

SCIENTIFIC AMERICAN

JANUARY 1989
\$2.95

New light on feeling SAD, carbohydrate craving and PMS.

Deep earthquakes: shocks from the planet's mantle.

Powerful algorithms weave efficient networks.



Patterns of complexity that emerge in the mixing of fluids have begun to yield insight into nonlinear physics.

That's a quote being heard more and more from new owners of the Ford Taurus. What is it about Taurus that made these former loyal import owners switch?

Design, advanced engineering, driver-oriented ergonomics and attention to detail all contribute to the answer.

In fact, attributes like these led *Road & Track* to name Taurus with 3.0L V-6 the best value in the world in its category...over cars like Honda Accord, Nissan Maxima and

Mazda 626. They've also led *Car and Driver* to name Taurus to its list of "10 Best Cars in the World" for three straight years. (Both distinctions given among cars sold in the U.S.) With accolades like these, it's little wonder Taurus is bringing import buyers back to the American sedan.

Transferable 6-Year/60,000-Mile Powertrain Warranty.

Covers you and future owners, with no

"I used to own an import."



Buckle up—Together we can save lives.

transfer cost, on major powertrain components for 6 years/60,000 miles. Restrictions and deductible apply.

Also, participating Ford Dealers stand behind their customer-paid work with the Lifetime Service Guarantee. If a covered repair must be fixed again, the repairing dealer will fix it free for as long as you own your vehicle.

Ask to see these limited warranties at your participating Ford Dealer.

Ford. Best-built American cars...eight years running.

Based on an average of owner-reported problems in a series of surveys of '81-'88 models designed and built in North America. At Ford "Quality is Job 1."

Ford Taurus



Have you driven a Ford...lately?



34

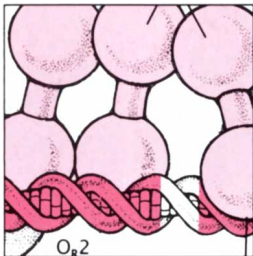


On Science Advice to the President

Jerome B. Wiesner

What worked for presidents Eisenhower and Kennedy will work for President-elect Bush. The author advocates reviving the President's Science Advisory Committee (which he headed under Kennedy) to give the new president the scientific community's best thinking on issues such as industrial policy, defense and the environment.

40

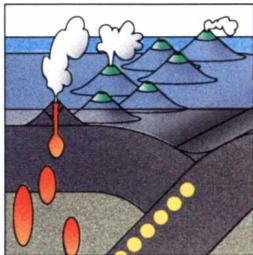


How Gene Activators Work

Mark Ptashne

A cell does not need all its genes all the time; it switches them on and off. Regulatory proteins bind to specific regions of DNA and start or stop the production of protein from nearby or distant genes. By studying gene regulation in viruses and yeast, the author and his colleagues have uncovered principles that also apply in more advanced organisms.

48

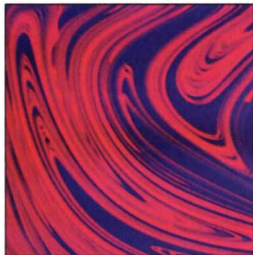


Deep Earthquakes

Cliff Frohlich

Earthquakes are often recorded at depths as great as 650 kilometers or more. These deep events mark regions where plates of the earth's surface are consumed in the mantle. But the earthquakes themselves present a conundrum: the high pressures at such depths should keep rock from fracturing suddenly and generating a tremor. What gives in deep earthquakes?

56



The Mixing of Fluids

Julio M. Ottino

Marbled endpapers, the eruption of a volcano and the making of puff pastry are all instances of mixing by stretching and folding. Laboratory mixing experiments offer insights into the stunning complexities of everyday mixing and serve as models of chaotic behavior: the time-varying flows wreak disorder that nonetheless has a certain symmetry.

