANIMALS on the west wall of the Passage cover six square meters. A total of 32 animals are shown in part or in their entirety; most of them are horses. Painting has been used to accent three of the horses (Nos. 120, 133 and 135). A few cattle heads and a bull's hindquarters appear to the left of the center. At the center and to the right of the center are the exaggerated tails of three horses. This illustration and the one below are tracings made by Abbé André Glory. His tracings of the engravings covered 117 square meters of paper.**



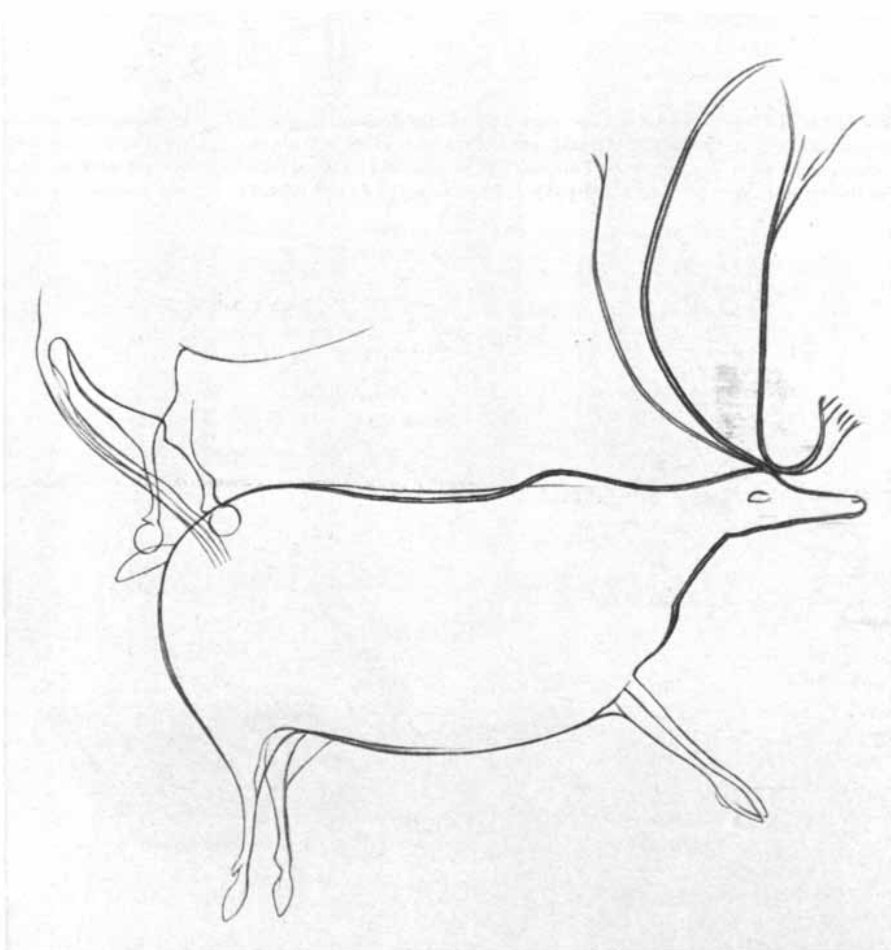
**DETAIL OF SIX DEER, all of them engraved but four fully painted and one with paint accents in outline, forms part of a panel four meters square south of the Well. The sixth deer, the head, horns and back of which are faintly visible above the painted deer second from the left, has the least exaggerated antlers. The dark angular lines visible on the painted deer second from the right may represent an arrow.**

ples in order to learn the nature and extent of Magdalenian processing.

Of Abbé Glory's many contributions to the study of Lascaux none is greater in scope and in importance than his documentation of the 1,500 Magdalenian engravings in the cave. Ranging in complexity from simple straight lines to renderings of entire groups of animals, these incised figures are found in the Passage, in the Nave and particularly in the Apse, where no fewer than 578 of them occupy every available rock surface. It is in these same parts of the cave that 27 flint burins, or engraving tools, have come to light. All of them show identical signs of wear: at the working corner the original sharp angle is worn smooth and round.

The Lascaux paintings, particularly those in the Hall of the Bulls and the Axial Alleyway, were widely reproduced in books and periodicals in the years following World War II. The engravings, however, except for the few that were most clearly visible, attracted little attention. As a result hundreds of significant Upper Paleolithic works of art remained unknown except to specialists. To rectify this neglect was the main task Glory set himself when in 1952 Abbé Breuil, whose eyesight was failing, enlisted Glory's aid. Glory adopted a technique pioneered by Breuil: sheets of transparent plastic were placed over the engravings and the incised lines were traced on the surface of the plastic. The tracings were subsequently transferred to sheets of tracing paper large enough to encompass an entire composition. Eventually Glory's transfers covered 117 square meters of paper.

Glory did much other work at Lascaux: excavation, the collection of artifacts, photography. In the early years he often pursued his researches at night when the flow of tourists, as many as 1,000 per day in season, had ended. Nevertheless, his greatest labor—the recording of the engravings—was also his most important. From the closing of the cave in 1963 to the time of his death Glory continued to prepare this great mass of graphic material for publication. In so doing he has given prehistorians a base for the study of Upper Paleolithic cave art and symbolism that exists nowhere else and may never be surpassed.



**ENGRAVED REINDEER, 70 centimeters long from its muzzle to the tip of its stubby tail, stands at one edge of a composition in the Apse that includes a painted yellow horse. Two rear hooves and a small part of the tail of an engraved horse overlap the reindeer's rump and tail. This copy of another of Abbé Glory's tracings is somewhat simplified for the sake of clarity.**

# The Structure and Function of Fish Schools

*Schooling serves to reduce the risk of being eaten. Each fish employs its eyes and lateral lines, which are sensitive to the displacement of water, to match the speed and the direction of all the other fish in the school*

by Brian L. Partridge

How do they do it? The question occurs naturally to anyone watching a school of silversides moving slowly over a reef in clear tropical waters. Hundreds of small silver fish glide in unison, more like a single organism than a collection of individuals. The school idles along on a straight course, then wheels suddenly; not a single fish is lost from the group. A barracuda darts from behind an outcropping of coral, and the members of the school flash outward in an expanding sphere. The flash expansion dissolves the school in a fraction of a second, yet none of the fish collide. Moments later the scattered individuals collect in small groups; ultimately the school re-forms and continues to feed, lacking perhaps a member or two.

Although the schooling of fish is one of the most familiar forms of animal social behavior, until recently it was little understood, partly because of the difficulty of observing minute changes of position and velocity in a school under natural conditions. The fact that a great many species of fish congregate in schools suggests that the behavior offers a considerable evolutionary advantage. How the school is formed and maintained, however, is only beginning to be understood in detail. My colleagues and I have approached this question by recording on videotape schools swimming in a large circular tank. It had been thought that each fish maintains its position in the school chiefly by means of vision. Our work has shown that the lateral line, an organ sensitive to transitory changes in water displacement, is as important as vision.

Our work has also shown that the fish school is not a regular geometric structure like a crystal lattice. In each species a fish has a "preferred" distance and angle from its nearest neighbor. The ideal separation and bearing, however, are not maintained rigidly. The actual distance and direction vary greatly, approximating the ideal only over a long period. The result is a probabilistic arrangement that appears more like a random aggregation than a lattice. The tendency of the fish to remain at the preferred distance and angle, however, serves to maintain the structure. Each fish, having established its position, uses its eyes and its lateral lines simultaneously to measure the speed of all the other fish in the school. It then adjusts its own speed to match a weighted average that emphasizes the contribution of nearby fish. The combination and comparison of information from the two sensory systems provides the basis of all the intricate maneuvers of the school.

Although most people have an intuitive sense of what a fish school is, students of animal behavior have spent much time trying to define the notion precisely. Do two fish constitute a school? Do three? Is a school that has a million members made up of half a million pairs? Does a school have a leader?

There seems to be an important qualitative difference between a pair of fish and a larger group. My analysis of videotapes of European minnows swimming in a tank shows that when there are two fish, one leads and the other follows. The follower adjusts its speed and direction to match those of the leader; the speed and direction of the leader, however, are not influenced by the movements of the follower. When a third minnow is added to the tank, the pattern changes: in a group of three or more fish there is no leader. Each minnow adjusts its speed and heading to agree with those of all the other fish, with the neighbors nearest to a given fish having the greatest influence on it. Thus in a sense the entire school is the leader and each individual is a follower.

One of the most striking qualities of a school of fish is its polarization: the parallel arrangement of the members. Polarization has been cited repeatedly in efforts to define the concept of a school. When fish feed, they often form a loose group, with the members facing in many directions. When the school is in motion, however, the polarized arrangement tends to prevail. Moreover, when the school is threatened, its members often move closer to one another and align themselves more uniformly with their neighbors. That the polarization of the school is more pronounced under a threat suggests it may be connected in some way with the adaptive advantage conferred by schooling behavior.

The role the school plays in the life of the individual fish varies greatly from one species to another. In some species fish spend all or almost all of their time in a school. In other species fish join schools only occasionally, spending most of their time as isolated individuals. Fish that spend all or most of their time in schools are often called obligate schoolers; those that form schools part of the time are called facultative schoolers. In much of the work done on fish schools it has been assumed that there is an important difference between obligate and facultative schools. My work with minnows and cod, which are facultative schoolers, and with herring, which are obligate schoolers, suggests on the contrary that in all three species the school is formed and maintained on the same principles. The only difference

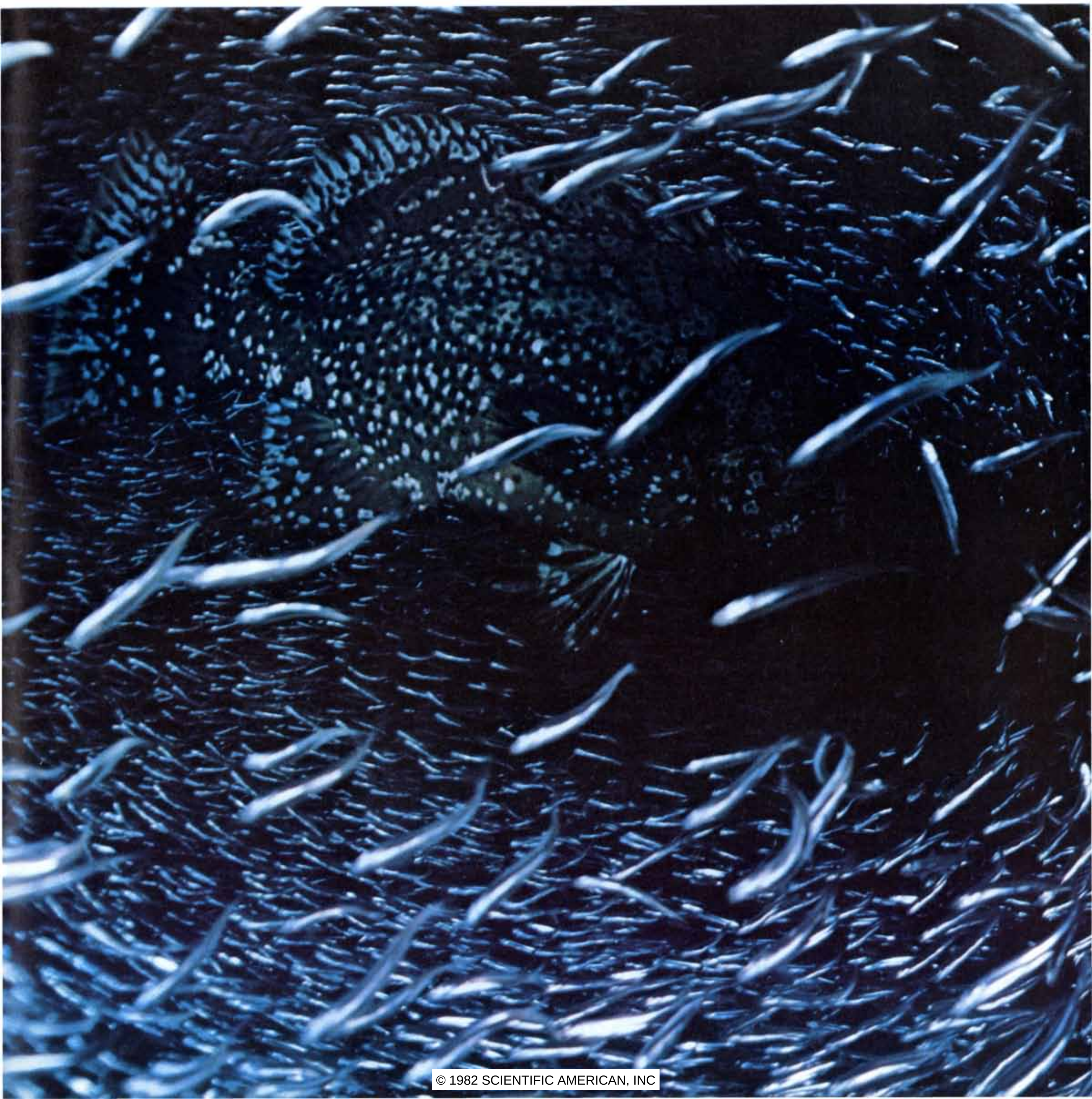
**ATTACK ON A SCHOOL OF SILVERSIDES is shown in a photograph made in a cave in the Florida Keys; the attacking fish is a grouper. More than 10,000 species of fish form schools. Most of them are small fish that are prey rather than predators. For prey the adaptive value of schooling lies in reducing the probability of detection by a predator and in reducing the risk of being eaten once the school has been detected. As the photograph suggests, one way the school reduces the risk of being eaten is by confusing the predatory fish as it makes its strike.**

seems to lie in the amount of time the fish spend in a school. From these observations it is possible to formulate a useful working definition of a school: It is a group of three or more fish in which each member constantly adjusts its speed and direction to match those of the other members of the school.

Evelyn Shaw of Stanford University has estimated that out of about 20,000 species of fish more than 10,000 species collect in schools during some part of their lives. The species that school, however, are not a representative sample. Most of the fish that form schools are small; it has generally been thought that the main evolutionary advantage of schooling lies in protecting such small fish from predators.

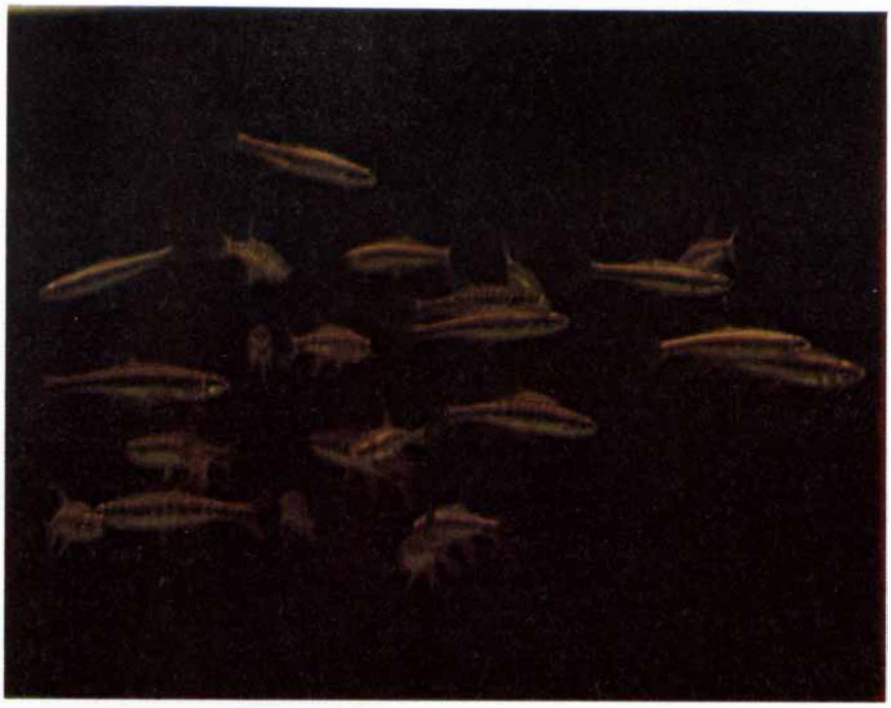
It might seem that a school made up of thousands or even millions of fish, however small the individuals are, would be highly visible; actually a school is not much more likely to be found by an ocean predator than an isolated fish is. The reason has to do with the optical character of the medium in which both the prey and the predator live. Contrast is extremely important for distinguishing an object from its background. In a large body of water the scattering of light by suspended particles and the absorption of light by the water itself greatly reduce the contrast. As a result, even in water of exceptional clarity the greatest distance at which an object can be seen is about 200 meters, and the distance does not depend on the size of the object. In practice the maximum is usually much less. (Scuba divers consider a visibility of from 30 to 50 meters to be exceptionally good.)

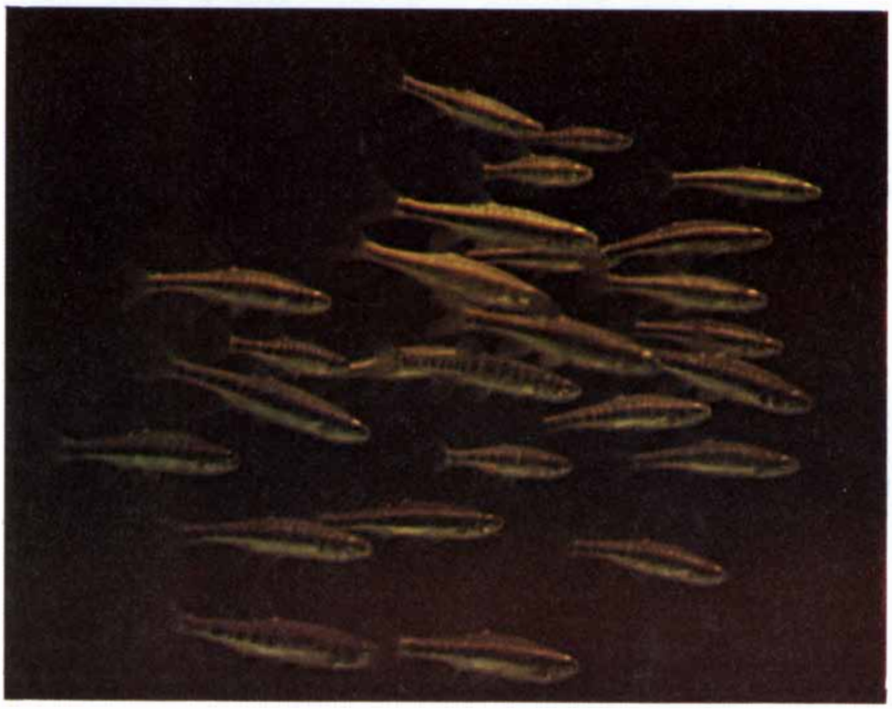
Consider three fish swimming close together in a simple school. The area within which each fish can be seen is defined by a sphere whose radius is the maximum distance of visibility. Since the fish are in a compact group, the spheres overlap to a great extent. The chance of a predator's finding the school



is therefore only very slightly greater than the chance of its finding a single fish. Indeed, the chance of a predator's finding the school is about one-third the chance of its finding at least one of the three fish if they were separated.

The example of a school with three members may appear trivial. It turns out, however, that in the open ocean a predator's chance of finding a school of 1,000 fish is only slightly greater than its chance of finding one fish. If the predator, on discovering the school, eats exactly one fish, then an individual fish's risk of being eaten is about a thousandth of what it would have been if the prey had been discovered on its own. The advantage afforded by being in a school is thus substantial, and it appears to increase with the size of the school.





**POLARIZATION, or parallel arrangement, is one of the most conspicuous features of a fish school. When the school is threatened, it becomes more highly polarized and more densely packed. The effect is shown in photographs of European minnows in a tank. In the top photograph the minnows are undisturbed. The bottom photograph shows them soon after a pike, which is a predatory fish, was put in the tank. Schools of some species are ordinarily more polarized than those of other species, a fact that was once thought to constitute an essential difference. It now appears that all schools are organized according to the same principles.**

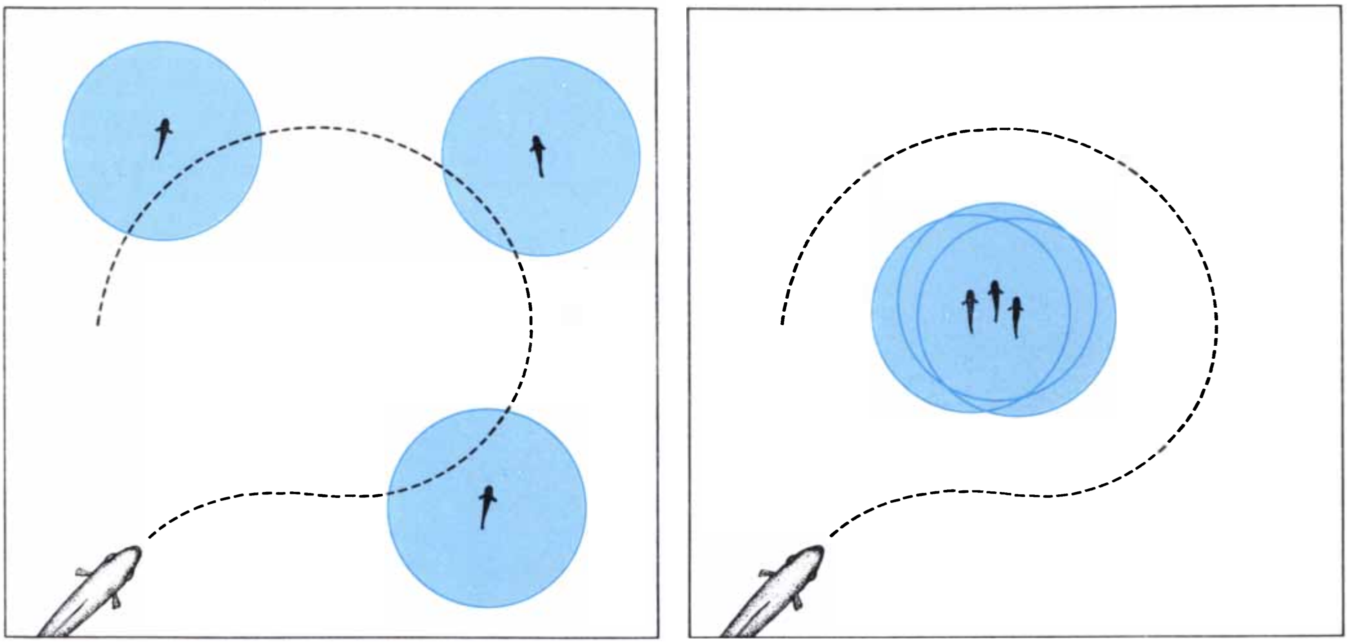
Several considerations suggest, however, that for the school to have flourished as a social form over a long evolutionary history it must provide some benefit beyond reducing the probability of detection. One confounding observation is that some fish form schools even in the presence of predators, where there can be no possibility of escaping detection. Benoni H. Seghers of the University of Western Ontario has shown that there are so many predators in the streams of Trinidad that guppies are constantly within sight of predators. The guppies continue to school. Moreover, their daily routes to and from feeding grounds take them past many predators. Many species of fish living on the coral reefs off the coast of Florida spend their entire lives within a few feet of predatory barracuda and groupers; nevertheless, schooling is common among the prey species. The notion of predators searching a limitless ocean for scarce prey clearly cannot explain such schools.

Furthermore, the school must be at least slightly more conspicuous than a single fish. It follows that an individual might reduce its chance of detection by leaving the school. The persistence of the school suggests that it continues to be of value to its members even after detection.

Another deficiency of the "detection" theory can be illustrated by again considering the three-fish school. Although the school is not much easier to find than a single fish, the survival of the individual fish is not necessarily enhanced. Suppose that when the school is found, the predator eats all three of its members; then none of the prey will have gained an advantage by being in the school. On the other hand, if the school were so large that a predator could not possibly consume all the fish, an advantage might remain.

Even if it can be established that more fish survive on the average when they swim in a school, it does not follow that schooling is advantageous for any particular fish. Natural selection acts on the individual, and in general there is no benefit to the individual in improving the welfare of other members of the group. There are exceptional circumstances, however, in which behavior that benefits the group might be favored by selection, namely when the members of the school are closely related. If an individual has many genes in common with the other fish in a school, the survival of those fish tends to perpetuate the individual's own genes.

By analyzing the structure of enzymes that exist in several chemical forms, Moira Ferguson and David L. G. Noakes of the University of Guelph in Ontario have shown that golden shiners (a freshwater species) in any one school are more closely related to one



**PROBABILITY OF DETECTION is reduced by forming a school. Because of the scattering and absorption of light in the ocean the greatest distance from which an object of any size can be seen is roughly 200 meters; the maximum is usually much less. In the diagram the circles represent the greatest distance from which a single fish can be seen. The illustration shows the view from above; in the ocean the radii would, of course, form spheres. If three fish are isolated from one another, the chance of a predator's finding at least one of them is fairly great (*left*). If the three fish form a school, the circles representing the maximum distance of visibility overlap to a large extent (*right*). The probability of a predator's finding the school is therefore only about one-third the probability of its finding at least one of the isolated fish. If on finding the school the predator ate all three of its members, schooling would have been of no benefit. For this reason schooling must somehow also reduce the likelihood of a fish's being eaten after the school has been discovered by a predator.**

another than they are to members of other schools. Hence the "kin selection" mechanism may be operating in this species. On the other hand, there is probably little genetic similarity among the ocean fish in a school because larvae from different parents are intermixed as they float freely in the water.

It seems that in many species schooling can offer a substantial evolutionary advantage only if it reduces the chance that an individual will be eaten once the school has been found. There are several ways it might do so. Albert Eide Parr, one of the first workers to study schools in a quantitative way, observed that a school is more densely packed and more highly polarized when it is under attack. Parr hypothesized that open-water fish respond to the lack of cover in the ocean by hiding behind one another; the school is the result. It has also been suggested that a predator might perceive a dense group of small prey as a large, frightening object, but one would expect natural selection to favor predators that are not fooled in this way.