68

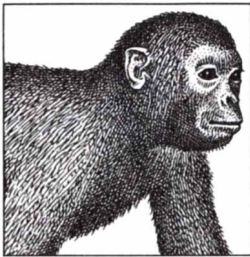


Carbohydrates and Depression

Richard J. Wurtman and Judith J. Wurtman

As the year wears on and the nights lengthen, many people become listless and depressed; they seek consolation in carbohydrate-rich snacks. Bright light can help, as can drugs that enhance the action of serotonin, a neurotransmitter. SAD (seasonal affective disorder) has some features in common with premenstrual syndrome and a form of obesity.

76

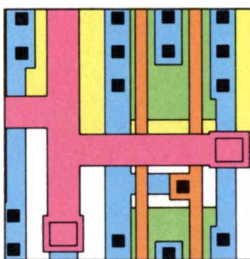


The Hunt for *Proconsul*

Alan Walker and Mark Teaford

Crushed, confused with pig bones and scattered among collections around the world, the remains of an apelike creature that lived in East Africa some 18 million years ago have now been reassembled. The slow-moving, tree-living hominoid that has taken shape probably represents the last common ancestor of human beings and the great apes.

84

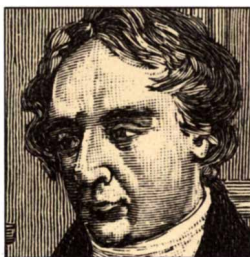


The Shortest-Network Problem

Marshall W. Bern and Ronald L. Graham

The problem is easy to state: Draw the shortest possible network of line segments interconnecting, say, 100 points. It is also unsolvable, in many cases. But its sheer fascination together with its practical importance to designers of telephone networks, for example, have led to the development of algorithms that yield rough solutions quickly.

90



André-Marie Ampère

L. Pearce Williams

In the early 19th century Ampère made pioneering contributions to the philosophy of science and the study of electricity and magnetism. His conviction that theoretical entities—themselves unobservable—could nonetheless be studied through their interactions guided his investigation of the magnetic effects of electric currents.

DEPARTMENTS

7 Letters

102 The Amateur Scientist

10



50 and 100 Years Ago

1889: A new narrow-gauge railway links mining centers in Colorado.

106



Computer Recreations

Can you tell what is on your head by looking at your companions?

12 Science and the Citizen

110 Books

98 Science and Business

114 Essay: *John Shattuck and Muriel M. Spence*



How To Get From The Zone

Dillen Pinnacles, Colorado, photographed on 4x5-inch KODAK T-MAX 100 Professional Film. ©Thomas L. McCartney, 1988
The moon, photographed with a 16-inch telescope on a KODAK T-MAX 100 Professional Plate. ©Richard Albrecht, 1988

©Eastman Kodak Company, 1989

© 1988 SCIENTIFIC AMERICAN, INC



The same advanced Kodak emulsions that are redefining photographic art are now poised to do the same for scientific photography.

New KODAK T-MAX 100 Professional Plates and KODAK Technical Pan Plates offer scientific photographers the same characteristics as the black-and-white films chosen by world-renowned fine-art photographers. Plus the ultimate in dimensional stability.

T-MAX Plates have very fine grain, high resolution, and conventional panchromatic sensitivity. Technical Pan Plates have extremely fine grain, extremely high resolution, and extended-red panchromatic sensitivity.

Between them, these plates comprise the widest available range of contrasts and applications. The T-MAX Plate excels



in the low- to medium-contrast range and the Tech Pan Plate, moderate to high contrast.

These new products complement the world's most comprehensive and widely available line of glass plates. They are stocked in all popular sizes and can be ordered in non-stocked sizes

System To The Solar System

and thicknesses to fit virtually any camera configuration.

If superior quality, consistency, and stability are important to you, you'll want to learn more about the newest KODAK Plates. For details, write the Kodak Information Center, Eastman Kodak

Company, 343 State Street, Rochester, NY 14650-0811. Or call toll free, 1 800 242-2424, Ext 12.