A more plausible explanation of the adaptive value of the tightly packed school is that it reduces the predator's chance of making a successful kill. A predator facing a large number of prey often has difficulty choosing a single fish to attack. The phenomenon has been designated the confusion effect, but it may result from two quite different processes. One process takes place in the central nervous system: the predator simply cannot make a choice among the members of the school. Many predators prefer to strike prey that are distinct from the rest of the school in appearance or behavior. Even very small differences are enough to overcome the predator's inability to make a decision, but in many schools the fish are almost identical in appearance, and the predator may have difficulty selecting one.

The second process may have its origin in the peripheral nervous system. It is the sensory confusion caused by a large number of prey moving around the predator. Even if the predator makes the decision to attack a particular fish, the movement of other prey in the vicinity can be distracting. The predator's difficulty can be compared to that of a tennis player trying to hit two tennis balls at once.

The mechanism of sensory confusion seems to me more likely than that of indecision to be responsible for the predator's dilemma. If there really were no reliable criteria for selecting a fish to attack, natural selection should favor predators that choose randomly and quickly. Sensory confusion, on the other hand, is an indirect consequence of perceptual sensitivity to movement. There should be strong selective pressure against a predator's evolving less sensitive movement detectors because they are needed to find prey.

Whatever the mechanism, there is substantial evidence that schools confuse predators. Sean Neill and Michael Cullen of the University of Oxford have studied attacks by pike and perch on schools of bleak and dace, which are European freshwater fish. Increasing the number of prey from one to six and then to 20 reduced the frequency of the attacker's strikes and the probability of success.

By diminishing the predator's chance of finding prey and confusing the predator once the prey is found the school is of benefit to each of its members. The advantages discussed so far can be attributed primarily to the form of the school itself rather than to the active cooperation of its members. By cooperating, the members of the school can reduce still further their chance of being eaten.

Schools of fish engage in several dramatic evasive maneuvers. The tactic adopted depends in part on how rapidly the predator is approaching. The tactics can be illustrated by the responses of various prey species to barracuda, a common predator on schools of small fish in tropical waters. The barracuda has evolved a shape well suited to a quick strike, which is its characteristic attack. It has a long, torpedo-shaped body with a pair of vertical fins near the tail. In a strike the fins act like a second tail, providing a powerful forward impetus. Such a body plan is not efficient for sustained high-speed swimming. The barracuda tends to sidle up to its prey and then strike in a single motion.

When the barracuda moves slowly up to a school, the prey may back away, creating a cavity around it. More of-

**SCHOOL OF PREDATORY FISH swims in a parabolic formation with the concave side of the parabola forward. The fish are giant bluefin tuna. The photograph was made by the U.S. National Marine Fisheries Service in a census of the tuna population. It has long been known that some predators, including tuna and barracuda, congregate in schools. It was thought, however, that schooling merely increased the predators' visual range: if one fish found prey, the others could join in the kill. The author's analysis of the tuna school suggests that they hunt in a more truly cooperative manner. The tuna apparently work together to drive schools of prey between the outstretched ends of the parabola, then surround and destroy the prey.**

**EVASIVE TACTICS employed by schools of prey are shown here and in the two illustrations on the opposite page. The photograph shows a school of dwarf herring forming a vacuole, or cavity, around a barracuda. The tactic adopted depends partly on how fast the predator is approaching. The torpedo-shaped barracuda with its pair of vertical fins near the tail is well adapted to a quick strike but not to sustained high-speed swimming. It often sidles up to its intended victims before striking. As the barracuda moves slowly toward the school the dwarf herring form the vacuole, maintaining sufficient distance to escape if the predator attacks.**

ten, however, the school splits into two parts in front of the predator. The halves of the school turn outward, swim around the barracuda and rejoin behind it. The tactic has been named the fountain effect by Geoffry Potts of the Laboratory of the Marine Biological Association at Plymouth, England. The result is that the predator is left with the school behind it. If the barracuda turns, the maneuver is repeated. By a succession of such movements the school can evade a predator it cannot outrun.

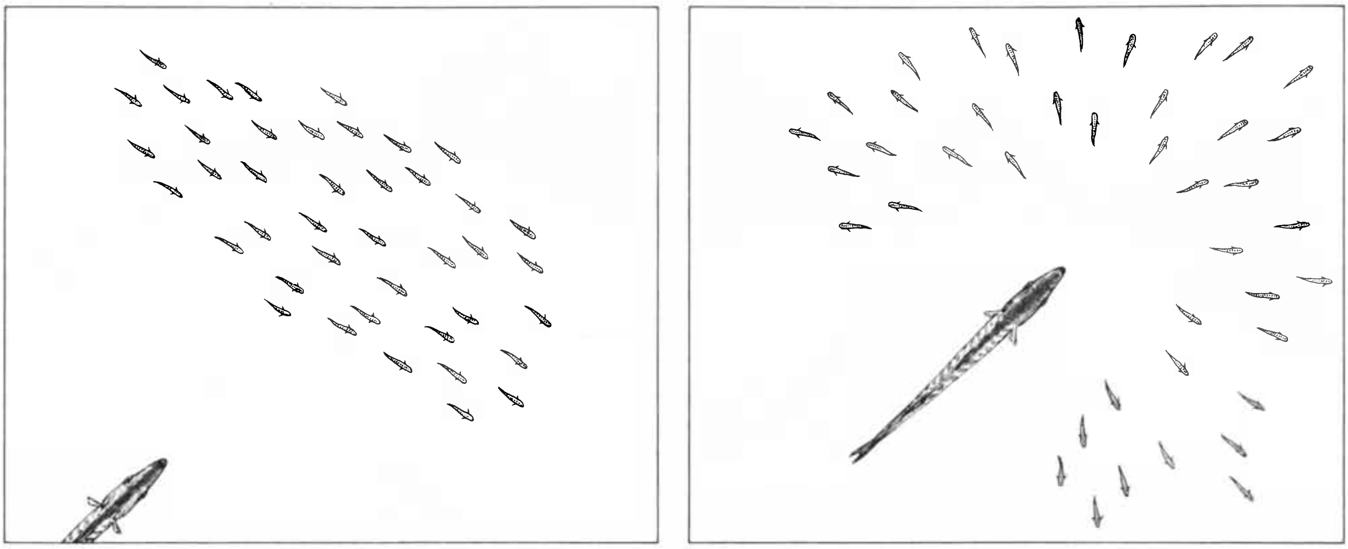
When the barracuda strikes, maneuvers as stately as the fountain effect are of little use. In response to a quick strike each fish darts from the center of the school. In such a flash expansion each fish moves radially outward, propelled by a single flick of the tail. The movement resembles a bomb burst. In as little as a fiftieth of a second each fish accelerates from a standing start to a velocity of between 10 and 20 body lengths per second. The entire expansion can take place in half a second.

Because the expansion is created by roughly simultaneous tail flicks throughout the school it seems it cannot be coordinated by any means that would require each fish to register the movements of its neighbors. In all probability each member of the school "knows" where the other members will go in the event of an attack. The hypothesis is given some support by the fact that collisions have never been observed in fish with all their sense organs intact.

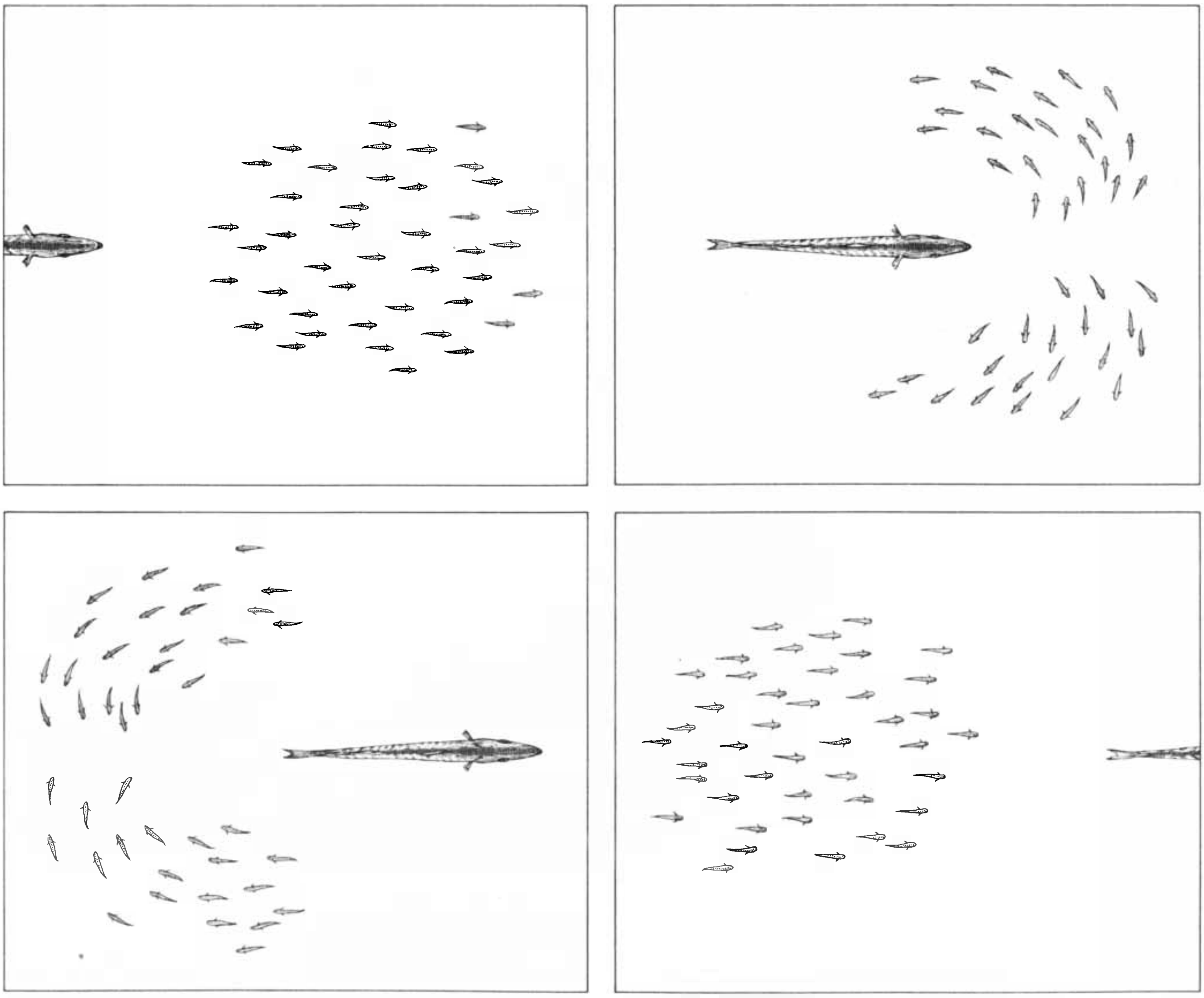
Although most work done on schools has concerned species of fish that are consumed rather than consumers, some predators also form schools; among them are the barracuda and the tuna. It has long been assumed that when predators school, they act more or less as a group of individual hunters. Forming a school could nonetheless have an adaptive value by increasing the search area of the hunter. If one member of the school finds food, the other members can take advantage of the find. If the members of the school remain barely in sight of one another, the search area is at a maximum. This is a much looser and more individualistic form of hunting than that seen among lions or wolves, where the pack can bring down the prey together, or among dolphins, which herd their prey into shallow water.

Recently I have begun to suspect that some predatory fish also coordinate their hunting in a cooperative way. I have analyzed the structure of schools of giant bluefin tuna (a fish that can reach three meters in length and more than 400 kilograms in weight). Aerial photographs of tuna schools made by the U.S. National Marine Fisheries Service for counting the tuna population show that the arrangement of the tuna in the schools is remarkably regular.

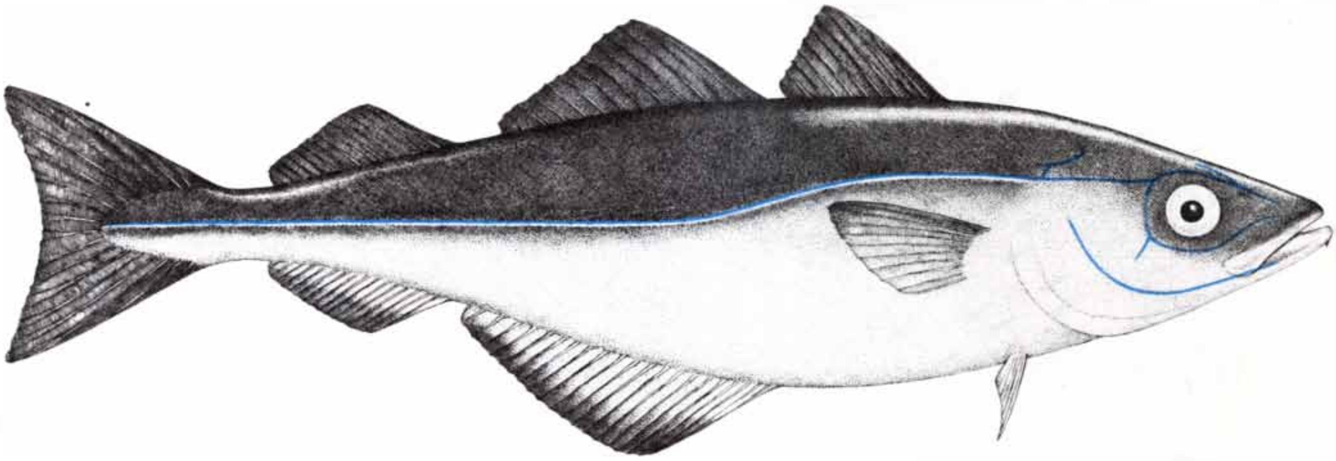
Tuna schools of 50 or more members



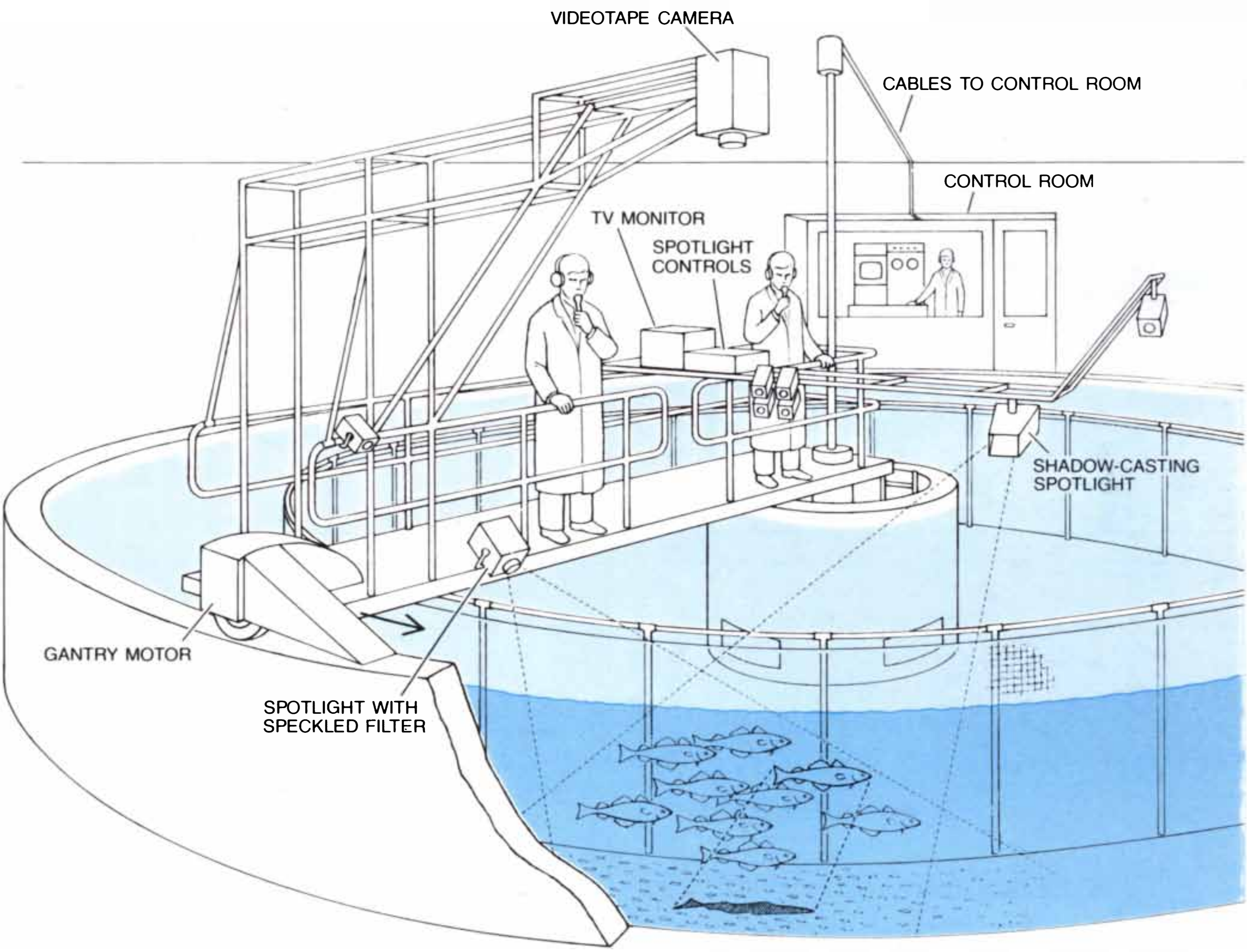
**FLASH EXPANSION is the most dramatic of the evasive tactics of the dwarf-herring school. As the barracuda strikes, the school expands in the form of a sphere. The entire expansion can take place in as little as half a second. It is accomplished by a single movement of the tail on the part of each member of the school. Collisions in the course of the expansion have never been observed in fish that have all their sensory organs intact, and so it appears that each fish must "know" where its neighbors will go in the event of an attack.**



**FOUNTAIN EFFECT is a tactic that enables a school of small, slow-moving prey to outmaneuver a predator it cannot outrun. As a barracuda moves toward a school of dwarf herring the school splits and flows in two groups behind the larger fish, which is carried forward by its own momentum. The school re-forms behind the barracuda. If the predator turns to face the school again, the maneuver is repeated.**



**LATERAL LINE,** an organ sensitive to transitory displacements of the water, provides information that helps a fish to maintain its position in a school. The lateral-line canal is shown in color on a drawing of a pollock (*Pollachius virens*), a saltwater fish closely related to the cod. The pollock has been utilized in much of the author's work. The lateral line is made up of gelatinous canals connected by pores to the external environment. Inside the canal are thousands of hair cells, which are much like the sound receptors in the ear of a terrestrial vertebrate. The response of the hair cells to the displacement of water gives the pollock information about the speed and direction of neighboring fish. The author's work has shown that schooling fish compare information from the eyes with information from the lateral lines.




**FISH SCHOOLS IN THE LABORATORY** were studied by the author with apparatus at the Department for Agriculture and Fisheries for Scotland in Aberdeen. The circular tank is 10 meters in diameter. A plastic fence created a circular channel 1.5 meters wide. The gantry projecting from the central stanchion held a videotape camera, other equipment and observers. Schools of from 20 to 30 fish with numbers branded on their side were trained to stay over a speckled spot projected from a light on the gantry. As the gantry rotated at a constant speed the school moved to keep pace with the speckled spot. A beam of red light was projected diagonally through the school from the stanchion. From above, the distance between a fish and its shadow in the red spot indicated the depth of the fish. Videotapes were made, and an observer gave a description of the positions of the fish. A plot of the coordinates of each fish in successive frames of the videotape was made with the aid of a computer. It yielded information about the adjustment of position that takes place in the school.

may divide into smaller groups when hunting. The smaller schools consist of between 10 and 20 fish spread out along a curve closely resembling a parabola, with the concave side forward. Achieving a regular spacing of individuals along a parabola is a difficult feat because the distance and the angle between each pair of tuna are different. That the form of the school is maintained in spite of this difficulty suggests it must provide a considerable advantage in hunting.

The possible nature of the advantage can be considered by means of an analogy with the functioning of a parabolic mirror in a telescope. Any light ray parallel to the axis of the parabola is reflected from the concave surface toward the focus. It is possible that a similar effect operates when a parabolic school of tuna swims parallel to its own axis. If the prey react to the curved school as if it were a solid wall, they will be driven into the focus of the parabola, which is the most convenient place for the tuna to surround and consume them.

For both predators and prey the value of the school thus depends on the ability of its members to coordinate their movements quite closely. Some of the advantage is derived from the geometric form of the school, but much of it comes from the tactics employed by the school. In addition the adaptive value of the school increases with its size. As a result of such evolutionary factors schools with a million members are not uncommon.

How is such a large group organized? Finding the answer required a means of recording very small changes in the position of each fish in the school over a long period. The record was made possible by a unique apparatus at the Department for Agriculture and Fisheries for Scotland in Aberdeen. The department's fish laboratory has a circular tank 10 meters in diameter and one meter deep. Putting a plastic fence in the tank yielded a channel one meter or 1.5 meters across, depending on the experiment.

Projecting from a stanchion in the center of the circular tank is a rotating gantry. The five-ton gantry is large enough to support two observers and many pieces of equipment. From the gantry we projected a speckled spot of light onto the floor of the tank. It is possible to train a school of fish to stay over such a spot. When a school was put in the tank, the first few days were spent training the school to stay over the spot as the gantry moved around the tank.

Because the school is a three-dimensional structure it is necessary to examine the vertical relations within it as well as the horizontal relations. This was done by mounting a second spotlight on the gantry. The light was projected downward at an angle through the school. When the school was observed from above, the distance between each fish and its shadow indicated the height of the fish above the bottom.

A videotape camera was mounted on the gantry. As the gantry followed the school the camera recorded the horizontal positions of the fish; measuring the distance to a fish's shadow yielded its depth. In each frame of the tape the position of each fish in three dimensions was recorded with the aid of a computer. More than 35,000 frames were analyzed, providing more detailed information on schools than had ever been gathered. Most earlier work was based on visual observation or on no more than a few hundred frames of film.

In addition an observer riding on the gantry as it followed the school gave a continuous commentary on the positions of the fish. The commentary was recorded and later coordinated with the video tape. To make the analysis of the tape easier the fish were branded with numbers. A cold metal brand was applied to the side of the fish, where it left a temporary mark. Utilizing this setup, my colleague Tony J. Pitcher and I observed schools of cod, herring and a third fish much like the cod, which is known as the saithe in England and the pollock in the U.S. In most cases the schools consisted of about 20 fish.

One of the most persistent misconceptions about fish schools is that they have a regular geometric form, such as the cubic lattice characteristic of some crystals. Such a regular form has not been observed in the schools of any fish species. Our work shows that the structure is a rather loose or probabilistic one, and that it results from each fish's applying a few simple behavioral rules. The first rule is that each individual maintains an empty space around itself. For each species there is a characteristic minimum-approach distance within which neighbors do not come. The absolute distance depends on the size of the fish; it is usually about three-tenths of a body length.

The minimum-approach distance is not, however, the distance that is generally maintained between the fish in a school. In each species there is a typical preferred distance to the nearest neighbor, which is usually about one body length. In general only one neighbor at a time is at the preferred distance from a particular fish. (In a cubic lattice several neighboring fish would all be at the same distance.)

The spatial relations among the fish in a school change constantly as the fish adjust their speed and direction. For this reason the distance to the nearest neighbor is not uniform, even for a single fish. The preferred distance is a statistical abstraction, found by averaging the actual distances over a long period.

Fish of any one species also tend to keep their nearest neighbor at a particular angle with respect to their body axis. Like the preferred distance, the preferred angle is a statistical quantity. At any given moment only a few fish may have their nearest neighbors at the preferred angle, but over a long period the preferred angle predominates.

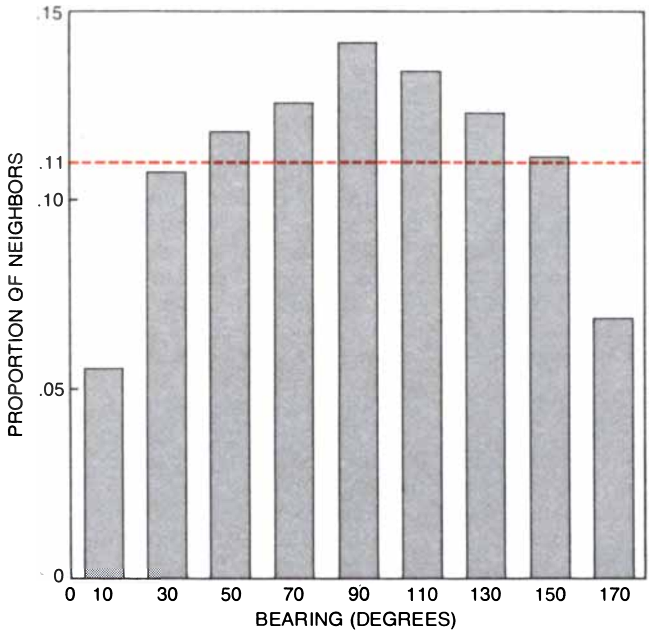
One useful measure of the degree of structure of a school is the average ratio between the distance to the second-nearest neighbor and the distance to the nearest neighbor. The closer the ratio is to 1, the more uniform the structure is. In a cubic lattice the ratio is exactly 1. The ratio varies considerably among species. For herring it is about 1.1, for pollock 1.3 and for cod 1.5. A ratio of 1.5 is only slightly less than the ratio of 1.6 that would be observed if the fish in the school took up positions at random.