THE COVER photograph shows the complex structure that a small blob of fluorescent tracer can produce in a viscous fluid if the blob is placed in a region of chaotic mixing (see "The Mixing of Fluids," by Julio M. Ottino, page 56). The mixing flow was generated by a periodic, counterrotating motion of two eccentric cylinders. Although such a flow is chaotic, it has certain symmetries. Here an "island" of unmixed fluid is captured breaking up into two smaller islands as it passes by the axis of symmetry.

THE ILLUSTRATIONS

Cover photograph by Paul D. Swanson and Julio M. Ottino

Page	Source	Page	Source
35	UPI/Bettmann Newsphotos	66	K. R. Sreenivasan, Yale University
36-37	Joan Starwood	67	Ichiro Sugioka and Bradford Sturtevant, California Institute of Technology
38	Wide World Photos	69	Robert Mankoff, © 1984 by The New Yorker Magazine, Inc.
40	Janice A. Fischer, University of California, Berkeley	70	Andrew Christie
42-47	Neil O. Hardy	71	Laurie Burnham
49	Neculai Mandrescu, Institute for Physics of the Earth, Bucharest	72	Andrew Christie
50-53	Joe Lertola	73	S. Varnedoe
54	Joe Lertola (<i>top left</i>), William A. Bassett (<i>top right</i>), Joe Lertola (<i>bottom</i>)	74-75	Andrew Christie
55	Joe Lertola	77	Alan Walker
57-61	C. W. Leong and Julio M. Ottino, University of Massachusetts at Amherst	78-82	Tom Prentiss
62	Bob Conrad	85	Quesada/Burke (<i>top</i>), Gabor Kiss (<i>bottom</i>)
63	Bob Conrad (<i>top</i>), C. W. Leong and Julio M. Ottino (<i>bottom</i>)	86-89	Gabor Kiss
64	Bob Conrad (<i>top</i>); John G. Franjione and Julio M. Ottino, University of Massachusetts at Amherst (<i>bottom</i>)	91	The Granger Collection
65	Paul D. Swanson and Julio M. Ottino, University of Massachusetts at Amherst	92	L. Pearce Williams
		93-95	Hank Iken
		102-104	Michael Goodman
		107-108	Andrew Christie
		109	Johnny Johnson

SCIENTIFIC AMERICAN

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Armand Schwab, Jr., *Managing Editor*; Timothy Appenzeller, *Associate Editor*; Timothy M. Beardsley; John M. Benditt; Laurie Burnham; Elizabeth Corcoran; Ari W. Epstein; Gregory R. Greenwell; John Horgan; June Kinoshita; Philip Morrison, *Book Editor*; Tony Rothman; Ricki L. Rusting; Russell Ruthen; Karen Wright

ART: Samuel L. Howard, *Art Director*; Murray Greenfield, *Associate Art Director*; Edward Bell, *Assistant Art Director*; Johnny Johnson

COPY: Sally Porter Jenks, *Copy Chief*; M. Knight; Michele Moise; Dorothy R. Patterson

PRODUCTION: Richard Sasso, *Vice-President Production and Distribution*; Managers: Carol Eisler, *Manufacturing and Distribution*; Carol Hansen, *Electronic Composition*; Leo J. Petrucci, *Manufacturing and Makeup*; Carol Albert; Madelyn Keyes; Nancy Mongelli; William Sherman; Julio E. Xavier

CIRCULATION: Bob Bruno, *Circulation Director*; Lorraine Terlecki, *Business Manager*

ADVERTISING OFFICES: NEW YORK: Scientific American, 415 Madison Avenue, New York, NY 10017; Robert F. Gregory, *Advertising Manager*; Kate Dobson, *Advertising Manager*; Lisa Carden; Jack Grant. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager*; Litt Clark, *Midwest Manager*. DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; William F. Moore, *Advertising Manager*; Edward A. Bartley, *Detroit Manager*. CANADA: Fenn Company, Inc. DALLAS: Griffith Group. PRINCETON: William Lieberman, Inc. WEST COAST: Frank LoVerme & Associates