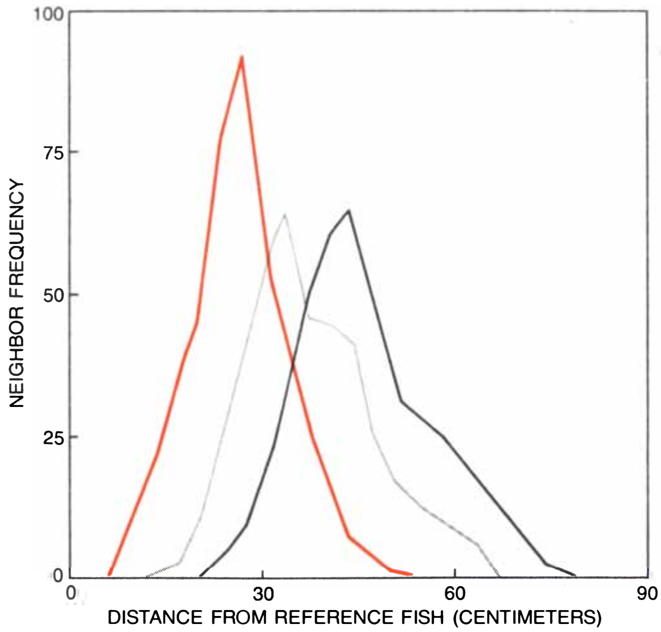
Several other measures suggest that schools of herring are organized in a more regular way than those of pollock or cod. Even when such differences are taken into account, however, it appears that most schools are organized on the same lines: by the maintenance of a preferred distance and angle. Attempts to distinguish species on the basis of whether they are facultative or obligate schoolers or of how highly polarized the school is now seem misguided; such categorical distinctions probably do not exist. Fish schools seem to vary along a continuum in their degree of organization. Other workers have shown that schools of squid, frog tadpoles and even flocks of certain birds are organized on the same principles.

In our work we were particularly interested in the function of the lateral line in determining the structure of the school. Most species of fish have a prominent lateral line on each side of the body. The displacement-sensitive receptors that make up the line are known as hair cells and are much like the receptors in the ear of a terrestrial vertebrate. The hair cells are placed in canals laid out in a complicated way on the head of the fish and in a roughly linear arrangement between the head and the tail.

Although it had been suggested that the lateral line plays a role in the formation of the school, most workers thought vision was much more important. In the 1920's Parr hypothesized that schooling was accomplished by vision alone. According to Parr's scheme, a fish is attracted by the sight of a member of its species but repelled if it comes too close. The cohesion of the school is thus the result of balanced attractive and repulsive forces, both originating in vision.

To test whether the lateral line might not also have some influence we observed schools of pollock that included fish that had been temporarily blinded or had had their lateral lines cut behind the operculum, the bony flap covering the gills. The pollock were blinded by placing opaque contact lenses over their eyes. When the blinded fish were placed in a school of unimpaired fish, they re-



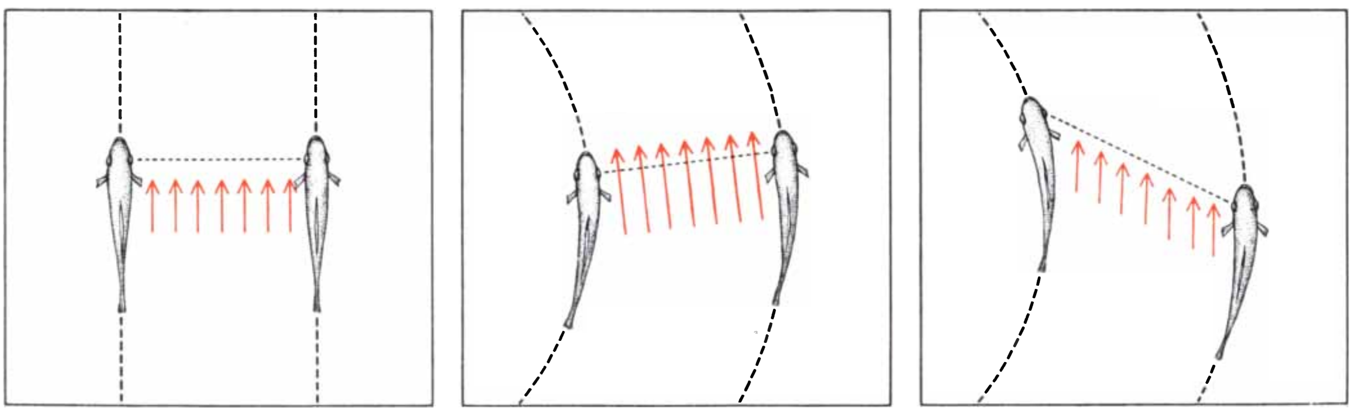




**STRUCTURE OF A SCHOOL is loose and probabilistic, unlike the regular arrangement of atoms in a crystal lattice. The structure is the result of the tendency of each fish to keep its nearest neighbor at a particular distance and angle. The graph at the left shows where the nearest neighbor is likely to be found in the area around a given pollock. Zero degrees is directly ahead of the fish; 180 degrees is directly behind. A pollock's nearest neighbor tends to be roughly alongside it, as the tallest bar at 90 degrees indicates. The broken line indicates where the neighbors would be found if the fish took up positions at random. The graph at the right shows how often the nearest neighbor (*color*), the second-nearest neighbor (*gray*) and the third-nearest neighbor (*black*) are to be found at various distances. The nearest neighbor is most commonly at a distance of from 25 to 30 centimeters, or about one body length. The horizontal spread of the curves shows that the preferred distance and angle are statistical abstractions: at any moment numerous fish are at other distances and angles.**

sponded to changes in speed and direction by the school and maintained their position among the other fish. Behavioral changes were observed, however: the blinded fish tended to swim somewhat farther from their nearest neighbor than pollock ordinarily do.

Fish whose lateral lines had been cut were also able to school. In contrast to the blinded fish, however, those whose lateral lines had been cut swam closer to their nearest neighbor than pollock generally do. Only if the fish were both blinded and had had their lateral lines cut did they fail to maintain position. The results suggest that information from both the eye and the lateral line is utilized when fish school. The distance maintained by the eyes alone is smaller than the distance maintained by the lateral lines alone; the preferred distance in the unimpaired fish lies between these values. Vision does seem to provide the attractive force between members of the school (the pollock swim farther apart without it); the repulsive force, however, appears to be provided by the lateral line (without it the fish swim closer to one another).

Other results suggest that vision is the more important sense for maintaining distance from and angle to the nearest neighbor. The lateral line appears to be most important for determining the neighbor's speed and direction. Strong evidence that both senses are being utilized at once comes from measurements of the correlation between the speed and direction of a particular fish and those of other fish in the school a short time before; such correlations can indicate what standard of reference each fish employs in adjusting its velocity. For neither speed nor heading is the correlation between a fish and its nearest neighbor very strong. Moreover, the results show that the school has no leader:



**CONTRADICTORY INFORMATION can be given by the eyes and the lateral lines. When the school swims in a straight line, the senses are in agreement. The eyes tell the fish on the left that its neighbor is keeping up; the lateral lines tell it that its neighbor is swimming at the same speed (*left*). When the school turns in an arc and the neighboring fish remain side by side on parallel courses, the eyes of the inner fish tell it that its neighbor is keeping pace. Because the outside fish must swim a greater distance, however, the lateral lines tell the inner fish that its neighbor is swimming faster (*middle*). If the inner fish accelerated to match its neighbor's speed as measured by the lateral lines, the school would disintegrate (*right*). When information from the senses is contradictory, vision apparently takes precedence.**

speed and direction are not closely related to those of any other single fish.

The strongest correlations are observed between the speed and direction of the individual and the average speed and direction of the entire school. The average that is most strongly correlated is not the simple arithmetic mean of the speeds and headings of the members of the school. A fish is much more strongly influenced by its near neighbors than it is by the distant members of the school. The contribution of each fish to the average is inversely proportional to either the square or the cube of the distance.

A correlation based on the square of the distance and one based on the cube are about equally accurate in accounting for our observations. If the school were maintained by vision, one would expect the correlation to depend on the square of the distance. Discriminations made by vision depend in part on the area of the perceived figure. The area decreases with the square of the distance.

If the school were maintained solely by the lateral line's sensitivity to water displacement, on the other hand, one would expect the correlation to depend on the cube of the distance. The volume of water displaced varies inversely with the cube of the distance. The fact that the correlation based on the square and the one based on the cube are about equally strong suggests that both senses are employed.

Although both the eyes and the lateral lines appear to be in use when fish school, there are times when the information from them is contradictory. Such a conflict arises when the school turns in an arc (as it does continuously in a circular tank). If two fish make the turn side by side, the fish receive conflicting information. For example, the eyes of the fish on the inner course tell it that its outer neighbor is just maintaining position. Because the fish on the outer course must cover a greater distance to keep up, however, the information from the lateral lines tells the inner fish that the outer neighbor is swimming faster. Our work shows that when the information from the two sensory systems is in conflict, the information from the eyes takes precedence.

In investigating particular forms of animal behavior biologists have tended to look for a single sensory explanation. It is now known that schooling is accomplished by comparing information from more than one sensory source. Certain other phenomena, such as the navigation of homing pigeons, also seem to require multiple sensory systems. This might have been expected for evolutionary reasons alone: selection would tend to favor the animal capable of exploiting the most information. When the intricate maneuvers of the fish school are completely understood, it may be found that still other senses participate.

# Brain Mechanisms of Visual Attention

*The process by which the brain decides that certain objects in the world are significant is studied by recording the activity of nerve cells in the brain of monkeys that are responding to visual stimuli*

by Robert H. Wurtz, Michael E. Goldberg and David Lee Robinson

Life is a bombardment of sights, sounds, smells, touches and tastes from which we select some information as a basis for action, ignoring all the rest. Thus we ignore the sensations evoked in our skin by our clothes but react immediately to the sensations evoked by an insect crawling under them. We ignore the din at a cocktail party but concentrate on our particular conversation. This ability to select objects of interest from the environment is called attention. William James described it well: "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others."

The process of visual attention, that is, the process of selecting important objects in the visual world, has been of particular interest to psychologists and physiologists because it has a reliable and measurable concomitant: the movement of the eyes. The reason is straightforward. The retina is distinctly nonuniform in its sensitivity to patterns of light. One tiny area of the retina, the fovea, which responds to light from the center of the visual field, has a high concentration of light-receptor cells; it therefore analyzes a pattern of light on a much finer scale than the periphery of the retina can. The brain rotates the eyes in their sockets so that the analytic power of the fovea is directed toward objects of interest. As you read these words, for example, your gaze jumps from one group of words to the next in a way that projects their image onto the fovea. Each such movement is called a saccade, from the French *saccader,* to jerk. Usually it lasts for less than 50 milliseconds. Nearly all the gathering of visual information that leads to visual perception occurs in the periods between saccades when the gaze is fixated on a visual target.

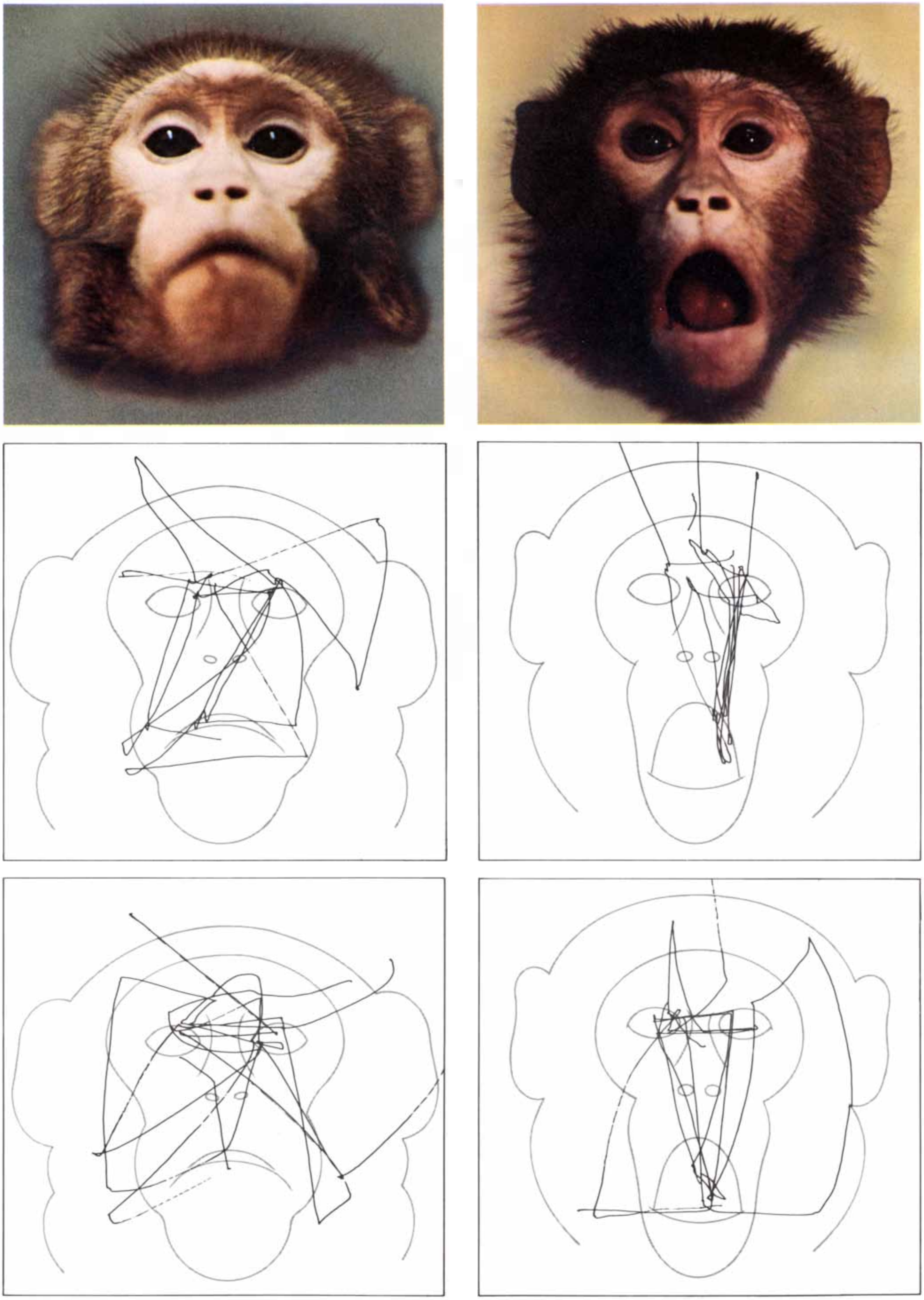
Since the process of visual attention is so closely linked to saccades, the question "What is a person attending to?" can be approximated by the question "What is the pattern of that person's eye movements?" Indeed, the Russian psychologist Alfred L. Yarbus, who recorded the patterns of people's eye movements as they looked at various pictures, has shown that the fixations tend to concentrate on a picture's most salient features. On the other hand, people also attend to objects they see "out of the corner of their eye." Michael T. Posner and his colleagues at the University of Oregon have done experiments in which the time it takes a person to respond to a target light by pressing a bar is taken to be an indication of how well the person attends to the light. People who know where the light will appear tend to respond more quickly to it even if they are expressly asked not to make a saccadic eye movement to it.

From these observations on visual attention one can devise a description of the process that makes it possible to investigate its mechanisms in the nervous system. First, visual attention implies the selection of a visual object from the visual field at the expense of other objects. Second, the selection must preserve the location of the object; obviously one must know where the attended object is. Third, the nature of the response to the selected object is not crucial; after all, one may gaze at an object, reach for it or simply notice it, but in each case one has attended to it.

In experiments that began at the National Institute of Mental Health and continued at the National Eye Institute we have searched for neural activity that meets these three criteria. Specifically, we have recorded the electrical activity of nerve cells in the brain of the rhesus monkey. We made these recordings at times when the monkeys were alert and able to move their eyes and attend to visual stimuli. We assume that for our purposes the monkey is an adequate model of man. One reason to think so is that the monkey's visual acuity, perception of color, perception of depth and pattern of saccades all resemble those of man. In addition the regions of the monkey's brain known to be involved with vision and with eye movement resemble their counterparts in the brain of man in both their position in the brain and their anatomical structure. We hope, therefore, that the results of anatomical, physiological and behavioral experiments with monkeys can yield some understanding of the neural basis of human behavior.

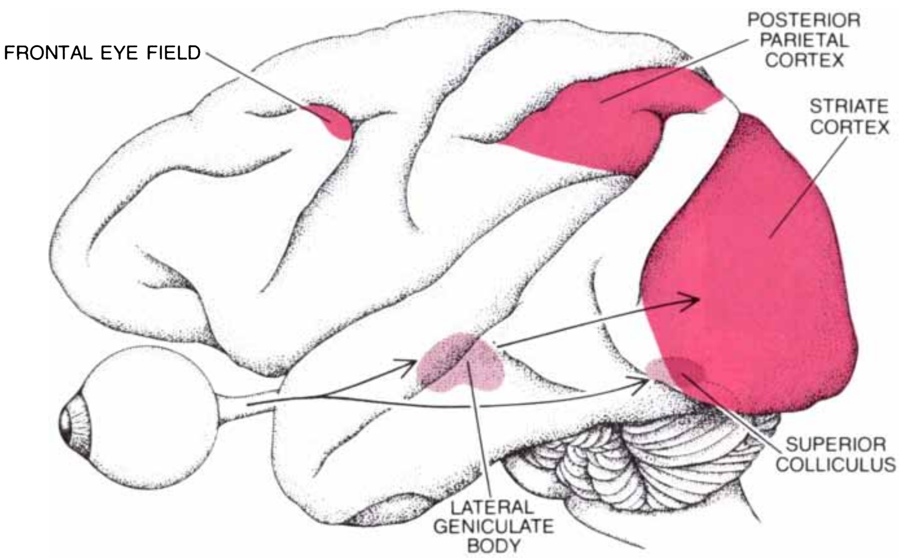
The attempt to record and interpret the electrical activity of nerve cells in the brain is facilitated by the fact that the activity has a characteristic pattern. Ordinarily a nerve cell has a resting electric potential: a voltage between the in-

**EYE MOVEMENTS made by rhesus monkeys as they viewed pictures of the faces of other monkeys reveal patterns of visual attention. Here two monkeys' faces (*top row*) were each shown for eight seconds to monkeys named Joe and B.F. One face was more or less expressionless; the other face had an expression known as open-mouthed threat. The open mouth got much attention. In general Joe's visual fixations (*middle row*) and B.F.'s (*bottom row*) are connected by rapid eye movements, or saccades. The fixations trace out the faces' salient features. The illustration is based on the work of Caroline F. Keating of Colgate University and Gregory Keating of the Upstate Medical Center at Syracuse of the State University of New York.**

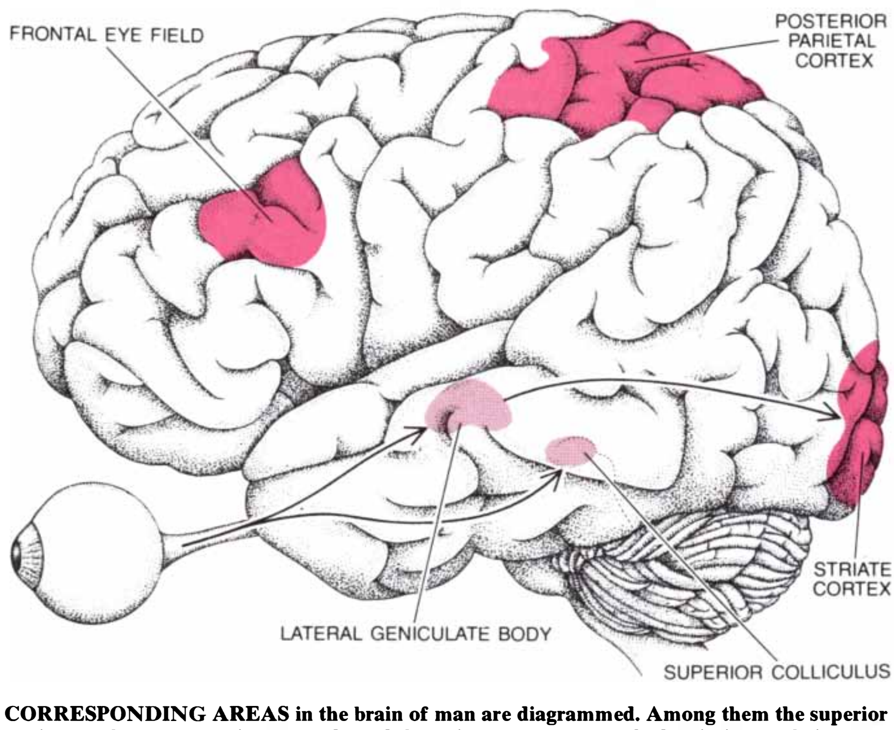


terior of the cell and the exterior. At certain times this potential changes dramatically; then a few milliseconds later it reverts to its resting value. The deviation from the resting potential, which is called an action potential, travels the length of the axon, the long fiber extending from the nerve cell. In this way the nerve cell sends messages to other nerve cells. The change in potential can readily be measured by means of a fine-wire microelectrode placed near the cell.

Suppose a particular nerve cell is shown to generate a burst of action potentials whenever a spot of light impinges on a certain part of the retina. The cell discharges sporadically at other times, but even so it is fair to conclude that the cell is related to the neural mechanisms underlying visual perception or the actions guided by vision. Similarly, suppose a nerve cell discharges in an intense and predictable way before a certain movement of the body. Presumably the cell is related to the neural generation of that movement.




**SEVERAL BRAIN AREAS in the rhesus monkey include cells that respond to visual stimuli (for example spots of light on a screen the monkey is watching) by changing their electrical activity. The superior colliculus gets signals direct from the retina. So does the lateral geniculate body, which communicates with the striate cortex. Two other cortical areas the authors studied also respond to visual stimuli. They are the frontal eye field and the posterior parietal cortex.**




**CORRESPONDING AREAS in the brain of man are diagrammed. Among them the superior colliculus, the lateral geniculate body and the striate cortex are markedly similar to their counterparts in the monkey. The great expanse of the cerebral cortex in the brain of man means, however, that the frontal eye field and the posterior parietal cortex are not easily delimited.**

In order to study the relation of nerve-cell activity to a monkey's behavior it is necessary to train the monkeys employed in our experiments in several visual tasks. First each monkey is taught to press a lever that makes a spot of light appear on a screen in front of it. A few seconds after the spot appears it dims and stays dim for half a second. If the monkey releases the lever while the spot is dim, it gets a drop of juice as a reward. The spot is so small and the dimming is so subtle that the monkey must train its fovea on it. (The people who try the task in our laboratory always do the same.) In fact, the monkey is sufficiently intent on watching the spot, which we call the fixation point, so that it seldom moves its eyes in saccades to look at spots of light we may flash elsewhere on the screen. If we move the fixation point, however, the monkey typically moves its eyes in a saccade from their former position to the new one.

When the monkey has learned these initial skills, it is placed under general anesthesia and surgically prepared for the making of neurophysiological recordings. We implant a device that restrains the monkey's head, sensors that measure its eye movements and a chamber through which microelectrodes can subsequently be introduced into the brain to monitor the activity of individual nerve cells. Then at the beginning of each day of experimentation the monkey is released from its home cage and allowed to climb into the experimental apparatus. Typically it does its tasks willingly. With a given monkey we usually study some hundreds of cells over a period of several months. Fortunately the brain is insensitive to pain. Indeed, the brain of human patients is often made the subject of electrical stimulation and recording while the patient is awake in order to render subsequent surgery (such as surgery to alleviate intractable epilepsy) more precise.

Since we wanted to study the physiology of visual attention, we began by examining the properties of nerve cells in areas of the brain that get signals from the retina by way of the optic nerve. In monkeys and in man the optic nerve divides into two major pathways. One of them leads to the lateral geniculate body, a collection of nerve cells whose axons project in turn to the striate cortex, the rearmost part of the cerebral cortex. Here the cortex begins the analysis of the visual world for color, shape, motion and depth.

The second main pathway leads to the

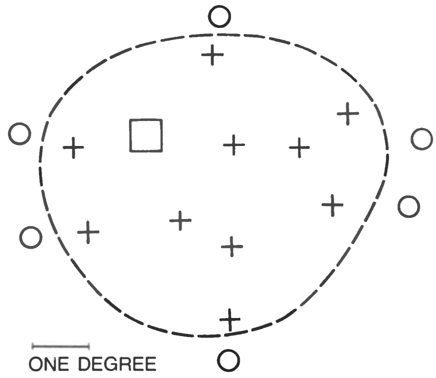
superior colliculus, an area in mammals at the top of the brain stem. The superior colliculus is the equivalent of the optic tectum, the area in less highly developed vertebrate animals that receives optic-nerve axons at the top of the brain stem. In animals such as frogs and fishes the optic tectum is the most important visual area.

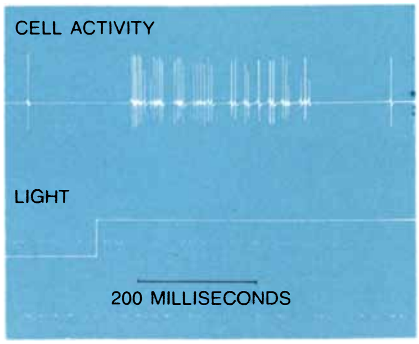
We began, then, with cells in the superior colliculus. It was already well established that such cells, like the great majority of cells in the brain that respond to visual stimuli, respond not to diffuse light but to light falling on a specific area of the retina. We can map this area, which is called the receptive field of the cell, by measuring the response of the cell to spots of light flashed at various positions on the screen while the monkey is looking at the fixation point. Since the monkey keeps its gaze steady, we assume that the position of the spot of light on the screen with respect to the fixation point is a measure of the position of the image of the spot on the monkey's retina with respect to the fovea. Spots in some locations on the screen cause the cell to respond with a burst of action potentials; spots in other locations do not.