ADVERTISING SERVICES: Laura Salant, *Sales Services Director*; Diane Greenberg, *Promotion Manager*; Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: FRANKFURT, GENEVA, LONDON, PARIS: Infopac. HONG KONG/SOUTHEAST ASIA: C. Cheney & Associates. SEOUL: Biscom, Inc. SINGAPORE: Cheney Tan Associates. TOKYO: Nikkei International, Ltd.

ASSOCIATE PUBLISHER/INTERNATIONAL: Peter B. Kennedy

ASSOCIATE PUBLISHER/BUSINESS MANAGER: John J. Moeling, Jr.

PRESIDENT OF MAGAZINE DIVISION AND PUBLISHER: Harry Myers

SCIENTIFIC AMERICAN, INC.

415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

EXECUTIVE COMMITTEE: Claus-G. Firchow; *Executive Vice-President and Chief Financial Officer*, R. Vincent Barger; *Senior Vice-President*, Harry Myers; *Vice-Presidents:* Linda Chaput, Jonathan Piel, Carol Snow

CHAIRMAN OF THE BOARD: Georg-Dieter von Holtzbrinck

CHAIRMAN EMERITUS: Gerard Piel

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1988 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 rates: one year \$24, two years \$45, three years \$60 (outside U.S. and possessions add \$11 per year for postage). Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51593.

LETTERS

To the Editors:

I thoroughly enjoyed Volker A. Mohnen's graphic description of the acid-rain problem ["The Challenge of Acid Rain," *SCIENTIFIC AMERICAN*, August, 1988], and I consider it to be the best general summary of the problem to date. There are, however, a few conclusions about the current status of acidified lakes in his final paragraphs that are not totally accurate, as well as recommendations regarding control of the problem with which the majority of ecologists would disagree.

There is no evidence that the acidification of lakes in the U.S. Northeast has slowed in recent years. Paleocological analysis of sediments from lakes in the Adirondacks by Donald F. Charles of the University of Oregon and his colleagues in the PIRLA project indicate that acidification of many lakes in the Adirondacks has been rapid since the 1930's and 1940's and is continuing. There are two probable reasons. For one thing, scientists at the Hubbard Brook Experimental Forest, which has a 30-year record of rainfall chemistry, have shown that while the deposition of sulfuric acid has decreased in the past two decades, deposition of nitric acid has increased, so that the total acidity of rainfall has decreased very little in the eastern U.S. Moreover, as Peter J. Dillon and his colleagues at the Ontario Ministry of the Environment have found in eastern Ontario, decades of acid deposition have impoverished the buffering capacity of sensitive watersheds. Hence acidification in very sensitive lakes can continue even if the acidity of the deposition is decreased. Geological and chemical similarities suggest that similar conditions may occur in some watersheds of the northeastern U.S. Continued acid precipitation would be expected to exhaust the buffering ability of such watersheds.

On the other hand, there is evidence from studies of lakes in Canada, Norway and Sweden that decreasing the acidity of precipitation now will allow the majority of lakes and streams to recover quickly. In brief, forgoing short-term controls of the emissions at fault in favor of long-term, less expensive measures does entail further risks to the environment.

As Mohnen's diagrams show, only a few thousand lakes in the eastern U.S. are sensitive to acidification, and only a fraction of those have been damaged. Although it is conceivable some

urban-oriented Americans would be prepared to sacrifice these lakes in pursuit of the monetary "benefits" of continuing industrial pollution, recent surveys indicate that most Americans favor preservation of the environment, even if the cost is high. Furthermore, the U.S. is responsible for 50 percent of the acid deposition in eastern Canada, where 350,000 of the 700,000 lakes are acid-sensitive.

At present Canada is doing a much better job of reducing the acidity of deposition than the U.S., without being preoccupied with finding the easy or cheap way out. Sulfur dioxide emissions in eastern Canada have been reduced by 45 percent since the early 1970's. Nitrogen oxide emissions have been stable. These substantial cuts have made it possible for many Canadian lakes to recover from acidification. Further reductions in the acidity of deposition and recovery of Canadian lakes will require U.S. controls.

Like Canada, Norway and Sweden have also reduced emissions of sulfur dioxide. These countries all realize that not reducing emissions itself carries a significant economic cost, which we would pass on to our beneficiaries. Economic and ecological considerations are truly given different priorities by different governments. The priorities of the U.S. should be a matter of great concern to Americans.

D. W. SCHINDLER

Freshwater Institute
Canadian Department of Fisheries
and Oceans
Winnipeg, Manitoba

To the Editors:

Scientists from the Ontario Ministry of the Environment recently published in the reviewed literature the results of their analysis of the acid-rain trend. They concluded: "There has been a significant decrease in sulfate and strong acid concentrations in bulk deposition collected in central Ontario over the past 10 years. Over the same period, nitrate deposition has not changed significantly. During this time there have been decreases in sulfur dioxide emissions in Ontario (36 percent), eastern Canada (30 percent) and eastern U.S. (17 percent). The sulfate concentration in precipitation over the 10-year period was strongly correlated with changes in regional sulfur dioxide emissions." Atmospheric scientists in the U.S. have also established that the ambient burden of sulfur compounds has fallen during

the same period in the northeastern U.S., and that—as was observed in Canada—the nitrate concentration in precipitation over the northeastern U.S. has remained essentially unchanged. Furthermore, surveys have shown that the sulfate levels in streams and lakes in the northeastern U.S. have declined significantly.

Based on this evidence I stated "that some breathing space remains" and that the nation can probably forgo the short-term solution of retrofitting existing plants with pollution controls for sulfur dioxide only in favor of the gradual but more comprehensive and economical approach of repowering. Why do we need this breathing space? Acid rain is one of many environmental problems associated with fossil-fuel combustion. Climate warming and the formation of low-level ozone are equally serious problems that must be considered in a holistic approach to global housekeeping. Increasing significantly the efficiencies for power production and use is the only near-term solution to the upward trend in carbon dioxide emissions. Cutting emissions of oxides of nitrogen is essential for reducing the ambient ozone levels as well as the acidity of cloud water and precipitation.

Any effort to reduce sulfur dioxide emissions from the 410 or so power plants that were built before 1975 (and hence lack proper controls) must be accompanied by a strong commitment to minimize the emissions of oxides of nitrogen and carbon dioxide. The repowering concept, using clean-coal technologies, is a very attractive solution since it both increases efficiency and at the same time reduces emissions of oxides of sulfur and nitrogen. It is not a cheap solution, however. Whether this strategy is implemented by an act of Congress or on a market-driven basis is a question of public policy.

I fully respect Dr. Schindler's concern. We both agree on the need to minimize cumulative ecosystem damage, but our approaches to the problem of acid rain may differ.

VOLKER A. MOHNEN

ERRATUM

"Signing Off?" (*Science and Business*, October, 1988) incorrectly attributed the invention of television to RCA in 1939. Actually RCA was one of several contributors to television technology. RCA did spur the commercialization of television in the U.S. with a landmark broadcast from the World's Fair in April, 1939.



This is not your fa

Cutlass Supreme

We think your Dad would readily agree. We also think he might be right behind you when you go to the showroom. Because this new

Cutlass Supreme is the most researched, refined and remarkable car in our history.

Who wouldn't love it? The power comes from a 2.8-liter V6 with multiport fuel injection.

As for handling, it's remarkably precise. Each wheel boasts a separate suspension system, with MacPherson struts up front and coil springs in the rear.