In the case of the superior colliculus it turns out that small spots of light flashed anywhere in a large part of the visual field (so that their image falls anywhere on a large part of the retina) can activate a cell. A receptive field near the fovea may correspond to an angle of four degrees in the visual field. Farther from the fovea the fields may each be 20 degrees in diameter. (The width of your forefinger at arm's length blocks out an angle of about one degree.)

While we were mapping the receptive fields of cells in the superior colliculus we made a fortuitous observation: whenever the monkey mistakenly glanced at a spot of light in the receptive field of the cell whose activity we were recording, the cell's response was much more intense than its response to such a spot if the monkey continued to fixate. We therefore modified our experiment so that the monkey was rewarded for these saccades with a drop of juice. We confirmed that many nerve cells in the superior colliculus discharge more intensely and more regularly when the spot of light in the receptive field is the target for an eye movement. In fact, roughly half of the cells in the superior colliculus that respond at all to visual stimuli showed this enhanced response. Typically the enhanced response begins about 50 milliseconds after the spot flashes onto the screen. The eye movement to the spot comes later: about 200 milliseconds after the spot appears.

Doubtless the monkey attended to the spot of light: its eyes moved in reaction to it. Doubtless too the vigorous discharge by cells in the monkey's superior colliculus implies that the activity of other brain cells (such as the ones to which the cells in the superior colliculus project their axons) was altered. Could the enhanced response therefore be part of a neural basis for a process by which "the mind takes possession of one object"?




**RECEPTIVE FIELD of a nerve cell is the specific part of the visual field to which the cell is sensitive. At the left a receptive field is mapped. Plus signs indicate points at which a spot of light made the cell discharge at an increased rate. Open dots indicate points at which a spot elicited no increase. The cell's response to a spot at the point marked by the square is at the right.**

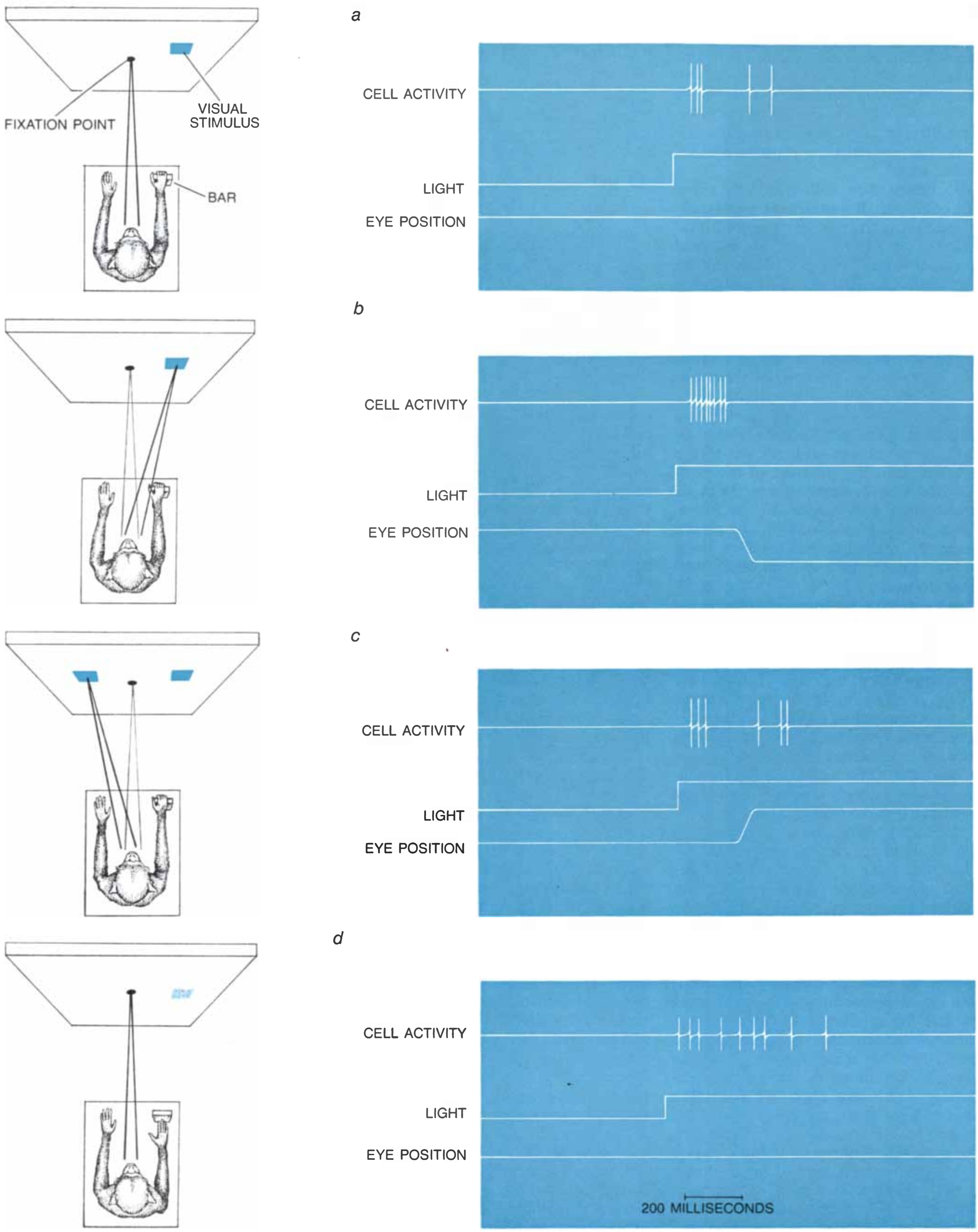
In order to answer this question we had to solve several problems. First we had to determine whether the enhanced response required a visual stimulus or whether it was merely activity preceding a spontaneous eye movement. It did require a stimulus. When the eyes of the monkeys moved in saccades in total darkness, the cells did not discharge. The enhanced response was not simply a precursor of any eye movement; it was a modulation of the reaction to visual stimuli.

Next we had to determine whether the enhanced response could be associated with a subsequent saccade that had another stimulus as its target. It was conceivable that the effort of making any eye movement increases the excitability of all the cells in the brain that respond to visual stimuli. We examined this possibility by presenting the monkey simultaneously with two spots, one in the receptive field of the cell we were monitoring and one far outside it. When the monkey moved its eyes to the spot in the receptive field, the cell anticipated the saccade with an enhanced response. When the monkey moved its eyes to the spot outside the field, the response of the cell was not enhanced. This means the enhanced response was spatially selective: it is related to stimuli in a specific part of visual space.

Finally we had to test the possibility that the enhanced response came only when the monkey reacted to the stimulus in a particular way, that is, by making the stimulus the target of a saccade. In order to do so we collaborated with Charles W. Mohler of our laboratory in designing an experiment in which monkeys reacted to spots of light in a different way. Again the monkey began by pressing a bar that made a peripheral spot of light appear. Here, however, the spot was large, and the monkey earned a reward by releasing the bar when either the spot or the fixation point dimmed. The fixation point was small enough and its degree of dimming was slight enough so that the monkey could perceive the dimming only with its fovea. In contrast, the dimming of the peripheral spot was evident to the monkey even if the monkey attended to the spot with the periphery of its retina. The monkey tended, therefore, to look at the fixation point, and when it reacted to the dimming of the peripheral spot, we could infer it had attended to the spot even though its eyes had not moved.

The experiment was conclusive: the cells in the superior colliculus that showed an enhanced response when the monkey made the spot the target of an eye movement gave no such response when the monkey attended to the spot by releasing the bar. It followed that the enhanced response in the superior colliculus was not a necessary concomitant of visual attention. More likely the enhancement is part of the neural processes involved in the initiation of eye movement.

In the other main visual pathway, the one that leads into the cerebral cortex, a similar sequence of experiments yielded quite different results. Typically a cell in the striate cortex showed only a slight increase in its rate of discharge when a spot of light in its receptive field was the target of a saccade. It showed this slight increase for a saccade to some other part of the visual field. Moreover, it showed the increase when the monkey attended to the spot of light by releasing the bar without an eye movement. Since the enhanced response in the striate cor-




**ACTIVITY OF SUPERIOR COLLICULUS** when a spot of light appears on a screen depends on how the monkey reacts to the spot. For the experiments diagrammed here a cell in the superior colliculus was monitored as the spot was flashed in the cell's receptive field. In *a* the monkey kept its gaze directed at a central point of light (*fixation point*) and ignored the stimulus; the cell responded to the spot with a few bursts of activity (action potentials). In *b* the spot was the target for a saccade; the cell's activity was more pronounced. Two further experiments with the monkey failed to yield a similar enhanced activity. In *c* a spot outside the cell's receptive field was the target for the saccade. In *d* there was no saccade, but the monkey attended to a spot in the receptive field by releasing a bar when the spot dimmed.

tex is the same for any saccade, it does not show that the saccade is directed toward a stimulus in a given cell's receptive field. That is, the enhanced response is not spatially selective. Hence it cannot be related to visual attention. Perhaps it is related simply to how alert the monkey is. The monkey's state of alertness affects the brain's processing of all stimuli equally.

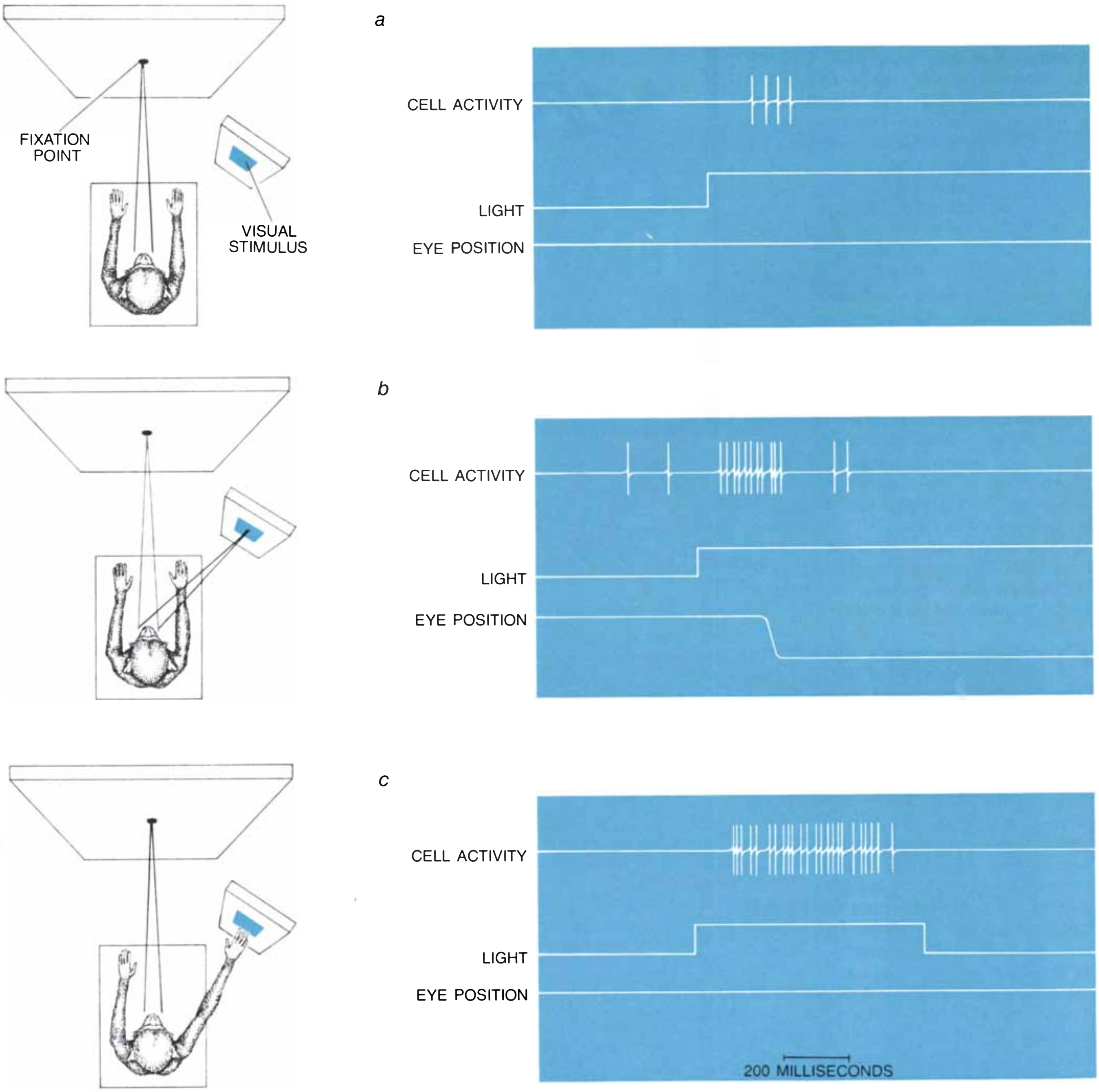
The results of the two sequences of experiments reinforce the differences already discovered between the two visual pathways. In the striate cortex the neural activity is strongly modulated by details of the pattern of light falling on the retina: the cells respond preferentially to slits of light at a particular place and orientation in the visual field. In addition the activation of some cells requires that the slit move in a certain direction or have a certain color. Our experiments show that these various responses are almost unchanged when the visual pattern is the target of a saccade or when it signals the monkey to release the bar. In other words, the striate cortex, the area of the cerebral cortex that is exquisitely sensitive to the details of the visual world, is essentially unaffected by the significance the details may have for the monkey.

The superior colliculus is different. Here the cells that respond to visual stimuli perform a rather rudimentary computation: in general their activation requires only a small spot of light somewhere in a relatively large part of the visual field. In the superior colliculus, however, the response of the cells is modulated quite clearly when the monkey makes the spot of light the target of a saccade. The visual processing may be rudimentary, but the cells are sensitive to the significance of the stimulus.

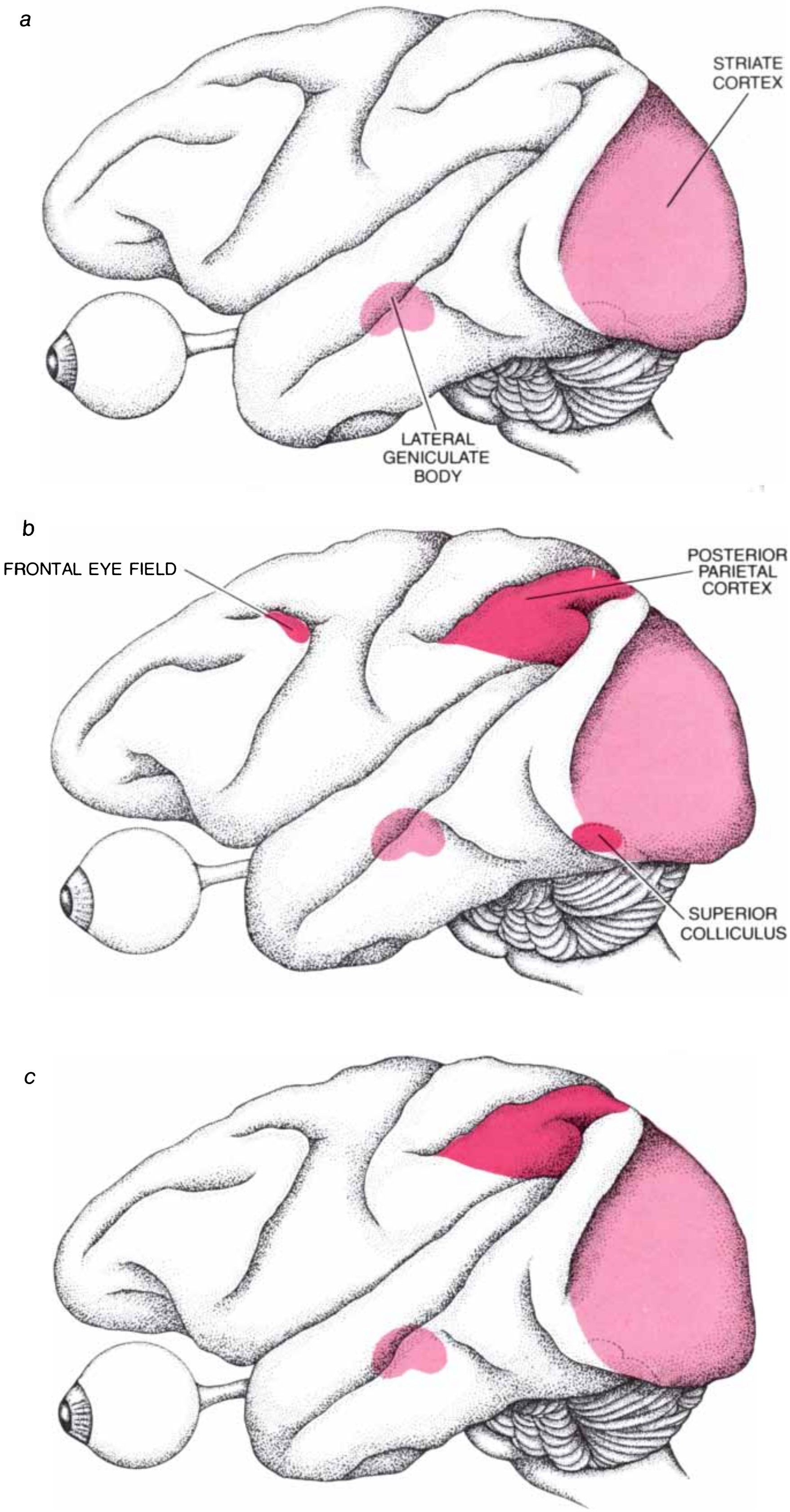
Our next experiments showed that cells in an area of the cerebral cortex respond to visual stimuli much as the cells in the superior colliculus do. The area lies in the frontal lobe in front of the motor cortex, a region important for the control of movements of the body. In the 19th century the Scottish physiologist Sir David Ferrier showed that in monkeys the electrical stimulation of this area of the cortex evoked eye movements. He called the area the frontal eye field. For some time investigators assumed that the frontal eye field controlled eye movements much as the nearby motor cortex controlled the movements of the rest of the body. Then, however, Emilio Bizzi, who was working at the National Institute of Mental Health, made electrical recordings of the activity of nerve cells in the frontal eye field. He could find none that reliably discharged in advance of a monkey's spontaneous eye movements. A few of the cells discharged during and after eye movements, but these cells could not, of course, be initiators of eye movement.

When we ourselves studied the frontal eye field, we found that many of the cells in the area discharge in response to small spots of light appearing anywhere in a receptive field that can be as large as a fourth of the entire visual field. (A receptive field of that size is larger than the receptive field of a typical cell in the superior colliculus and many times larger than the field of a typical cell in the striate cortex.) The nerve cells in the frontal eye field of the monkey showed no response anticipating spontaneous saccades in the dark. On the other hand, they showed an enhanced response when a spot of light in a given cell's receptive field was the target for a saccade. They did not show an enhanced response before a saccade that directed the gaze somewhere else or before the monkey attended to the spot without making a saccade.

It follows that these nerve cells, like the nerve cells in the superior colliculus, cannot account for visual attention. The reason is the same: attention can be dissociated from eye movement but these cells' enhanced discharge cannot. Perhaps, however, the nerve cells we have studied in the frontal eye field account for Ferrier's finding. When we apply a small electric current to a cell in the frontal eye field that is responsive to visual stimuli, a saccade moves the eyes to the center of the cell's receptive field. We assume the cell normally supplies visual information to the neural mechanisms that make the eyes move; the electrical stimulation passes along this chan-




**ACTIVITY OF POSTERIOR PARIETAL CORTEX differs from the activity of the superior colliculus (and also the striate cortex and the frontal eye field) in that enhanced activity was found when the monkey attended to the spot of light regardless of how it attended to it. In *a* the monkey showed no reaction; a posterior parietal cell responded to the spot with a few action potentials. In *b* the spot was the target for a saccade; the cell's activity was enhanced. In *c* the monkey touched the spot, and again the cell's activity was enhanced.**




**PATTERNS OF ENHANCED ACTIVITY in the areas of the brain responsive to visual stimuli can be surmised from the authors' experiments at the National Institutes of Health. Visual stimuli activate the pathways leading through the lateral geniculate body to the striate cortex (*a*). Visual stimuli that become the target for a saccade activate preferentially the superior colliculus, the frontal eye field and the posterior parietal cortex (*b*). Visual stimuli to which the monkey attends in any manner activate preferentially the posterior parietal cortex (*c*). This suggests that the posterior parietal cortex is part of the neural substrate for visual attention.**

nel of communication so that the eyes of the monkey move in a saccade as if there has been a visual target.

If the superior colliculus, the striate cortex and the frontal eye field do not mediate visual attention, what does? Clinical evidence implicates a part of the cerebral cortex designated the posterior parietal cortex. Specifically, people with damage to the posterior parietal cortex on the right side of the brain tend to ignore objects in the left half of their visual field. Such people can see the objects, but they do not attend to them. A striking example of this deficiency is seen in a series of self-portraits made by the German artist Anton Räderscheidt after he had suffered a stroke that damaged the parietal cortex on the right side of his brain [*see illustration on page 134*]. A self-portrait he made early in his recovery from the stroke omits the left side of his face.

Experiments done with monkeys by Juhani Hyvärinen and his colleagues at the University of Helsinki have shown that cells in the posterior parietal cortex are active when the monkey's behavior is guided by vision, for example when the monkey reaches for an object or tracks the motion of an object with its eyes. Experiments done by Vernon B. Mountcastle and his colleagues at the Johns Hopkins University School of Medicine have shown that cells in the posterior parietal cortex discharge just before eye movements, and Mountcastle has proposed that the activity of such cells is related to visual attention. On the basis of these efforts and the clinical evidence we decided (with Gregory Stanton and Catherine Bushnell of our laboratory) to make the posterior parietal cortex a subject of our own experiments. We found that cells there have receptive fields similar to the receptive fields of the cells in the frontal eye fields. The cells in the posterior parietal cortex respond best, however, to large stimuli, for example spots of light five degrees in diameter.

When our monkeys made such a stimulus the target for a saccade, the cells we were monitoring in the posterior parietal cortex showed an enhanced response. The response was spatially selective: we observed it when the saccade was directed to a stimulus in the cell's receptive field but not when the saccade went elsewhere. In this respect the enhanced response resembled the one we have noted in the superior colliculus and in the frontal eye field.

We next required the monkey to gaze at a fixation point and attend to the dimming of a peripheral stimulus by releasing a bar. We found an enhanced response. To confirm that this enhancement in the posterior parietal cortex is independent of eye movement we de-

# The lens that sets all standards in photography is standard on only one video camera.

Toshiba has been responsible for so many firsts in home video, its really no surprise that weve created the first video camera equipped with a Nikon lens.

Its the Toshiba IK-1900 color video camera, with a Nikkor 12.5mm ~ 100 mm, f1.6 macro zoom lens. And its going to change recorded history. Because rather than just revamp the video camera, Toshiba reinvented it.

The electronic viewing system has been eliminated (which, incidentally, decreases weight while increasing battery life by 25%) and replaced with our new True Image system.

You look directly through the lens at virtually 100% of the image, something even network studio cameras cant promise. Since you focus on precisely what you see, and not on a black and white electronic image, the subject literally shows its true colors.


Thanks to the True Image system, even the way you adjust the focus is superior. There is no split image to rob you of light.

Instead, a fresnel field system sharpens the image to crystal clarity with only the slightest rotation of the focusing barrel.

Working with Toshiba, Nikon also perfected macro/zoom capability all the way to 0.0 cm. focusing, a feature found on no other video camera. It means, although we dont suggest you try it, that you can actually touch the lens to the subject while keeping it in crisp focus.

Toshiba also gives you the highest resolution of any compact home video camera. And even a unique diopter adjustment feature, which allows people who wear eyeglasses to remove them while using the camera.

The Toshiba IK-1900 color video camera with the amazing Nikkor TV 8:1 macro zoom lens. Anyone interested in the very best home video should consider it standard equipment.

**TOSHIBA**
Again, the first.

Toshiba America, Inc., 82 Totowa Road, Wayne, NJ 07470

Simulated viewfinder image

**SELF-PORTRAITS** made by the German artist Anton Räderscheidt are evidence that the posterior parietal cortex is involved in brain mechanisms of visual attention. The first self-portrait (*upper left*) was made two months after Räderscheidt suffered a stroke that damaged the parietal cortex on the right side of his brain. Half of the face is omitted and half of the paper is blank. It is as if the artist could not attend to the left half of the world. The second self-portrait (*upper right*) was made three and a half months after the stroke, the third one (*lower left*) six months after the stroke and the fourth one (*lower right*) nine months after the stroke. The self-portraits were collected by Richard Jung of the University of Freiburg and are published through the courtesy of Gisele Räderscheidt, the widow of the artist.

vised another task. As before, the monkey was required to gaze at the fixation point. On some trials, however, the fixation point dimmed; the monkey released the bar to earn its reward. On other trials the monkey had to touch an illuminated panel without taking its eyes off the fixation point. If the panel was in the receptive field of the cell we were monitoring, the cell showed an enhanced response. The cell's activity was as intense as it had been when the monkey made a saccade to the panel.

The enhanced response of cells in the posterior parietal cortex is spatially selective, and it is independent of the particular action the monkey takes toward the stimulus. This makes it different from the enhanced response of cells in the superior colliculus and the frontal eye field. It also fits the clinical evidence that the posterior parietal cortex is involved in visual attention.

The phenomenon of enhancement suggests the outlines of a neural mechanism by which the monkey chooses objects from the visual world and makes them the targets for saccades. The retina can be taken to send code to stations in the brain describing the visual world. If the monkey is alert but not attending to anything in particular, the response of the nerve cells in these stations is relatively uniform. When the monkey begins to attend to some object, however, the nerve cells in the posterior parietal cortex that are related to the object because it is in their receptive field begin to discharge more intensely. If the monkey decides to make a saccade toward the object in order to examine it more closely, the cells in the superior colliculus and in the frontal eye field that are related to the object will also discharge intensely. The enhanced response has changed the uniform state of activity in the visual areas into a state in which the activity of nerve cells responding to one object is greater than the activity of nerve cells responding to other objects. This change happens even if the selected object is no more striking than the rest of the visual field. The only difference is that the monkey has decided to attend to the object.

We do not know how the enhancement arises, and in the absence of this knowledge it is tempting to say that the enhanced response amounts in itself to visual attention. It remains quite possible, however, that the enhancement is only a correlate of visual attention. That is, it may accompany visual attention, just as an eye movement may, but not be part of the neural mechanism whose product is attention. Nevertheless, the demonstration of brain activity whose enhancement is related to attention means we have at least begun to translate a psychological concept into physiological terms.

# The Lunar Society of Birmingham

*Science and technology in late 18th-century Britain were transformed by the activities of this provincial band of manufacturers, inventors and natural philosophers, who timed their meetings by the full moon*

by Lord Ritchie-Calder

In 1764 William Small returned to Britain from Virginia, where he had been Thomas Jefferson's professor of natural philosophy at the College of William and Mary, and began looking for a place in which to set up a practice as a physician. Benjamin Franklin, who was then in London to oppose the Stamp Act, gave Small a letter of introduction to his friend Matthew Boulton, a Birmingham manufacturer. The letter (which commended Small as "an ingenious Philosopher & a most worthy honest Man") was to lead to the formation of the Lunar Society of Birmingham, one of the most important coteries in the history of science and technology. The members of the society, in addition to achieving many individual distinctions, personified the conjunction of three great historical events: the American Revolution, the French Revolution and the Industrial Revolution.

The name chosen by the group came from their custom of meeting for dinner on the Monday nearest to the night of the full moon. They met at that time of the month so that the moon would light their way home. Perhaps inevitably, the members came to be known as the Lunatics. The three original members were Small, Boulton and Erasmus Darwin, each of whom had an important role in spreading the influence of the society.

Small was born in Scotland in 1734 and was educated at Marischal College in Aberdeen, where he learned mathematics, natural philosophy, classical languages, chemistry, anatomy, *materia medica* (pharmacology), surgery and midwifery; these were the prerequisites of a Scottish medical degree. At the age of 24 he went to William and Mary to teach mathematics and soon became professor of natural philosophy, expounding Newtonian physics. Small's teaching methods were popular with the students and an outrage to the faculty: he lectured in the common language instead of in Latin or Greek. One of his ablest students was Jefferson, who had been admitted to the college at the age of 16. Jefferson became the protégé of Small and George Wythe, a member of the Virginia bar. Jefferson ultimately chose the law and politics over natural philosophy, but in his autobiography he acknowledged that "Small probably fixed the destinies of my life."

The young Scottish professor may have done more than that. As Woodrow Wilson pointed out, the U.S. Constitution, which bears the imprint of Jefferson's thought, is based on a theory of political dynamics that owes much to the Newtonian view of the universe. The Constitution establishes a system of government in which action and reaction are equal and opposite, poised by the balance of forces acting on them.

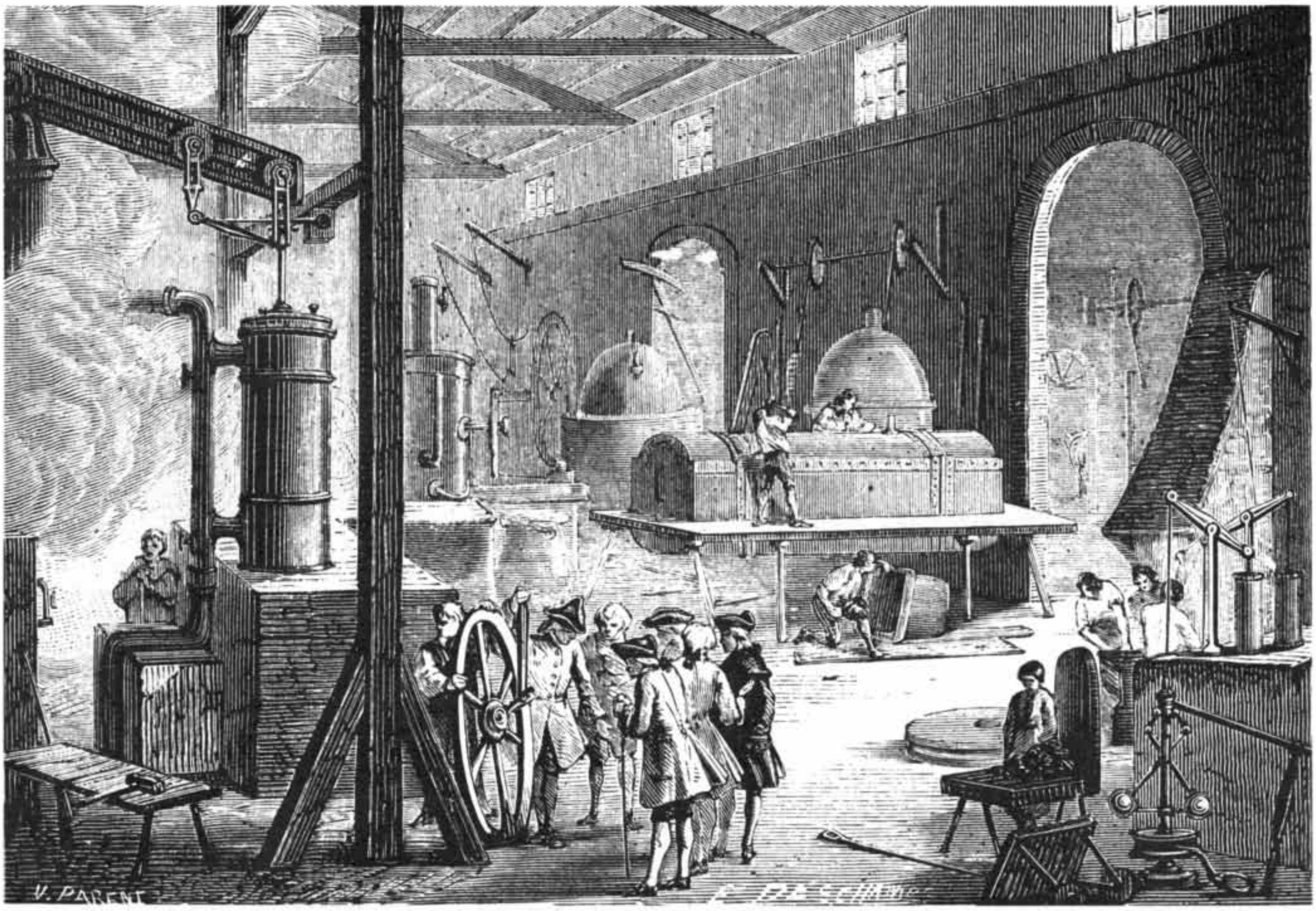
After six years in the colonies Small's health seems to have suffered (perhaps more from the climate of the faculty than from that of Williamsburg) and he returned to Aberdeen to complete his clinical training. With the help of Franklin and Boulton he then took up medical practice in Birmingham.

Boulton's friendship with Franklin dated from a few years earlier, when Franklin visited Birmingham to collaborate with Boulton on some electrical experiments. Thereafter they kept up a correspondence. Boulton's chief occupation at the time was the manufacture of various metal products, such as buckles for shoes and breeches, mainly for the profitable export trade. He also minted coins. By the standards of the day he was a wealthy man, having inherited a flourishing business and having married two coheiresses. (His second wife was his deceased first wife's sister.)

At about the time Small arrived in Birmingham, Boulton was expanding his business. He had built a large factory at Soho outside of Birmingham and was mechanizing it for mass production. His source of power was a stream he had harnessed to drive a water mill. In the summer there was not enough water, but Boulton had an ingenious idea: he would lift the water from the tailrace back into the mill pond. For that he needed a pump. He thought of a way of adapting the Savery engine, an early steam engine in which the steam itself acted as a piston, applying pressure directly to the surface of the water. In 1766 he sent his ideas (and a model) to Franklin in London, but by then Franklin had become too preoccupied with political matters to offer any direct assistance. Franklin's advice, however, would have won the approval of present-day environmentalists: he enjoined Boulton to arrange the grate of the boiler "in such a manner as to burn all your smoke."

Darwin, the third founding member of the society, was a prosperous, enlightened and humane physician who had studied medicine at the University of Edinburgh. He was also a poet, an ingenious if impractical inventor and an evolutionist 60 years before his grandson Charles Darwin upset the theologians by publishing *On the Origin of Species.* In *Zoonomia; or, The Laws of Organic Life,* Erasmus Darwin gave plausible reasons for believing species originated by transmutation. As Charles Darwin pointed out, however, his grandfather "anticipated the views and erroneous grounds of opinion of Lamarck," not the principle of natural selection. Through the daughter of a second marriage Erasmus Darwin was the grandfather of Francis

**TWO MAJOR ACTIVITIES of the members of the Lunar Society of Birmingham—technological innovation and philosophical discussion—are depicted in the wood engravings on the opposite page, which are reproduced from a 19th-century French book on the steam engine. The top engraving shows the interior of the Soho Manufactory near Birmingham, where James Watt and his business partner Matthew Boulton produced parts for their steam engines from 1775 to 1800. Both men were leading members of the Lunar Society. The bottom engraving shows a meeting of "le cercle des lunatiques" in Watt's house. The society took its name from the practice of meeting on the Monday nearest to the full moon (visible in the background).**


V. PARENT


YAN' D

Galton, the eugenicist and statistician. By all accounts Erasmus Darwin was a boon companion and a great conversationalist, in spite of a stammer, which he maintained was a convenience: "It gives me time for reflection and saves me from asking impertinent questions."

Over the years the Lunar Society had as many as 14 members, but no more than 10 ever sat down to dinner together at one time. The gathering customarily began at two o'clock in the afternoon and lasted until about eight. The meetings were held in the houses of various members, but in general it was Boulton who served the notice, typically as follows: "Pray remember that the full moon will be on Saturday, March 3rd." No record was kept of the proceedings lest the informality of the discussions be inhibited; there were no agendas, no transactions, no resolutions and no programs of action. (Boulton's one attempt to formalize the meetings, in 1776, failed.) What is known of the topics of discussion has been gleaned from the correspondence of the participants and from the published journals of Mary Anne Schimmelpenninck, née Galton. Mrs. Schimmelpenninck was the daughter of Samuel Galton (Francis Galton's paternal grandfather), a wealthy Quaker at whose home, Great Barr, the group often met.

It would be difficult to exaggerate the influence of the Lunar Society in terms of the men it brought together. They were a company of "merchants of light," as Sir Francis Bacon (in *The New Atlantis*) called the 12 Fellows of the House of Salamon, whose commission was to search everywhere for the facts of nature and of human experience. The role of the Lunatics in the Industrial Revolution was decisive. It is true that some events of the revolution had preceded them: the spinning and weaving inventions of Richard Arkwright and Edmund Cartwright had made possible the mass production of threads and textiles, and the enclosure of common lands had driven the landless proletariat into the mills and factories. The motive power of industry up to then, however, was still the water wheel. Boulton's dissatisfaction with the water wheel for his die-pressing operations was to have momentous consequences.

**WILLIAM SMALL (1734–75) was one of the three original members of the Lunar Society. A native of Scotland, he taught mathematics and natural philosophy from 1758 to 1764 at the College of William and Mary in Virginia. When he returned to Britain, he established a medical practice in Birmingham. There he became a scientific adviser and business associate of Boulton's, and he played a central role in setting up the firm of Boulton & Watt. This drawing, which hangs in the Birmingham Assay Office, is the only known contemporary likeness of Small.**

Franklin (a corresponding member of the Lunar Society) had not been very helpful on this occasion, but the versatile Erasmus Darwin knew someone who might be: John Roebuck, who had also studied medicine at the University of Edinburgh. Roebuck had abandoned medicine for industry, and with Samuel Garbett (another member of the society) had founded the first sulfuric acid factory, at Prestonpans on the outskirts of Edinburgh, and the Carron Ironworks, farther south in Scotland. Roebuck's activities had involved him with a scientific-instrument maker at the University of Glasgow named James Watt.

Watt, a native of Greenock, west of Glasgow, had only a parish-school education and had not yet completed his craft indentures as a mechanic. His technical skill and intellectual ability, however, impressed the professors at the university. John Anderson, who taught natural philosophy, gave Watt a model of the Newcomen engine to repair. In this early steam engine the steam was let into a cylinder, moving a piston, and the cylinder was then cooled, so that the steam condensed and the piston returned to its original position.

Watt recognized the inefficiency of this operating cycle and discussed the matter with another client and friend, Joseph Black, a lecturer in chemistry. Black explained to Watt something he had discussed so far only with his students: the theory of latent heat. The practical significance of the new idea became clear to Watt as he walked across Glasgow Green. "I had not walked further than the Golf House," he recalled, "when the whole thing was arranged in my mind: the waste of heat could be avoided by keeping the cylinder at steam heat and condensing the steam in a separate boiler." He made a model (including his wife's thimble), and Black was so impressed that he lent him £1,200 to continue his experiments. The result was the prototype of the condensing steam engine.

Roebuck, who was consulting Black about the production of synthetic soda (sodium carbonate) as an industrial chemical, heard about Watt's ideas. He was about to go into coal mining and saw the advantages of Watt's engine for pumping out flooded pits. He paid Watt's debt to Black and provided him with a workshop in which to construct a full-scale engine in return for a two-thirds share of the patent rights.

Meanwhile the impecunious Watt had to earn a living, which he did by working as a surveyor for the canal interests.

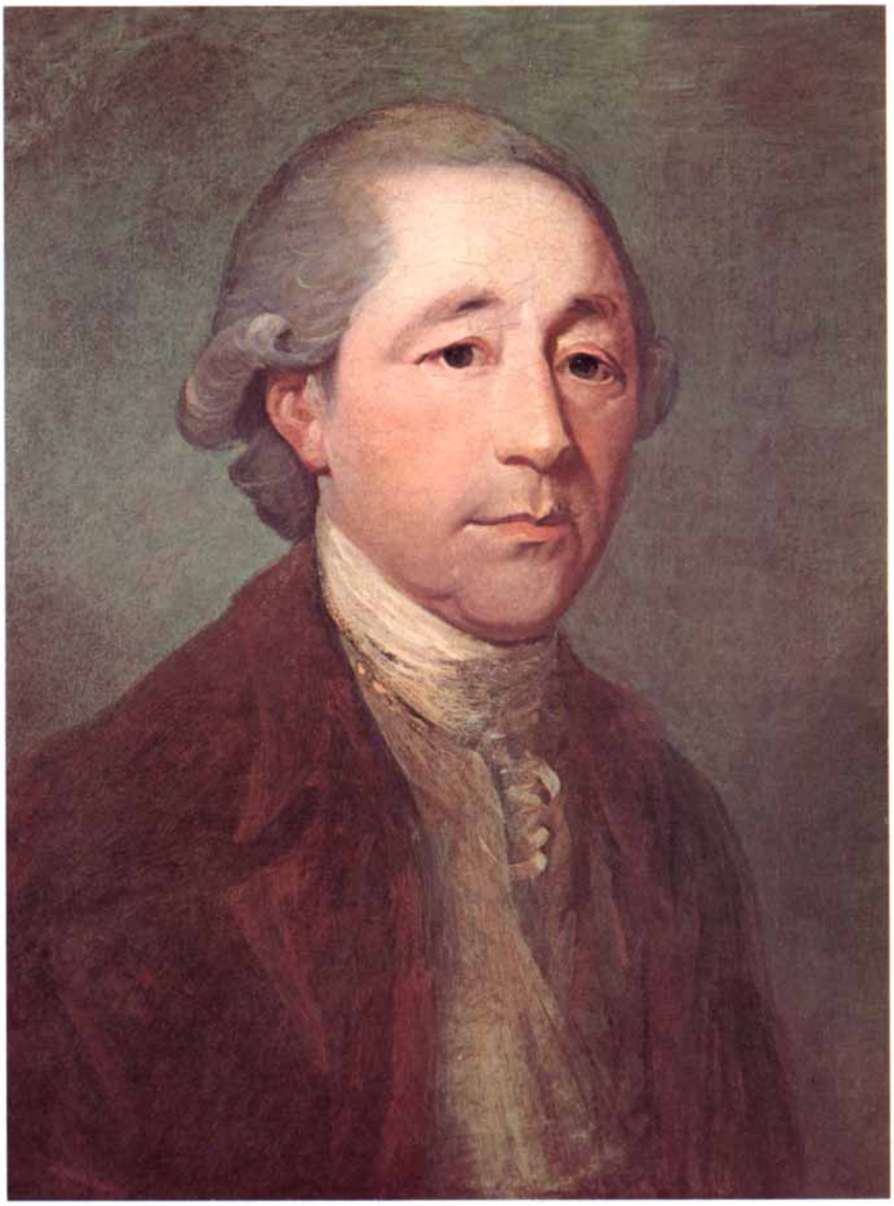
On his way back to Scotland from giving evidence before the canal commission in London he stopped in Birmingham, where Small showed him around the Soho works and introduced him to members of the Lunar Society. On a subsequent visit Watt met Boulton and inaugurated a historic association. In 1769 Roebuck was financially overextended and sold his interest in Watt's patent to Boulton and Small, who had become Boulton's partner.

The subsequent engineering success of the firm of Boulton & Watt would not have been possible without another member of the coterie: John Wilkinson, a cannon maker who built superior machine tools for boring gun barrels. It was Wilkinson who supplied the means of achieving the required precision in the cylinders of the Boulton & Watt steam engines. These engines, improved in various ways by Watt over the succeeding years, became the driving force of the later phase of the Industrial Revolution.

Watt soon became a regular member of the Lunar Society, which had already begun a period of significant growth in its intellectual capacity. Franklin had introduced his close friend Joseph Priestley to the group. Priestley, a teacher in a religious dissenters' school at Warrington in Cheshire, had become interested in Franklin's work on electricity and had made the long stagecoach journey to London to see him. They had interests in common not only in science but also in politics: Priestley was deeply engaged in support of the American colonists. At Franklin's instigation Priestley wrote *The History and Present State of Electricity,* published in 1767. The book included the first detailed account of Franklin's kite-and-lightning experiment in Philadelphia. (Franklin himself never committed the story to paper.)

Franklin secured Priestley's appointment as librarian to the Earl of Shelburne, Secretary of State for the Colonies, in whose house Priestley carried out the experiments that isolated oxygen (which he called "dephlogisticated air"). Years later, when Priestley's relations with Shelburne became strained—not over science or politics but over religion—it was Franklin who rescued him by writing to Boulton, just as he had done for Small. Boulton and his wealthy friends in the Lunar Society arranged to have Priestley appointed minister of the Unitarian church in Birmingham. They not only underwrote his stipend as minister but also arranged to secretly provide him with funds to carry on his research.

One of Priestley's secret sponsors was Josiah Wedgwood, the potter. Wedgwood, who was also a member of the Lunar Society, needed the moonlight more than most because he lived 40 miles away at Hanley in Staffordshire.



**MATTHEW BOULTON (1728–1809) was a successful manufacturer of metal products in Birmingham at the time he formed his association with Watt. He was also a founding member of the Lunar Society and generally notified the other members of the meeting dates. After Small's death Boulton made an unsuccessful attempt to formalize the society's proceedings. This oil painting, by an unknown artist, is now in the National Portrait Gallery in London.**

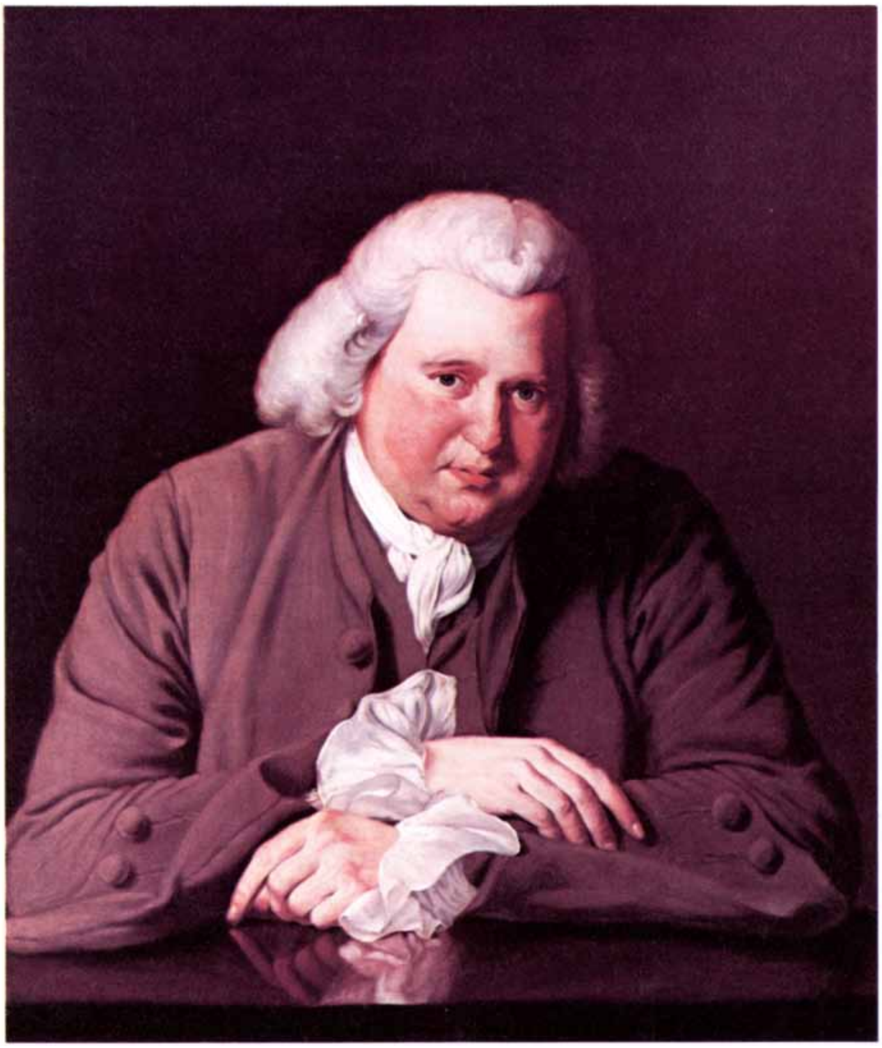
He had started in trade at the potter's wheel, but he had a brilliant and inquiring scientific mind, which he applied to solving the age-old mysteries of pottery. His participation in the Lunar Society was not mere intellectual virtuosity. While Priestley was obscuring his own discovery of oxygen in his persistent defense of the phlogiston theory, Wedgwood and a fellow Lunatic, William Withering, were describing in the Lunar Society discussions their findings on the significance of oxidation. Wedgwood was well acquainted with the rust-colored clays of domestic crockery. With Withering he set out to find nonferrous clays to use in making chinaware of high quality.

Withering is best known in the history of medicine as the physician who introduced digitalis as a treatment for dropsy, the accumulation of fluids that is commonly a symptom of heart disease. He picked up the idea on his rounds as a country doctor from a woman in Shropshire who administered infusions of foxglove (*Digitalis purpurea*) for cases of dropsy. Withering was called, in his time, "the flower of physicians." He was an epidemiologist who described and documented the spread of scarlet fever. He was a student of animal breeding. He was an accomplished musician on the flute and the harpsichord. He was a geologist whose name is perpetuated in mineralogy as the discoverer of witherite, a carbonate of barium that he distinguished from barite. (Witherite is used in the casehardening of steel, the refining of sugar, the manufacture of clay products, the enameling of iron and glass and as a paint filler.)

The clays and minerals Withering and Wedgwood found in their field work led to the development of the famous Wedgwood jasper ware and fine porcelains. Wedgwood later became a Fellow of the Royal Society, not for his work on the mineralogy, chemistry or aesthetics of pottery but for his invention of the pyrometer, a device he employed to measure the temperature of kilns. He was also to become (through his daughter) the grandfather of Charles Darwin and (through his son) of Darwin's wife, the former Emma Wedgwood. In addition to his many other distinctions the elder Wedgwood was an early worker in the field of study that would now be called operations research or systems analysis. He made comparative studies of the breakage of his crockery consignments when they were transported by road and by water. As a result he became a vigorous advocate of the canal system that soon laced Britain, with the Grand Trunk Canal running from the Mersey through the Wedgwood Potteries to the Trent and thence to London.

Another noteworthy member of the Lunar Society was James Keir. He had been educated at Edinburgh High School and had been a fellow medical student of Erasmus Darwin at the University of Edinburgh. After a period in the army he settled in Birmingham, where Darwin introduced him to Wedgwood and Boulton. Watt, after he came to know Keir in the Lunar Society, described him as "a mighty chemist and a very agreeable man." His major contribution to the Industrial Revolution was his scheme for the production of industrial alkalis.

Still another Lunatic was William Murdock, whom Sir Walter Scott described as "that madman who is proposing to light the streets of London with smoke." Murdock, a miller's son with a bent for engineering, was a latecomer to the society. After Watt had made his reputation in Birmingham, Murdock walked from Scotland to the Soho factory to join him. Boulton interviewed Murdock but was unimpressed by the awkward youth until he dropped his top hat. It made a strange sound. Boulton picked it up and found it was made of wood. Murdock explained that he had made it himself on a machine he had built for turning both internal and external ovals. He was hired; moreover, he soon lost his bashfulness and became a controversialist in the Lunar Society. His principal accomplishment was to develop methods for producing, purifying and storing gas made from coal. The Soho factory was illuminated by a blaze of gaslights to celebrate the Peace of Amiens in 1802. The firm of Boulton & Watt, with which Murdock was to remain for the rest of his life, did not encourage his other invention: a locomotive that would have anticipated Richard Trevithick's achievement of putting Watt's engine on wheels.