And remember, behind it all



ther's Oldsmobile.

is GM's new 3-year/50,000 mile Bumper-to-Bumper Plus warranty. See your dealer for terms of this limited warranty.

On this car, everything looks good for a highly technical reason.

A whole new generation of Cutlass watchers will be snapping heads when this beauty whizzes by.

Want in on the fun? Visit your Oldsmobile® dealer for a test drive. Or for more information, call toll-

free 1-800-242-OLDS, Mon.-Fri., 9 a.m. to 7 p.m. EST.



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1939: "Television, problem-child of the laboratory, appears to be ready to turn the long-promised corner. Announcement has been made that television (now called 'video') programs will be presented on regular schedule by two or more eastern stations, beginning next spring."

"Radium is a gleaming sword in the treatment of disease, but it is a two-edged sword that has an unfortunate habit of getting lost. If the tiniest particle disappears, it not only is costly to replace (radium is worth 24,000 times its weight in pure gold) but also becomes a menace to the lives of those who may unwittingly come in contact with its destructive rays. The potential danger lurking in a few milligrams of lost radium is so great that scientists turn to ingenious devices to recover it: the gold-leaf electroscope and the Geiger-Müller counter."

"North Carolina is the first state in the nation to have a birth-control program sponsored by the state health department. Nearly half of the coun-

ties of the state now have birth-control clinics. There has been no local opposition to the service or the method adopted for rendering it. Social, religious and other civic leaders have given their full endorsement and cooperation. The patients have been selected from poor married women who need to limit the size of their families or space their children for the sake of their own and the children's health."

"Drinking water containing fluoride, a worry to residents of certain localities because it causes a permanent discoloration of the teeth of children, known as mottled teeth, may actually be a blessing in disguise. Children drinking this water are relatively immune to tooth decay, it is revealed by a four-year study described in *Public Health Reports* by Dr. H. Trendley Dean, dental surgeon of the United States Public Health Service."

SCIENTIFIC AMERICAN

JANUARY, 1889: "Columbia College, New York, has decided to have a special course in electrical science, and not a moment too soon, for this has long been seen to be a department by itself, and, while allied to other branches of natural philosophy, requires, at least from those who would adopt it as a profession, an undivided attention. In a practical age like this, the most valuable college instruction would seem to be the one that best

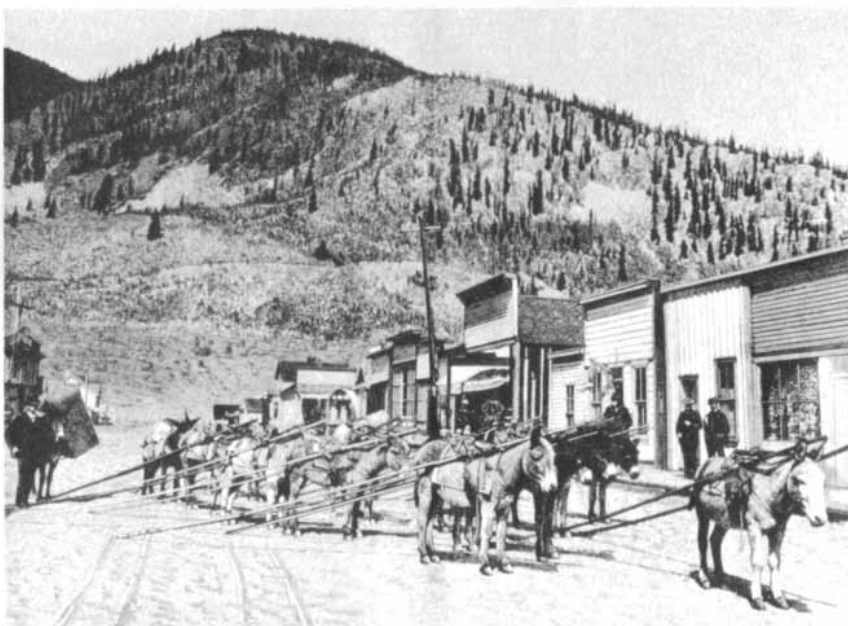
resembles what its recipients are expected to accomplish outside of it."