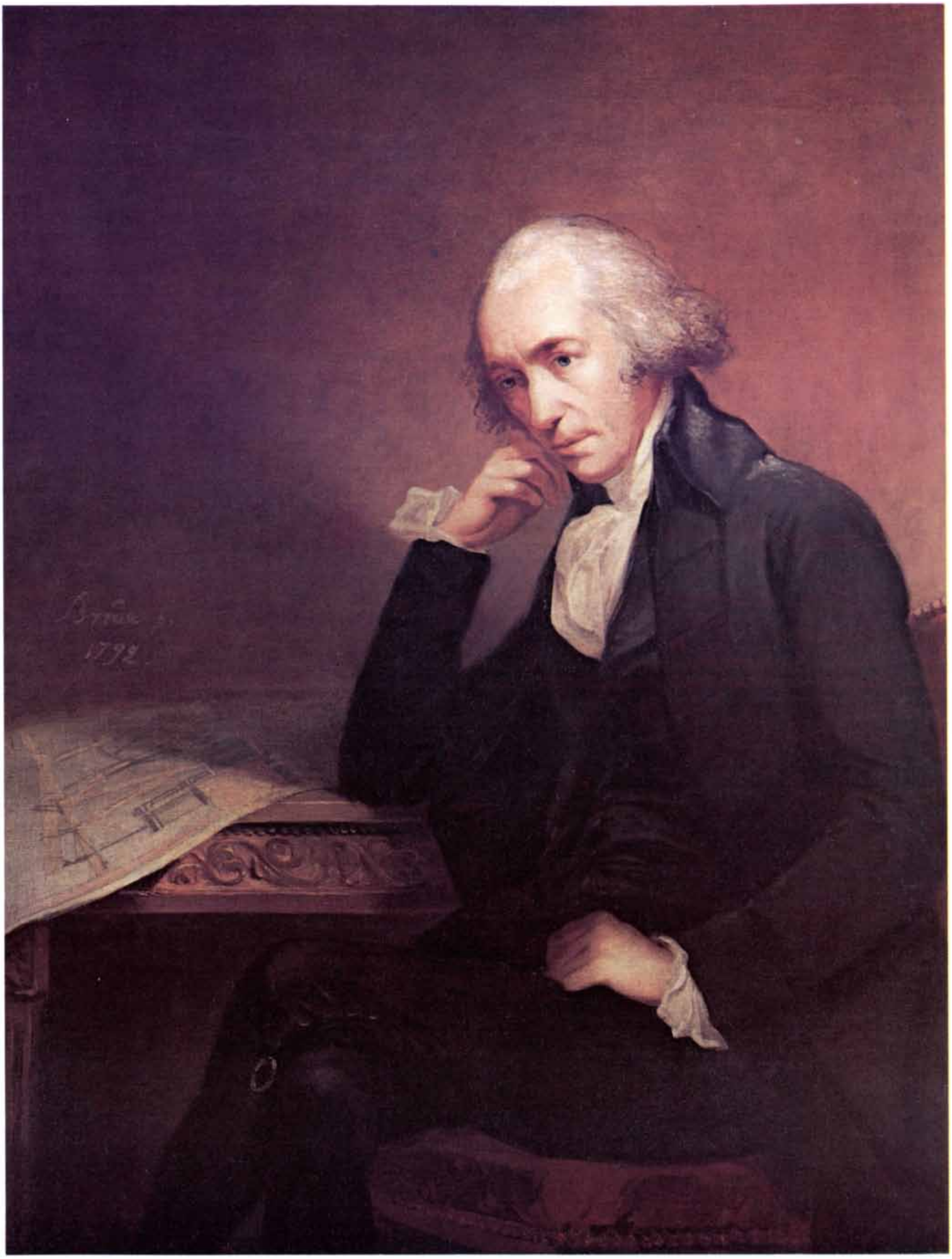


**ERASMUS DARWIN (1731–1802) was the third founding member of the Lunar Society. A physician, poet, inventor and writer on various scientific topics, he was by all accounts a lively contributor to the Lunar Society discussions. In addition he was the grandfather of two leading 19th-century scientists: Charles Darwin and Francis Galton. This oil portrait, painted by Joseph Wright in 1770, is also part of the collection of the National Portrait Gallery in London.**

On occasion a meeting of the Lunar Society must have seemed like a reunion of Scottish medical-school alumni: Small, Darwin, Roebuck, Keir and Withering. In addition Watt and Murdock were Scots, albeit without university education, and Priestley had received his doctorate in law from Edinburgh.

The association of medicine and industrial science can be explained, and so can the preeminence of Edinburgh. The universities of Oxford and Cambridge were then dominated by theological interests, although their teaching was sound in the mathematical sciences, including not only mathematics itself but also astronomy, mechanics and optics. Physics, however, was speculative and chemistry was empirical; the life sciences were a study of collections; the earth sciences were virtually nonexistent, still being circumscribed by Archbishop James Ussher's dating of the Creation in 4004 B.C. Only in two European universities was instruction given in both the theory and the practice of science: Edinburgh and the University of Leiden. In the medical schools of those universities as much attention was paid to the science of medicine as was paid to the art.

Earnest medical aspirants in 18th-century Britain therefore headed for Edinburgh or Leiden (or both). Both universities had been founded after the Reformation, Leiden in 1575 and Edinburgh in 1583. They were secular and open to all faiths. From the beginning there was a close association between them, but by the 18th century their teaching was virtually reciprocal. In



JAMES WATT (1736–1819) conceived the principle of the condensing steam engine in 1764, when he was an instrument maker at the University of Glasgow. He later moved to Birmingham to join Boulton and his associates, becoming a partner in the engineering firm of Boulton & Watt and a regular member of the Lunar Society. He is shown here, in another painting in the National Portrait Gallery in London, with a design for one of his engines on the table in front of him; the painting was done by Carl Frederick von Breda in 1792.

1692 Alexander Pitcairne, who founded the Royal College of Physicians of Edinburgh, was appointed to the chair of medicine at Leiden. One of his students was Hermann Boerhaave, who succeeded him at Leiden and whose lectures in chemistry were to have a profound effect on European thinking. By the end of the 18th century some 40 holders of chairs in Scottish universities had been educated at least in part at Leiden.

The medical faculty at Edinburgh provided a thorough grounding in the sciences. Many of the medical graduates found ways other than medical practice to apply their scientific learning. James Hutton, for example, took a medical degree but used the chemistry he learned to such good effect that he commercialized a process for making ammonium chloride; he never practiced as a physician but made a fortune and became one of the founders of the science of geology. Of the other Edinburgh medical graduates Roebuck gave the Industrial Revolution its sulfuric acid (among other things) and Keir gave it its alkalis.

Richard Lovell Edgeworth was a regular member of the Lunar Society. He was a wealthy landowner with estates in Ireland and England, but he found the "merchants of light" of the industrial Midlands congenial. His technical interests were modest: inventing, or rather designing, elegant carriages, notably the phaeton, a vehicle light in construction with large wheels, high seats and no doors. He was on visiting terms with Jean Jacques Rousseau, at whose instigation he wrote the book *Practical Education* in collaboration with his daughter Maria. He argued that education should be based on an understanding of the child's mind and should be treated as an experimental science in which conversation was noted down and analyzed. His 22 children (by four wives) provided the material for his own experiments.

**JOSEPH PRIESTLEY (1733–1804) was introduced to the Lunar Society by one of its corresponding members, Benjamin Franklin. In 1780 Priestley moved to Birmingham, where Boulton and his friends arranged to have him appointed minister of the Unitarian church and to provide him with funds to carry on his research. Priestley is best known as the discoverer of oxygen, but in debates with his fellow members of the Lunar Society he persisted in defending the phlogiston theory, which regarded the newly isolated gas as "dephlogisticated air." This pastel drawing, also in the National Portrait Gallery in London, was done by Ellen Sharples.**

Edgeworth's appraisal of the role of the Lunar Society was perceptive. He wrote: "A society of literary men and a literary society may be very different. In the one, men give the results of their serious researches and detail their deliberate thoughts. In the other, the first hints of discoveries, current observations and mutual collision of ideas are of important utility. The knowledge of each member of such a society becomes in time disseminated among the whole body, and a certain *esprit de corps,* uncontaminated with jealousy, in some degree combines the talents of the members to forward the views of a single person." He was making the point that the Lunar Society was not a "learned society" but a free-for-all in which there was no hierarchy of learning. There were heated discussions in which any of the members might start an inflammatory exchange without resentment on anyone's part.

This quality was illustrated by the society's discussions of the phlogiston theory, originally propounded in 1697 by the German physician and chemist Georg Ernst Stahl. The theory held that a hypothetical substance, phlogiston, is the principle of combustion: substances that burn are compounds of phlogiston, and combustion is caused by phlogiston's leaving the substance. Priestley maintained this position against Keir, who was a confirmed antiphlogistonist.

Because of Priestley's preoccupation with the "fiery principle" he did not appreciate the significance of his own findings, including what he described as "a mere random experiment made to entertain a few philosophical friends who have formed themselves into a private society of which they have done me the honour to make me a member—to wit, the Lunar Society." In the experiment Priestley had exploded "inflammable air" (hydrogen) and "dephlogisticated air" (oxygen) with an electric spark, causing a dew to form on the glass container.

In April, 1783, Priestley told Watt that the weight of the water thus formed was equal to the weight of the two gases. Watt immediately replied that the experiment showed water is not an ele-


Renault Fuego. From the first people to turbocharge Grand Prix racing and win on three continents. Wind-smooth styling that simmers with the performance of a fuel injected 1.6 litre engine. EPA's that give you 24 est. MPG, 36 est. hwy.*
Renault Fuego. An optional Turbo for performance that's more responsive yet offers EPA's of 26 est. MPG, 39 est. hwy.*
Renault Fuego. Aerodynamics more slippery than the Porsche 928. Agile front-wheel drive. Five forward speeds or optional three-speed automatic. And standard Michelin radials. All give
F
ar
New. Rac
RENAULT

ment but a compound. He subsequently wrote to Priestley: "What are the products of your experiment? They are water, light and heat. Are we not, thence, authorised to conclude that water is a compound of the two gases, oxygen and hydrogen, deprived of a portion of their latent or elementary heat; that oxygen is water deprived of its hydrogen but still united to its latent heat and light; if light be only a modification of heat, or a simple circumstance of its modification, or a component part of hydrogen, oxygen gas will be water deprived of its hydrogen but combined with latent heat."

This was an exchange between two cronies in a club. It was left for Henry Cavendish to describe the composition of water (still in terms of phlogiston) to the Royal Society in 1784, and for Antoine Laurent Lavoisier to show that Cavendish was wrong in thinking that the element water preexisted in the gases and to redefine it in terms of oxygen and hydrogen. Nevertheless, Priestley persisted in defending phlogiston, and as late as 1800 he published his *Doctrine of Phlogiston Established, and that of the Composition of Water Refuted*.

The Lunatics would range over everything—poetry, religion, art, politics, music, science—with minds unbuttoned like their breeches bands. Theirs was the spirit of universal inquiry in the days before the common ground of understanding was fenced off into faculty estates.

Their politics got them into trouble. The French Revolution was in progress and the members of the Lunar Society, who had been indulgent toward the colonists in the American Revolution, were likewise on the side of the revolutionaries in France. Watt's son had been denounced by Edmund Burke in the House of Commons as a French agent. Priestley forsook his chemistry and became a vigorous supporter of the National Assembly, which offered him French citizenship and nominated him as a member of the National Convention, an honor he declined. He too was denounced by Burke.

On July 14, 1791, Keir held a dinner in the principal inn of Birmingham to celebrate the second anniversary of the fall of the Bastille. It was attended by some 80 sympathizers. A mob gathered outside the inn shouting "Church and King!" and smashed the windows. Their main target, however, was Priestley, who was at home. They went on to the New Meeting House, where Priestley was the pastor, and set it on fire. Then they made for his house at Fairhill, a mile and a half away, but Priestley had been warned and escaped with his family. The mob destroyed his apparatus and set the house on fire. His manuscripts, the work of 20 years, were scattered about the countryside.

The other Lunatics were also targets of the mob. Boulton and Watt retreated to the Soho Manufactory and armed

**DESTRUCTION OF PRIESTLEY'S HOUSE at Fairhill in 1791 was provoked by his support for the republican ideals of the French Revolution. The mob formed on the evening of July 14 to protest a dinner being held in Birmingham to celebrate the second anniversary of the fall of the Bastille. Priestley did not attend the dinner but nonetheless soon became the main target of the rioters. Although he escaped with his family, his scientific apparatus was destroyed, his papers were scattered and his house was set on fire. Priestley never returned to Birmingham and later emigrated to America. This lithograph, reproduced here by courtesy of the Birmingham Public Libraries, is a copy of a painting attributed to a contemporary local artist named Exted, who is said to have drawn the scene on the spot.**

their workers to withstand a siege. The mob had found houses with well-stocked wine cellars, however, and were otherwise occupied. Priestley left Birmingham and took his family to London. He later emigrated to America, where he settled in Northumberland, Pa. From there he dedicated his treatise on *Experiments on the Generation of Air from Water* to his "valued friends, the members of the Lunar Society at Birmingham."

The Lunar Society had revitalized British science, which after the great scientific outburst of the mid-17th century identified with Newton and his contemporaries had gone into decline. The *virtuosi* who had founded the Royal Society, under the patronage of Charles II, had become the *dilettanti* of the first half of the 18th century. The Royal Society needed a transfusion of red blood (to compensate for blue blood) and urgent ideas, both of which came with the Industrial Revolution and the Lunatics who helped to make that revolution. The Royal Society revived when men such as Boulton, Watt, Priestley, Wedgwood and Franklin were imported from the Lunar Society.

In the meantime extramural institutions were burgeoning. In Glasgow, Watt's friend Anderson had established his "anti-toga" classes, which led to adult education for the working classes. The artisans attended in their working clothes (hence "anti-toga") and in large numbers. More than 200 attended his natural-philosophy course, which met for five hours per week on Tuesday and Thursday evenings. Anderson's initiative (now embodied in the University of Strathclyde) led to the establishment of mechanics institutes throughout Britain. It also served as an example for the Royal Institution in London, which was founded by Benjamin Thompson, the American Tory who became Count Rumford. It was established "for the teaching by regular courses of philosophic lectures and experiments; the applications of the new discoveries of science to the improvements of arts and manufactures and in facilitating the means of procuring the comforts and conveniences of life." It was in the gallery of the Royal Institution that Michael Faraday, then a bookbinder's apprentice, listened to the lectures of Sir Humphrey Davy.

By the turn of the century the Lunar Society was no longer active, but its influence continued. The Whig politician Francis Horner commented in 1809: "The impression which they made is not yet worn out, but shows itself, to the second and third generations, in a spirit of scientific curiosity and free inquiry, which even yet makes some stand against the combined forces of Methodism, Toryism, and the love of gain."

# THE AMATEUR SCIENTIST

*The essence of ballet maneuvers is physics*

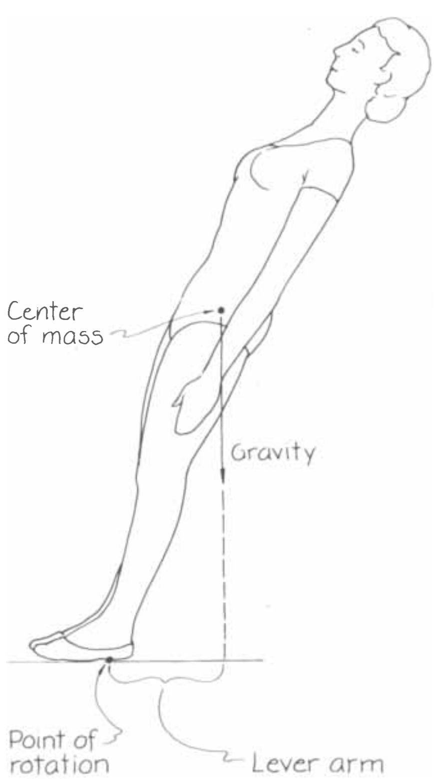
by Jearl Walker

Ballet is a blend of beauty and the physics of motion. A man who is interested in both components is Kenneth Laws, professor of physics at Dickinson College and a student of ballet with the Central Pennsylvania Youth Ballet. Much of the following analysis of certain positions and movements of ballet is based on information that he provided me with.

A large part of the early training of a ballet student is aimed at teaching her (or, to be sure, him) to maintain her balance as she moves gracefully through the ballet forms. Balance is achieved when an area of support lies in an appropriate place below the dancer's body. If the support area is off to one side, gravity pulls or turns her toward the floor.

Gravity, of course, pulls constantly on every part of the body, but the concept of a center of mass of the body helps to simplify the picture. The center of mass is a mathematical point whose position is determined by the distribution of mass within the body. The combined pull of gravity on all the parts of the body is said to operate through the center of mass.

When that center is not over a support area on the floor, the pull of gravity creates a torque on the body. A torque is the product of a force (here gravity) and a distance called the lever arm. Lines representing gravity and the lever arm appear in the illustration on this page, which shows a tilting human figure. The torque arising from the tilt tends to rotate the body about the feet and toward the floor. The greater the tilt, the longer the lever arm and the stronger the torque. If the dancer stood upright, the length of the lever arm would be zero and gravity would create no torque; the position would be stable.



*The gravity problem*

The routine for the beginning student seeks to develop a sense of balance gradually. For example, she learns the position called the arabesque in a simple first version and is then advanced to the first arabesque penchée. In the simple version (arabesque à terre) she puts her right leg forward and her left leg back with its toes touching the floor; her weight is therefore on the right leg. The right arm reaches forward, the left arm slightly back.

Moving the left leg to the rear shifts the center of mass in that direction. Without a compensating movement the dancer would fall over backward. To shift the center of mass back over the support area of the right foot she must lean forward and extend her right arm. This motion serves two purposes: it is graceful and it enables her to maintain her balance.

In the first arabesque allongée the dancer continues the motion until her torso and right arm are almost horizontal and her rear leg is tilted upward. In the first arabesque penchée her torso and right arm are tilted downward and her rear leg is tilted upward by 45 degrees or more. In both forms of the position the mass moved rearward by the rear leg must be matched by mass shifted forward by means of leaning the torso and extending the arm. Only then will the center of mass remain over the support area so that the dancer is stable.

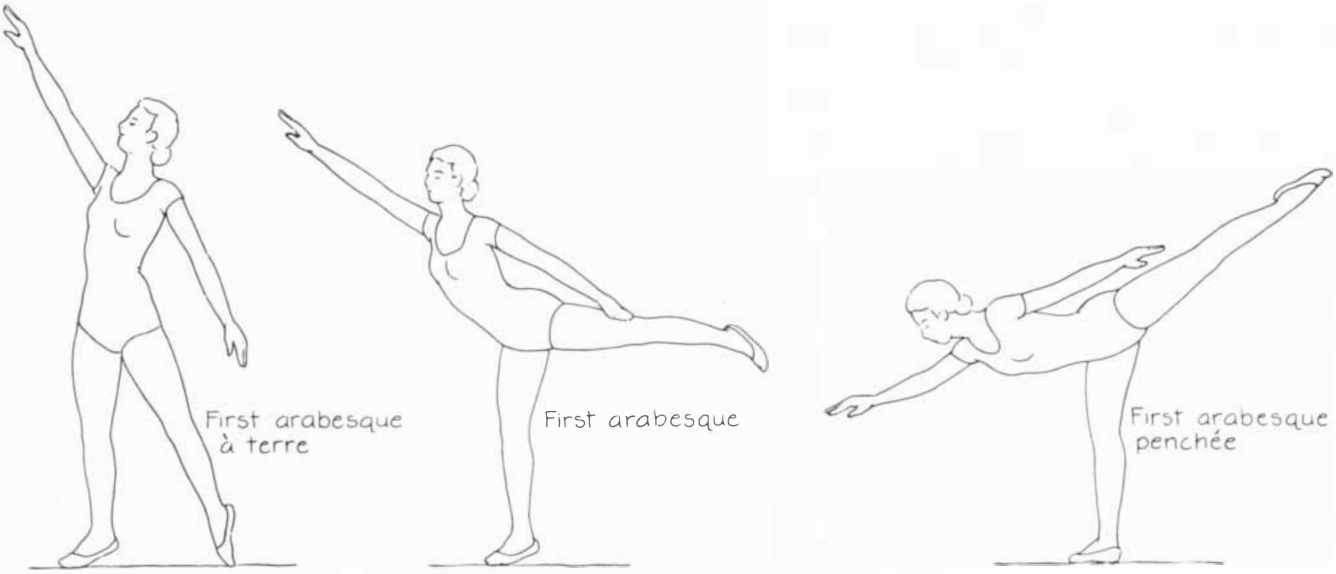
Some of the fascination of ballet stems from illusions in which physical laws seem to be momentarily suspended. An example is provided by the grand jeté, a forward leap. According to Laws, a properly executed grand jeté suggests that the gravitational pull on the dancer somehow weakens when she is near the top of the leap.

Two things contribute to the illusion. First, some slowing does take place near the top of even an ordinary jump because of the rules governing motion. Although the dancer's horizontal speed remains constant throughout the jump, her vertical speed is zero at the peak. Just before and just after that point her vertical movement is slow. As a result for about half of the time required for the entire leap she is within a fourth of the leap's maximum height.

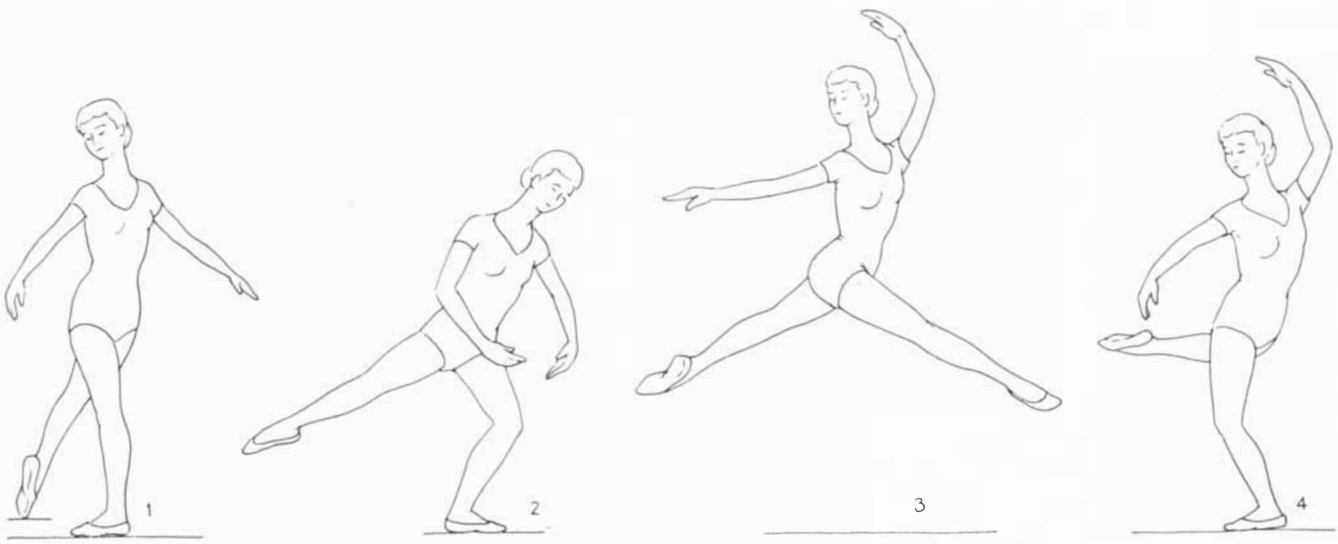
In a grand jeté another element is added to enhance the illusion that the dancer is hovering near the peak. Because of a certain shift of her arms and legs while she is in the air her path appears to be flattened at the top. The illusion depends on a shift of her center of mass as she moves her arms and legs. She launches herself with her arms downward. As she approaches the top of the leap she lifts and spreads her legs and arms. Her center of mass is then at a higher position in her body. Since the center of mass follows a fixed trajectory, the shift means the head and torso do not rise as far above the floor as they would have risen. In descending the dancer lowers her legs and arms, restoring the center of mass to its usual position.

In jeté en tournant, a turning leap, the dancer jumps into the air with no apparent spin around her vertical axis, yet near the top of the leap she begins to rotate. Impossible. One of the firm rules of physics is that the angular momentum of an object remains constant unless a torque acts on the object. If the dancer was not spinning when she left the floor, she cannot begin to spin in midair.