"Among the dangers peculiar to life in New York are the injuries to person and property resulting from the carelessness of employes connected with the elevated street railways, of which some forty miles are now in operation within the city. The railway people think nothing of piling up the coal on their locomotives in such a way that more or less of it rolls down into the street twenty feet below, to the danger of the crowds of people; while showers of hot water, oil, and live coals are not uncommon."

"The Lowth telephone is a new and in some respects remarkable instrument, by which speech is transmitted, without making use of sound waves as in the Bell and other forms of electrical telephones. In the Lowth telephone the transmission is effected by means of an electrical plug which is placed against the neck of the operator, near the vocal organs. The vibrations of the neck produced by the act of speaking shake the plug, thereby giving rise to corresponding electrical undulations, which pass over the wire. A valuable feature of this instrument is that the operator may be surrounded by all manner of loud noises and only his voice will be transmitted, and then he may speak almost in a whisper."

"Subscribers to whom are rented phonographs can have left at their door every morning the waxy tablets known as phonograms, which can be wrapped about a cylinder and used in the phonograph. On these tablets will be impressed from the clear voice of a good talker a condensation of the best news of the day, which subscribers can have talked back at them as they sit at their breakfast tables."

"If reference be made to a map of Colorado, it will be seen that the town of Ouray is the terminus of one branch of the Denver and Rio Grande Railroad. It will also be observed that Silverton, which is separated from Ouray by high mountains and deep valleys, is also the terminus of another branch of the Denver and Rio Grande. Work on a railroad to connect these two towns is now being prosecuted. The method of transporting the rails is illustrated in the engraving. The rails are strapped to the backs of donkeys, the ends dragging on the ground behind. The donkeys are not provided with either bridles or bits, but follow one another up the mountains single file."



Donkeys in the streets of Silverton, laden with rails for the new railroad

Revolutionary computer architectures have the potential to achieve massive parallel processing capabilities beyond that of the fastest conventional supercomputers. Under development by Hughes Aircraft Company for the U.S. Army Strategic Defense Command, these new architectures are designed to mimic the brain's vastly complex neurobiological structure. Using this technology, a new generation of computers may provide the solution for real-time processing problems like automatic target recognition, weapons allocation, automatic speaker identification and multi-sensor data fusion.

Voice and data communication to and from vehicles virtually anywhere in North America will soon be possible through a satellite system under development by Hughes and seven other companies that form the American Mobile Satellite Consortium. The system would allow drivers unrestricted contact with any telephone anywhere in the world. Current cellular telephone systems require drivers to be within range of special two-way radio towers, leaving about 15 percent of the United States population without service. Initial customers for the new system will be trucking companies, fire fighters, search and rescue teams, and personnel working in remote areas. The service will also be available to aviators and mariners.

A new superprojector provides large-screen display of computer data in full color. Designated Model 1000, the projector is designed for applications where real-time computer information must be viewed by large numbers of people. High-intensity xenon arc lamps, combined with the Hughes-developed liquid crystal light valve, generate a display with resolution in excess of 1,000 TV lines. Data is seen crisply and clearly, through front or rear projection, in normal room light. The superprojector is compatible with virtually all currently available computer sources and is derived from sophisticated color projection systems developed by Hughes for military command and control centers.

Using a special space-borne sensor's data, meteorologists are able to measure wind speed at the ocean's surface and the extent and thickness of ice for ship routing. The Hughes-built imaging sensor, designated SSMI, is also designed to measure soil moisture content, which may be used to predict potential flooding. Knowing the water content trapped in mountain snow packs could also lead to better water management and help control flood damage during ice thaws by permitting early precautions to be taken. SSMI, which uses microwave energy emitted by the earth and atmosphere to record this information, is flying aboard a U.S. Air Force weather satellite.