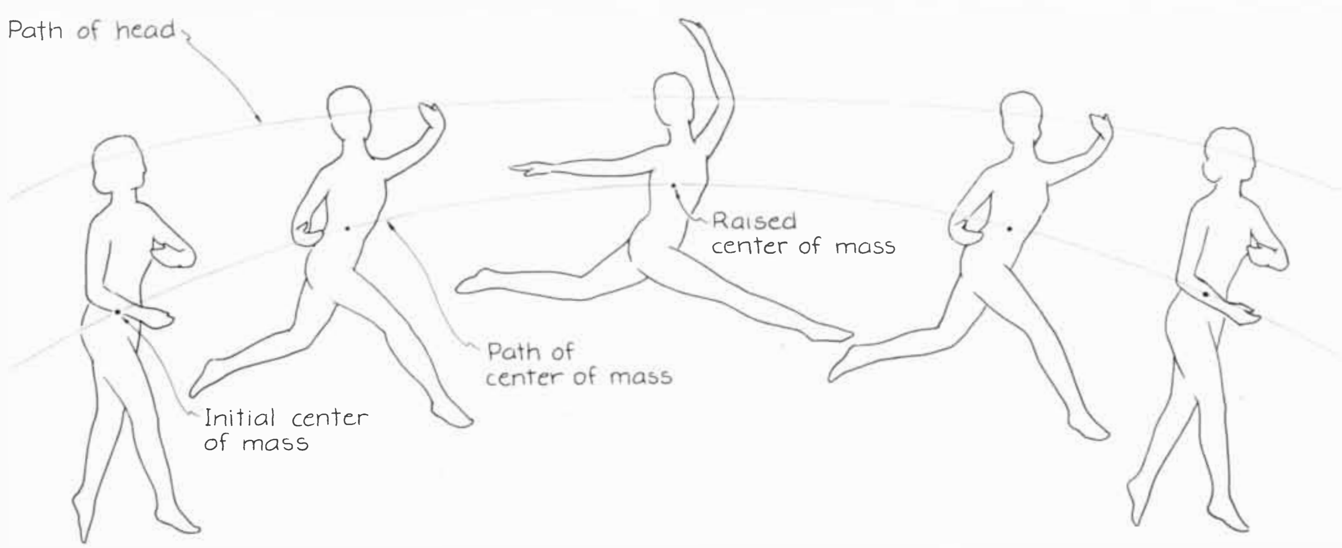
The explanation of the illusion is that the dancer does have a small spin at the beginning of the leap because of a torque she receives from the floor as she launches herself into the air. The spin is too slight for an observer to notice it. As the dancer rises, however, she pulls in her arms and brings her legs together. The effect is to decrease her moment of inertia. Since her angular momentum is fixed once she is in the air, the decrease in her moment of inertia increases her spin rate. One sees the same thing when



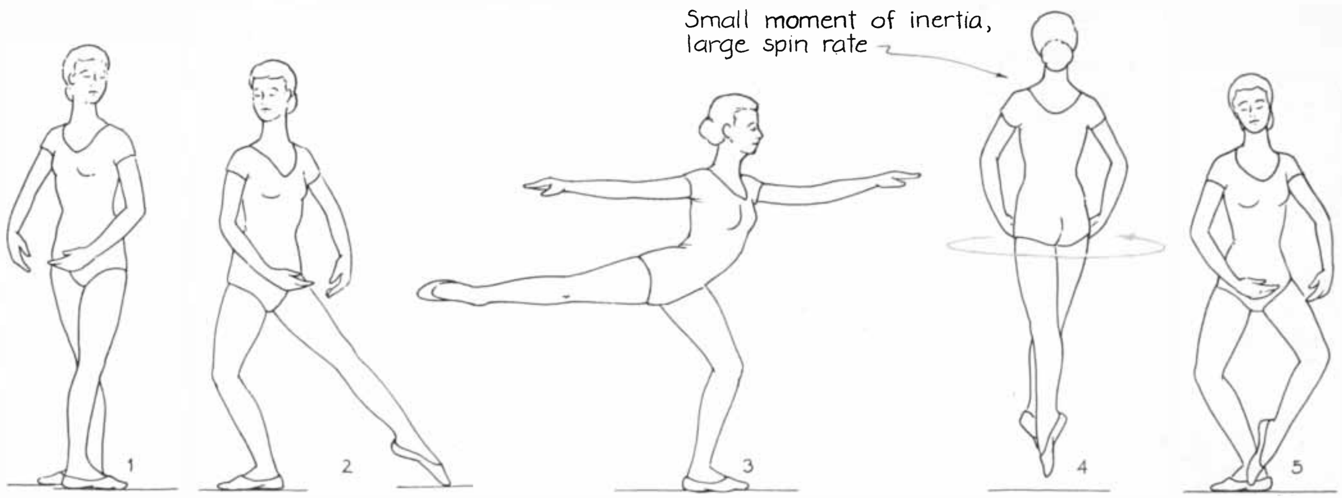

*Three forms of the arabesque*




*The grand jeté*




*How the dancer's center of mass shifts in the grand jeté*

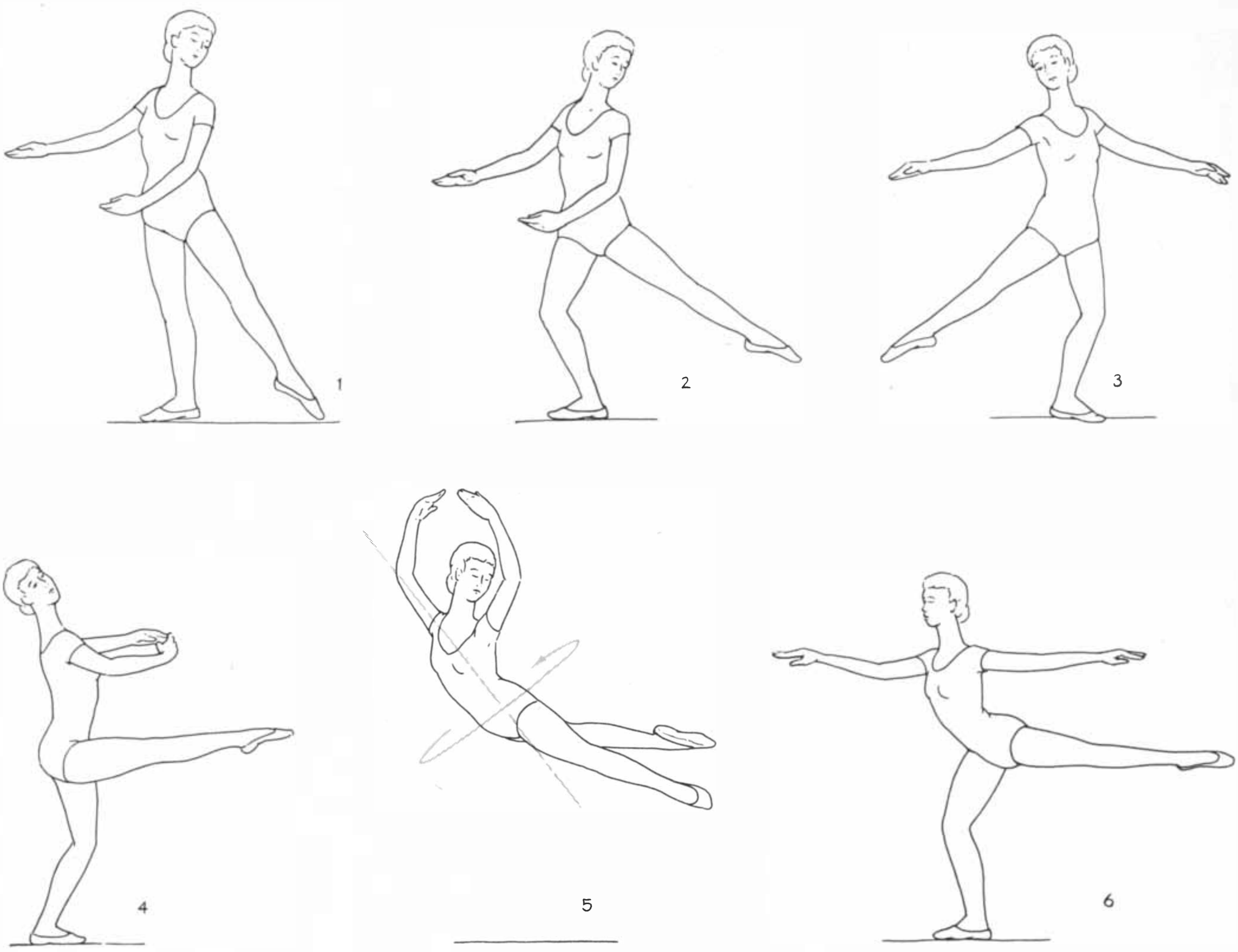



*The jeté en tournant*

a spinning figure skater pulls in her arms and so spins faster.

The dancer begins a jeté en tournant with her feet in what is called the fifth position. Her feet are parallel and pointing to the side in opposite directions, with the left foot in front. Each heel is near the toes of the other foot. The dancer slides the left foot to the left and quickly comes into an arabesque. She bends her left knee in what is called the demi-plié position.

Now she quickly brings her right leg downward and forward, simultaneously launching herself with her left leg. In flight she not only rises above the floor but also travels horizontally. At the beginning of the leap her arms and one leg are extended, providing the basis for her subsequent movements to increase her rate of spin.




*The grand jeté en tournant entrelacé*

# SCIENCE/SCOPE

Besides providing cloud pictures every 30 minutes day and night, Japan's newest weather satellite serves as a vital link in monitoring tides, tidal waves, and ice conditions in the western Pacific Ocean. GMS-2 (Geostationary Meteorological Satellite) relays data from remote sensors on buoys and fixed land locations to weather centers. The sensors transmit data at regular intervals, when readings exceed certain limits, or when they receive a command signal relayed through the satellite. GMS-2 was built by Hughes for and with the help of Nippon Electric Company. It was launched on Japan's N-II rocket, becoming the first geosynchronous satellite built in the U.S. and launched in another nation.

Ambitious manned missions in space will become possible with development by NASA of advanced solar-power platforms capable of electric-power output as high as 100 kilowatts or more. Hughes, under contract to the NASA Lewis Research Center, is building a breadboard 25-kilowatt dc-to-dc power converter that uses a transistorized series-resonant inverter. This module provides the basic building block required to match solar-array characteristics to payload requirements. The 25-kilowatt converter will utilize technology demonstrated earlier by Hughes in development of a lightweight 10-kilowatt converter that operated at over 91 percent electrical efficiency.

A new software system can translate naval tactical messages into understandable form. Messages within a command, control, and communications ($C^3$) system are typically hard to understand because they are transmitted in telegram form and often omit subjects, direct objects, articles, prepositions, and punctuation. If grammatical errors creep in, messages can be rendered unintelligible. While conventional computer techniques can't make sense of a garbled message, a Hughes message understanding system called GRACIE can. Using artificial intelligence techniques, GRACIE understands general descriptions of flights of aircraft over ships, of attacks, and of encounters with hostile ships. It constructs grammatical sentences based on what it expects messages to be, referring when necessary to a "rule book" of examples. It can be adapted for other than naval use.

Small and disadvantaged businesses support U.S. space and defense efforts by supplying parts and serving as subcontractors. Hughes has been a leader in tapping the skills and capabilities of these increasingly important businesses. In fiscal year 1981, Hughes awarded 70.6 percent of all purchase orders and 47.3 percent of the company's procurement dollar to small business. During the same period, $38.2 million went to disadvantaged business.

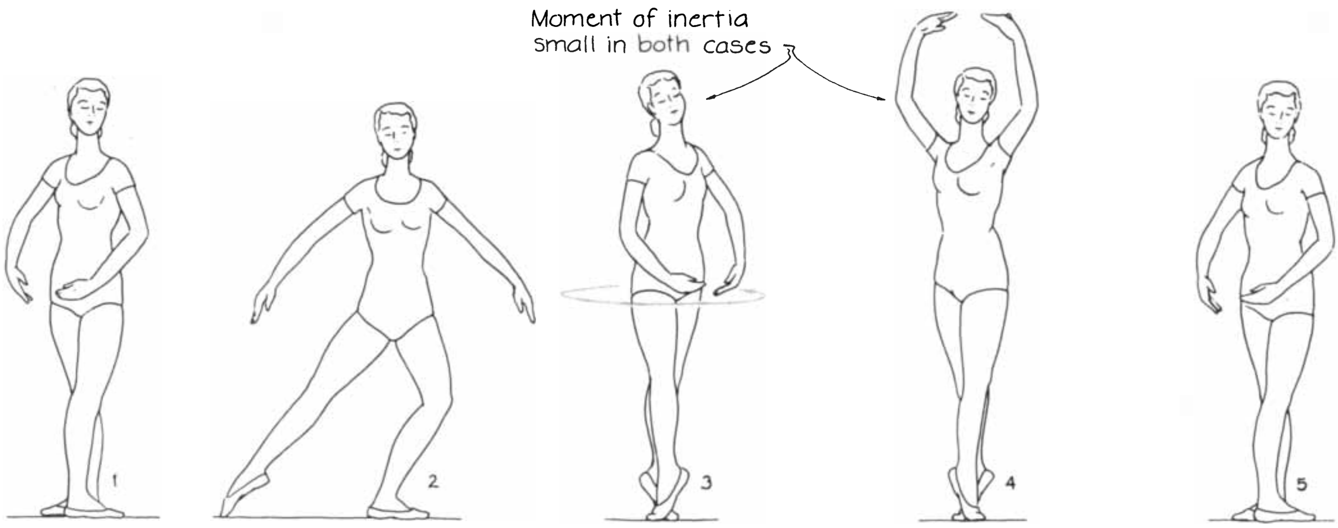
Hughes is seeking engineers to develop advanced systems and components for many different weather and communications satellites, plus the Galileo Jupiter Probe. Immediate openings exist in applications software development, data processing, digital subsystems test, microwave/RF circuit design, power supply design, digital communications, signal processing, spacecraft antenna design, system integration test and evaluation, and TELCO interconnection. Send your resume to Tom W. Royston, Hughes Space & Communications Group, Dept. SE, Bldg. S/41, M.S. A300, P.O. Box 92191, Los Angeles, CA 90009. Equal opportunity employer.

*Creating a new world with electronics*

**HUGHES**

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

(213) 670-1515 EXTENSION 5964




*The soutenu en tournant*

The dancer lands on her right leg and in demi-plié. She should be facing the audience. With experience she learns to control the leap and her moment of inertia to achieve this orientation.

An equally beautiful but more difficult leap is the grand jeté en tournant entrelacé (also called the grand jeté en tournant or simply the tour jeté). The dancer starts with her right leg in demi-plié and her left leg lifted to the side at an angle of about 45 degrees with the floor. Next she steps backward onto her left leg, bringing her right leg upward and around to the front so that it points to her left just as she turns her body to the left. She jumps from the left leg. Once she is in the air she rotates about an axis that tilts from the vertical. She holds her arms above her head and close to the rotational axis, also bringing her left leg up near her right leg so that they both rotate about the axis. She lands on her right leg and comes into the first arabesque.

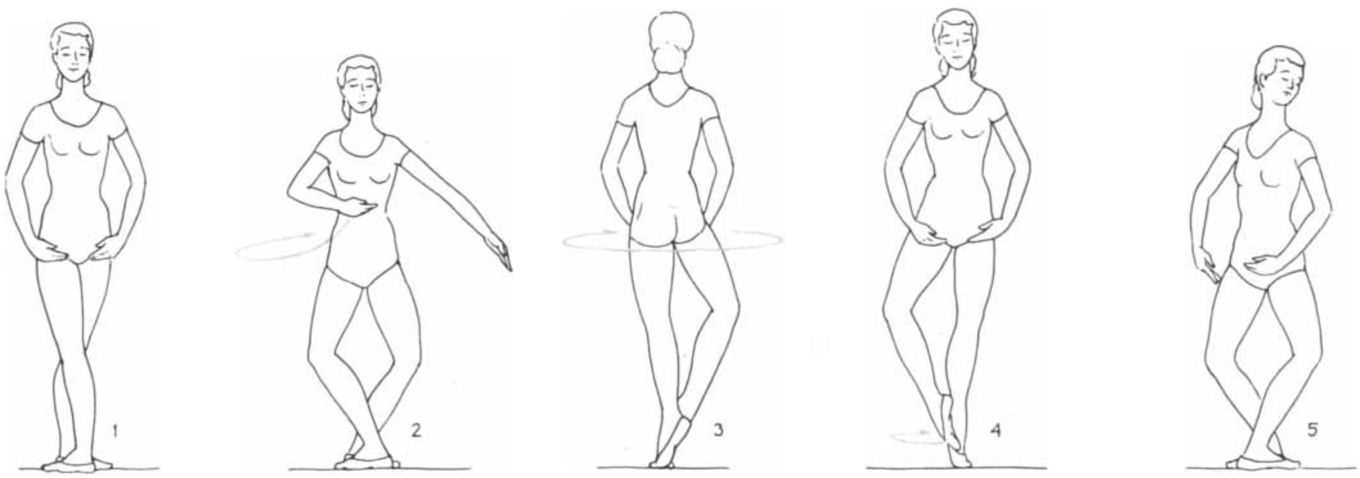
The rotation of the dancer in midair is rapid because her moment of inertia is small. As soon as she leaps she brings her arms and legs into line with the rotational axis. Once she has made her turn she effectively stops her rotation by spreading her arms and extending her right leg for her landing.

Classical dance has many techniques for whirling. One of them is embodied in the soutenu en tournant. Starting with her feet in the fifth position, left foot in front, the dancer goes into a demi-plié with her weight on her left leg and her right leg extended to the side. She brings her right foot back to the fifth position in front of the left foot and rises on her toes, twisting her feet to develop a torque that begins her twirl to the left.

A pirouette is a more ambitious twirl in which the propulsion also comes from torques on the feet. Consider first a quarter-turn pirouette. From the fifth position (right foot in front) the right foot is moved to the side while the arms are brought forward and then spread out to the sides. Next the right hand is moved to the front, the right foot to the rear. Now the dancer pushes against the floor with her right foot to propel herself around to the right. Simultaneously she rises on her left foot, which acts as a pivot.

The rotation brought about by the torque from the right foot is helped by the fact that the dancer draws in her left arm. The movement not only adds grace to the turn but also decreases the moment of inertia, enabling her to do the quarter turn rapidly. After dropping back into the fifth position she can immediately begin another quarter turn.

A complete pirouette differs mainly in the motion of the head and in the strength of the angular acceleration. A larger torque is required for a full turn. At the beginning of the rotation the dancer continues to face straight ahead. When her body has turned about 90 degrees, she suddenly whips her head in the direction of the turn, bringing it sharply to the front again when her body has turned about 270 degrees. Soon




*Portions of a pirouette en dehors*

thereafter her body is also facing forward once more.

A grand pirouette calls for the dancer to turn with one leg and both arms held horizontally to the side. A mathematical analysis of this pirouette is difficult unless the shape of the human body is simplified. Laws has done so in a model that appears in the illustration at the right. The "body" consists of an upper section of mass $M$ and length $L$ and a two-part leg section. The "shank" and the "foot" have a combined mass of $m/3$ and a length of $L'/2$; the "thigh" has the same length but a mass of $2m/3$. For a male dancer the ratio $M:m$ is about 3.8 and $L'$ is approximately equal to $L$.

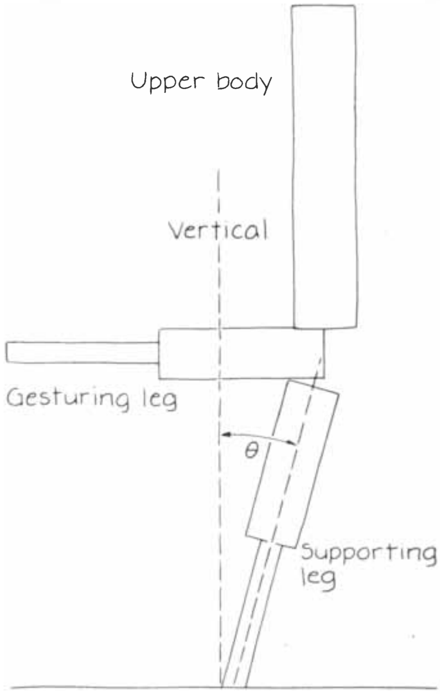
The upper body is vertical throughout the maneuver. In the model of the pirouette one leg is held horizontally to the side. The supporting leg is rigid and makes an angle theta ($\theta$) with the vertical axis through the supporting foot.

Laws was interested in the value of the angle required for a stable spin around the vertical axis. He first calculated it for a stationary dancer. As I have mentioned, the requirement for stability is that the dancer's center of mass be over the point of support. Since one of her legs is held horizontally to the side, she must tilt her other leg in order to shift mass and so reposition the center of mass. Stability is achieved when the angle between the supporting leg and the vertical is about 4.4 degrees.

Is it the same when the dancer is rotating? I would have thought so, but Laws's calculations indicate that the angle is instead about 3.5 degrees. Rotation imposes additional constraints on stability. The upper body must be closer to the vertical axis and the extended leg must stretch farther from the axis. Hence the dancer's center of mass is slightly off the vertical axis, and yet she is stable.

A more complicated rotation is part of the fouetté turn. (In Act III of *Swan Lake* the Black Swan does 32 consecutive turns.) The turn is essentially a full pirouette except that the source of the torque on the dancer is hidden. She continues to turn as though she were a toy top propelled by magic.

Once the pirouette has been started the right leg is not returned to the floor until the end of the turn. During most of the pirouette the right foot is held near the knee of the supporting left leg. The dancer is on the toes of her left foot. Just as her body begins to face forward again she thrusts her right leg forward and opens her arms to the audience. The heel of the left foot is brought down to the floor and the left leg is bent. Then the right leg, still pointed outward, is brought around to the left side, continuing the rotation, while the rest of the body remains facing the audience. When the right leg has been thus rotated through about 90 degrees, the right foot




*A simplification of the grand pirouette*

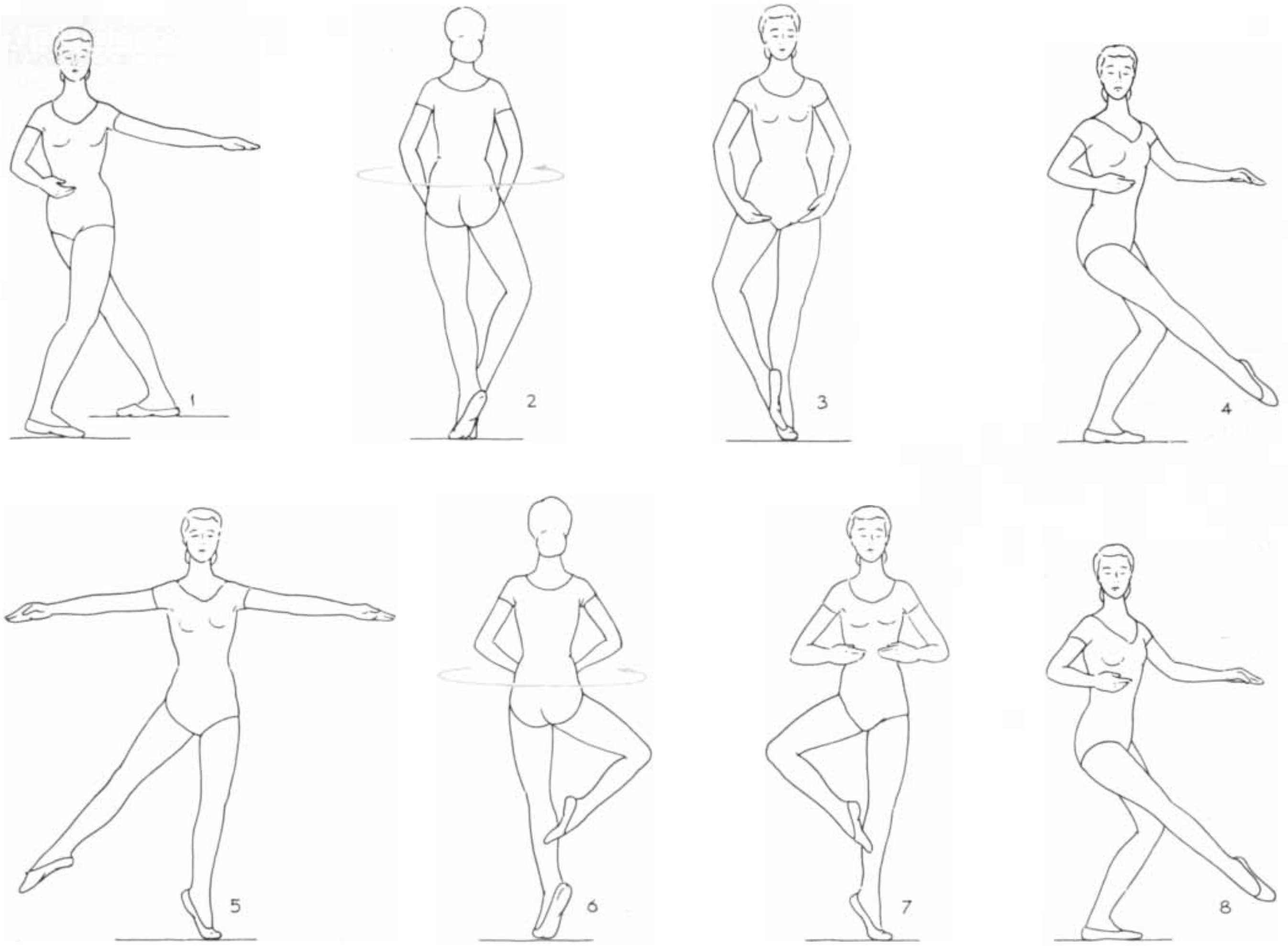
is brought back to the left knee and she comes back en pointe (on her left toes). She then repeats the entire procedure.

The key question is how the dancer generates enough torque to continue the rotation. With each turn the friction between the supporting foot and the floor robs her of angular momentum and spin. The secret is in the movement of the right leg. When it is brought forward and rotated to the right, it takes up whatever angular momentum the dancer still has. Her rotation stops, except for the right leg, giving her a moment to put her left foot flat on the floor. In this moment she can push against the floor, providing the torque for another rotation. To assist it she draws her right foot toward her left knee to reduce her moment of inertia as she spins.

The fouetté is difficult for dancers. Laws points out that a novice is likely to botch the turn by thrusting the right leg directly to the side instead of positioning it so that it absorbs the angular momentum properly. The novice might also extend the leg only partly forward toward the audience. It must be fully extended to absorb all the angular momentum.

The moment of inertia of the extended leg is about 1.7 times more than the body's. Hence when the leg absorbs the body's angular momentum, it does not turn as rapidly as the body. If the normal pirouette is at the rate of two revolutions per second, the extended leg turns at only 1.2. If the leg absorbs all the dancer's angular momentum, she has about .3 second to push off for another turn.

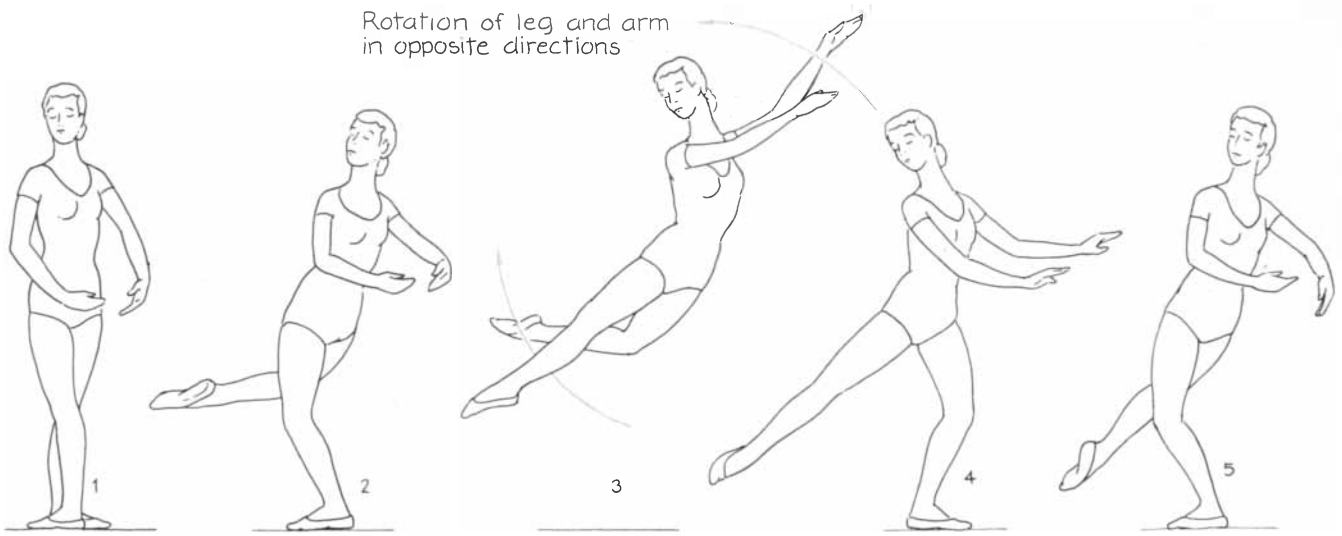
In the grand pas de chat the applica-




*A fouetté en tournant with two discrete turns*

tion of rotational dynamics is needed to enable the dancer to leap while she maintains the orientation of her body. From the fifth position of the feet she does a demi-plié on the right foot, bringing the left leg behind the right one at an angle of about 45 degrees from the vertical. The jump is made from the supporting leg. Once the dancer is in the air she brings the right leg into line with the left leg.

This alignment of the legs requires a rotation of the left leg about the dancer's center of mass. Yet her angular momentum in midair is essentially zero. How does she manage a rotation while keeping her angular momentum constant and maintaining the orientation of her torso? She rotates her arms in the direction opposite to the movement of her left leg. The arms are moved just enough




*The pas de chat*

to make the combined angular momentum of the arms and the right leg zero. By means of these countering rotations the dancer can rotate part of her body in the air.

Laws has analyzed the subtleties of balance achieved in the promenade en attitude derrière. Here a female dancer is in the position derrière en pointe: raised on the toes of one foot with the other leg poised in the air. One of her hands is held by a male partner, whom she faces. The other hand is raised in a graceful arc. The sustained pose is difficult because of the balance required. If the female dancer shifts into this pose from another movement, she is likely to be off balance.

The dancer could correct her balance by stretching her body in order to shift the center of mass back over the supporting foot, but that would disturb the grace of the movement. She could also push or pull her partner's hand, but that would be likely to generate a torque that would make her turn on her supporting foot, rotating her away from her partner when she should face him.

A better method is to achieve the proper application of forces between the touching hands. According to Laws, both dancers' hands should be horizontal and the elbows should be raised. Then the female dancer can apply forces to her partner's hand in such a way that a turning torque does not develop. The forces on her hand are indicated in the illustration at the right. Two oppositely directed forces come from the male dancer's hand, each one displaced from the center of the hand by a distance ***d***. The center of the hand is at a distance ***D*** from the vertical axis running through the female dancer's supporting foot.

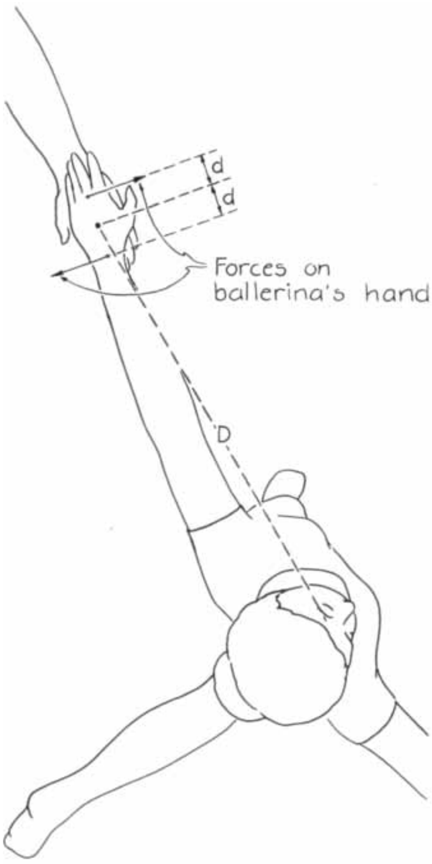
The forces on the hands generate torques on the female dancer, but they tend to turn her in opposite directions. By properly controlling the forces she can make the two torques cancel. She makes the nearer force slightly larger than the farther one. Since torque is the product of a lever arm and a force and the lever arm to the nearer force is smaller, the torques cancel. The forces on her hand therefore do not cause her to rotate about her pointed toes. By twisting her hand in just the right way, however, she can move toward or away from her partner's hand, thereby adjusting the position of her center of mass in such a way that the audience does not realize what she is doing.

Several of the male dancer's leaps call for his beating his calves together while he is in flight. One is entrechat quatre, which begins with the feet in the fifth position. The dancer descends into a demi-plié and then propels himself directly upward. In the air he opens his legs to the side, brings them briskly together, separates them and then brings them together again as he lands in the fifth position demi-plié. An experienced dancer with a powerful leap might be able to complete two beats or more.

Laws has investigated the difficulty this movement presents for a large dancer. In general such a dancer cannot beat his legs together at the rate and amplitude that can be achieved by a smaller dancer. Even though the larger dancer may be stronger, his legs are more massive and therefore more difficult to rotate. He also has more mass to lift and more difficulty leaping to the same fraction of his height.

I have only touched on the movements of classical dance. You could do a good deal more. Try analyzing a ballet movement. It would be helpful to make a series of photographs with stroboscopic flash. A motion picture that can be run slowly or stopped periodically would be even better. None of these techniques will be easy because the shape and motions of the human body are complex. You might simplify the shape of the body as Laws has done in some of his analyses.

A study of ballet is also complicated by the requirement of grace and style. If you want to determine the physical principles underlying certain movements, you must distinguish the components of the movements that are done only for style. Both Laws and I shall be interested in hearing about what you find out.




*The balancing forces in derrière en pointe*

# 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


The Antinomy of Parliamentary Sovereignty. Ilamar Tammela in *Archiv für Rechts- und Sozialphilosophie,* Vol. 44, pages 495–513; 1959.

Self-Referring Laws. H. L. A. Hart in *Festskrift Tillägnad Karl Olivecrona.* Kungliga Boktryckeriet, P. A. Norstedt & Soner, Stockholm, 1964.

On Self-Reference and a Puzzle in Constitutional Law. Alf Ross in *Mind,* Vol. 78, No. 309, pages 1–24; January, 1969.


## PHOSPHATE ROCK


Fertilizer Mineral Potential in Asia and the Pacific: Proceedings of the Fertilizer Raw Material Resources Workshop. Edited by R. P. Sheldon and W. C. Burnett. East-West Resources Systems Institute, East-West Center, 1980.

Fertilizer Raw Material Resources, Needs and Commerce in Asia and the Pacific: Proceedings of the Fertilizer Flows Conference. Edited by R. P. Sheldon, Saleem Ahmed and Y. H. Yang. East-West Resources Systems Institute, 1980.

Paleoceanography. Thomas J. Schopf. Harvard University Press, 1980.

Ancient Marine Phosphorites. Richard P. Sheldon in *Annual Review of Earth and Planetary Sciences,* Vol. 9, pages 251–284; 1981.


## RADIO ASTRONOMY BY VERY-LONG-BASELINE INTERFEROMETRY


Intercontinental Radio Astronomy. K. I. Kellermann in *Scientific American,* Vol. 226, No. 2, pages 72–83; February, 1972.

Introduction to Very-Long-Baseline Interferometry. Marshall H. Cohen in *Proceedings of the IEEE,* Vol. 61, No. 3, pages 1192–1197; September, 1973.

Quasars and Active Nuclei of Galaxies: Proceedings of the Copenhagen Symposium on "Active Nuclei," June 27–July 2, 1977. Edited by O. Ulfbeck in *Physica Scripta,* Vol. 17, No. 3, pages 135–385; March, 1978.


## CALMODULIN


Calmodulin and Its Role in the Second-Messenger System. Jerry J. Wang and David Morton Waisman in *Current Topics in Cellular Regulation,* Vol. 15, pages 47–107; 1979.

Calcium and Cell Function, Vol. 1: Calmodulin. Edited by Wai Yiu Cheung. Academic Press, 1980.

Calmodulin. C. B. Klee, T. H. Crouch and P. G. Richman in *Annual Review of Biochemistry,* Vol. 49, pages 489–515; 1980.

Calmodulin and Cell Function. Edited by D. Martin Watterson and Frank F. Vincenzi in *Annals of the New York Academy of Sciences,* Vol. 356, pages 1–443; 1980.

Calmodulin Plays a Pivotal Role in Cellular Regulation. Wai Yiu Cheung in *Science,* Vol. 207, No. 4426, pages 19–27; January 4, 1980.

Discovery and Recognition of Calmodulin: A Personal Account. Wai Yiu Cheung in *Journal of Cyclic Nucleotide Research,* Vol. 7, No. 2, pages 71–84; 1981.


## THE QUANTUM MECHANICS OF MATERIALS


Pseudopotentials in the Theory of Metals. Walter A. Harrison. W. A. Benjamin, Inc., 1966.

The Pseudopotential Concept. Volker Heine in *Solid State Physics,* Vol. 24, pages 1–36; 1970.

The Chemical Bond and Solid State Physics. James C. Phillips in *Physics Today,* Vol. 23, No. 2, pages 23–30; February, 1970.

Bonds and Bands in Semiconductors. James C. Phillips. Academic Press, 1973.

Introduction to Solid State Physics. Charles Kittel. John Wiley & Sons, Inc., 1976.

The Pseudopotential Panacea. Marvin L. Cohen in *Physics Today,* Vol. 32, No. 7, pages 40–47; July, 1979.


## THE ARCHAEOLOGY OF LASCAUX CAVE


Lascaux. H. Breuil in *Bulletin de la Société préhistorique française,* Vol. 47, pages 355–363; 1950.

Lascaux: Paintings and Engravings. Annette Laming. Penguin Books, 1959.

Débris de corde paléolithique à la grotte de Lascaux (Dordogne). A. Glory in *Mémoires de la Société préhistorique française,* Vol. 5, pages 135–169; 1959.

Lascaux inconnu: XII^e^ supplément à "Gallia Préhistoire." Arl. Leroi-Gourhan, J. Allain, L. Balout, C. Bassier, R. Bouchez, J. Bouchud, C. Couraud, B. and D. Delluc, J. Evin, M. Girard, A. Laming-Emperaire, A. Leroi-Gourhan, M. Sarradet, F. Schweingruber, Y. Taborin, D. Vialou and J. Vouvé. Éditions du Centre National de la Recherche Scientifique, 1979.


## THE STRUCTURE AND FUNCTION OF FISH SCHOOLS


Why Do Fish School? D. H. Cushing and F. R. Harden Jones in *Nature,* Vol. 218, No. 5145, pages 918–920; June 8, 1968.

Schooling in Fishes: Critique and Review. Evelyn Shaw in *Development and Evolution of Behavior: Essays in Memory of T. C. Schneirla,* edited by Lester R. Aronson, Ethel Tobach, Daniel S. Lehrman and Jay Rosenblatt. W. H. Freeman and Company, 1970.

The Sensory Basis of Fish Schools. B. L. Partridge and T. J. Pitcher in *Journal of Comparative Physiology,* Vol. 135, pages 315–325; 1980.


## BRAIN MECHANISMS OF VISUAL ATTENTION


Behavioral Modulation of Visual Responses in the Monkey: Stimulus Selection for Attention and Movement. Robert H. Wurtz, Michael E. Goldberg and David Lee Robinson in *Progress in Psychobiology and Physiological Psychology,* Vol. 9, pages 43–83; 1980.

Behavioral Enhancement of Visual Responses in Monkey Cerebral Cortex, I: Modulation in Posterior Parietal Cortex Related to Selective Visual Attention. M. Catherine Bushnell, Michael E. Goldberg and David Lee Robinson in *Journal of Neurophysiology,* Vol. 46, No. 4, pages 755–772; October, 1981.


## THE LUNAR SOCIETY OF BIRMINGHAM


The Life of Mary Anne Schimmelpenninck. Edited by Christiana C. Hankin. Longman, Green and Robert, London, 1860.

Lives of Boulton and Watt. Samuel Smiles. J. B. Lippincott and Company, 1865.

The Lunar Society of Birmingham. Robert E. Schofield. Oxford University Press, 1963.


## THE AMATEUR SCIENTIST


An Analysis of Turns in Dance. K. L. Laws in *Dance Research Journal,* No. 11–12, page 16; 1978–1979.

Physics and Ballet: A New Pas de Deux. Kenneth Laws in *New Directions in Dance,* edited by Diana Theodores Taplin. Pergamon Press, 1979.

Precarious Aurora—an Example of Physics in Partnering. Kenneth Laws in *Kinesiology for Dance,* No. 12; August, 1980.

# Offshore Arctic oil Until Exxon learned

## Dr. Anton Prodanovic is designing Arctic drilling islands that open a new world of energy.

Conventional offshore oil-drilling techniques simply won't work in Alaska's Beaufort Sea. Forces generated by Arctic ice could endanger the most rugged platforms now being built. But, America needs the Beaufort's oil. And Dr. Anton Prodanovic, Senior Research Specialist of Exxon Production Research Company, is helping us reach it. Dr. Prodanovic and a team of Exxon civil engineers are designing man-made islands which allow us to drill safely in the ice-covered waters of the Beaufort.

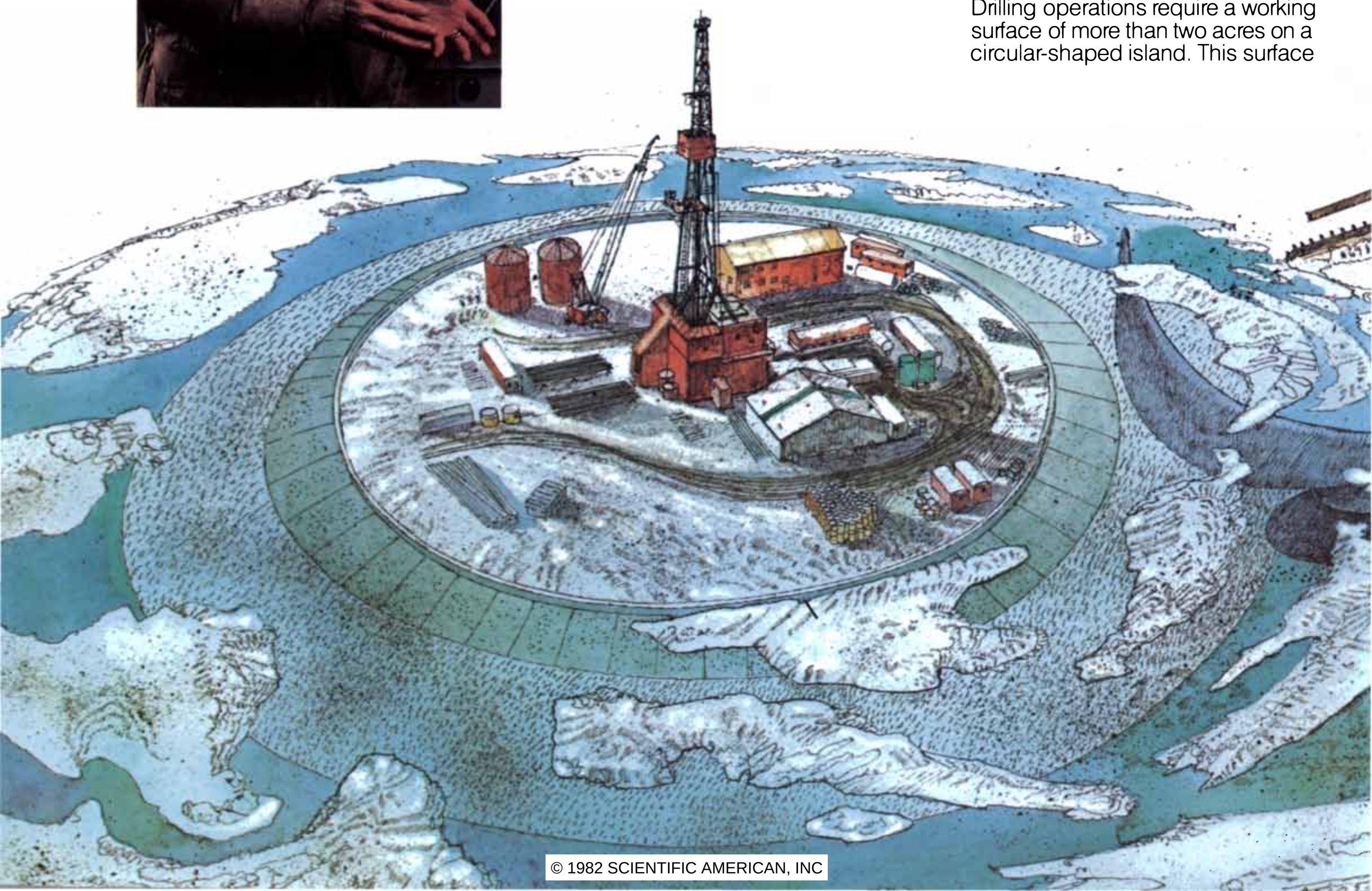
### Developing an I.Q. for ice

Key to determining the design of these islands (establishing such criteria as slope angles, freeboard elevations and berm heights) is getting to know the ice that will surround them. Since 1968 Exxon has pioneered in learning how Arctic ice grows, moves and behaves, and the forces it can exert.

Much of Exxon's Arctic ice research actually happens at −25°F in, of all places, Houston, Texas. Here, Dr. Prodanovic and other "island" engineers work closely with Exxon "ice" scientists. By studying, analyzing and testing core samples of ice flown down from the Beaufort Sea, a specific island design can be planned, taking into account a multitude of sea ice factors: sea ice structure (is it granular, columnar, or leached?); sea ice strength (compressive, tensile, shear or flexural); sea ice loads, pressure and movement in winter and summer; and whether the ice is multiyear or annual.

### Island design and construction

Once the ice environment, oceanographic and geotechnical aspects of a particular site have been studied, Dr. Prodanovic and his team can begin the design for a new island. Drilling operations require a working surface of more than two acres on a circular-shaped island. This surface



# was unreachable. how to break the ice.

also provides sufficient space for a crest-raising berm and a buffer zone behind the berm to accommodate potential ice-override pileups.

Island construction can be accomplished during the winter. Trucks carry gravel to the island site over an 8-foot-thick ice road constructed on top of the sea ice. The fill must be carefully placed and gradually built up from the sea floor to a safe freeboard elevation. The slope angle of gravel island beaches is 1:3, shallow enough to minimize fill erosion and give the island a design resistance greater than potential ice forces.

## Island safety and defense

For a few brief weeks each summer, the ice along the Beaufort coastline melts. To prevent slope erosion and protect the island from wave attack, a layer of hydraulic filter cloth is used to stop fill washout and the buildup of hydrostatic pressure in the underlying fill. On top of this, a cover layer of sandbags holds the filter cloth in place and further minimizes erosion.

While the island's mass is its first defense against winter ice attack, a secondary defense is ice slotting or trenching. An earth-trenching machine is used to cut a circumferential slot approximately 8 inches wide in the ice sheet. This defensive slot weakens the ice so that it collapses before ice loads become too large.

In addition to these defensive measures, monitoring stations and pressure sensors placed near the island provide continuous surveillance of ice movement and behavior.

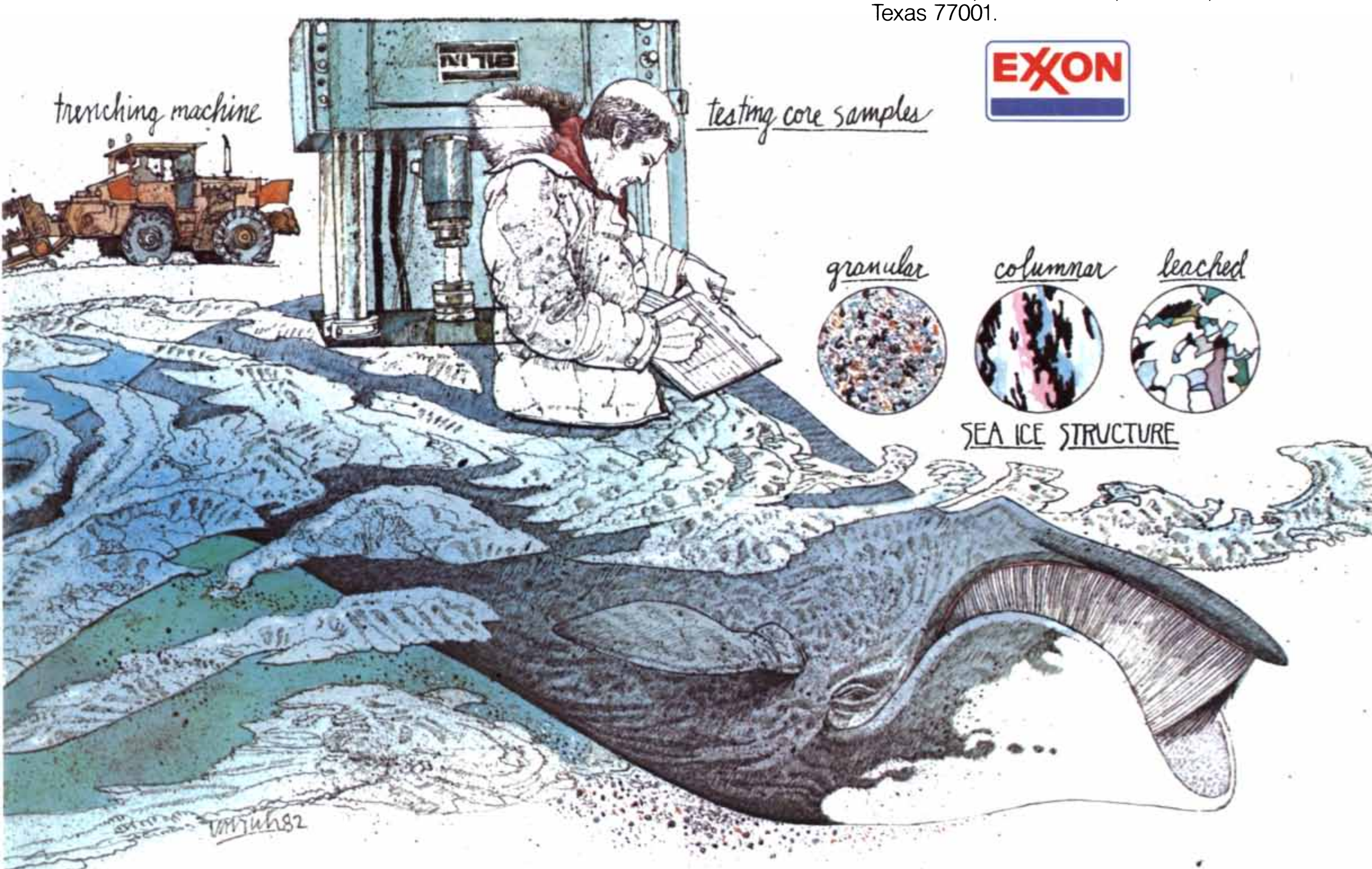
Over the past 4 winters, Exxon has built and drilled on 3 islands in the shallow, close-to-shore areas of the Alaskan Beaufort Sea. They have proven to be an efficient, effective way to tap energy that was once unreachable.

## Exxon Production Research Company

Offshore drilling systems and ice research are just two of the activities at EPRCo. A wholly owned subsidiary of Exxon Corporation, the company employs over 1000 scientists and engineers working in the fields of petroleum exploration and production, aimed at finding, producing and improving the recovery of worldwide energy sources.

If you would like more information about Dr. Prodanovic's work and EPRCo, write J. A. Rickard, Exxon Production Research Company, Room 605, P.O. Box 2189, Houston, Texas 77001.

EXXON




IMPORTED
Seagram's
V.O.
CANADIAN WHISKY
A BLEND
OF CANADA'S FINEST WH
THIS WHISKY IS SIX YE
86.8 PROOF
Taste!
Whether you mix it in a whisky sour, or pour it on the rocks, you'll see
why more V.O. is bought than any other imported distilled spirit in America.
More than any Scotch, Canadian, Rum, Gin, Vodka or Tequila.
VO
Enjoy our quality in moderation.
CANADIAN WHISKY. A BLEND OF CANADA'S FINEST WHISKIES
6 YEARS OLD. 86.8 PROOF. SEAGRAM DISTILLERS CO., N.Y.C.