Engineers and scientists are eligible for approximately 100 Hughes Fellowships awarded for the pursuit of Master's and doctoral studies in Engineering and Science. All Fellows work full-time at Hughes during the summer, with Work-Study Fellows working part-time during the academic year and Full-Study Fellows attending classes full-time. Fellows receive full academic expenses plus stipends for studies at approved universities. Additionally, Hughes offers a two-year, entry-level rotation program that enables qualified BS and MS graduates to diversify their engineering experience. For more information contact the Hughes Corporate Fellowship Office, Dept. S2, C1/B168, P.O. Box 45066, Los Angeles, CA 90045-0066. U.S. citizenship may be required. Equal Opportunity Employer.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

The logo consists of the word "HUGHES" in a bold, white, sans-serif font, centered within a solid black rectangular background.

Subsidiary of GM Hughes Electronics

SCIENCE AND THE CITIZEN

New Deal?

The Bush Administration may be taking science seriously

Discreet lobbying by an inner circle of advisers from industry and academia apparently has persuaded President-elect George Bush to shake up the advisory apparatus of the White House in order to strengthen scientific advice and energize technology policy throughout the Federal bureaucracy.

Many leaders in science, technology and industry have been unhappy with what they have considered to be a weak Office of Science and Technology Policy (OSTP) under President Reagan. They have argued, for example, that better high-level scientific advice could have avoided some of the widely discounted claims made for Reagan's "Star Wars" program. There is also concern about the management of civilian technology: Robert M. White, president of the National Academy of Engineering, recently gave a speech complaining that "we have a science policy, but no technology policy." He further argued that "the Department of Defense has become the nation's de facto Ministry of Technology and Industry by default" and suggested that the department "is not where it should be." White sees a need to strengthen coordination by the OSTP as well as to give spending authority for technology management to an operating-level agency.

Although science and technology received barely a mention during most of the election campaign, in August a coalition of scientific and professional societies did urge both candidates to bolster science and technology advice in the White House. The subject got some attention in the Republican platform in the form of two pages of commitments to strengthening science, technology and the space program. That achievement is credited largely to Edward O. Vetter, a Texas management and political consultant whose association with Bush goes back to the Ford Administration. Bush's chief of staff, John H. Sununu, who has a background as an engineer, and Edward E. David, Jr., who was fired from the post of Science Adviser by President Nixon, also helped to ensure that technology found a place on the Republican agenda. James E. Carpen-



17 PHYSICAL SCIENCES 22 TECHNOLOGY 26 BIOLOGICAL SCIENCES

ter, a former senior OSTP analyst and now a Government consultant, chaired "issue groups" on science and innovation. They recommended that an effective science advisory council be constituted and that the early appointment of a Science Adviser to the President be given high priority in the new administration.

Bush cheered science-policy aficionados by speaking on the subject in Ohio in late October, in tones that generally reflected those influences. Bush pledged to continue strong support for basic research and to upgrade the office of the science adviser (held for the past two years by William R. Graham) to the rank of assistant to the president. Such a change, often suggested, would provide the adviser with direct access to the president and allow him to take part in Economic Policy Council and national-security deliberations. (Graham and his immediate predecessors reported through the White House staff.)

Bush also undertook to establish a President's Council of Science and Technology Advisers "composed of leading scientists, engineers and distinguished executives from the private sector." OSTP-watchers expect the new council to play a more active role than that played by the White House Science Council in recent years. The Bush campaign released a document expanding on the Ohio speech and promising to make permanent the research and development tax credit that Congress has been renewing periodically. That move had been vigorously championed by the Council on Research and Technology (CORETECH), a Washington lobbying group supported by industry and universities.

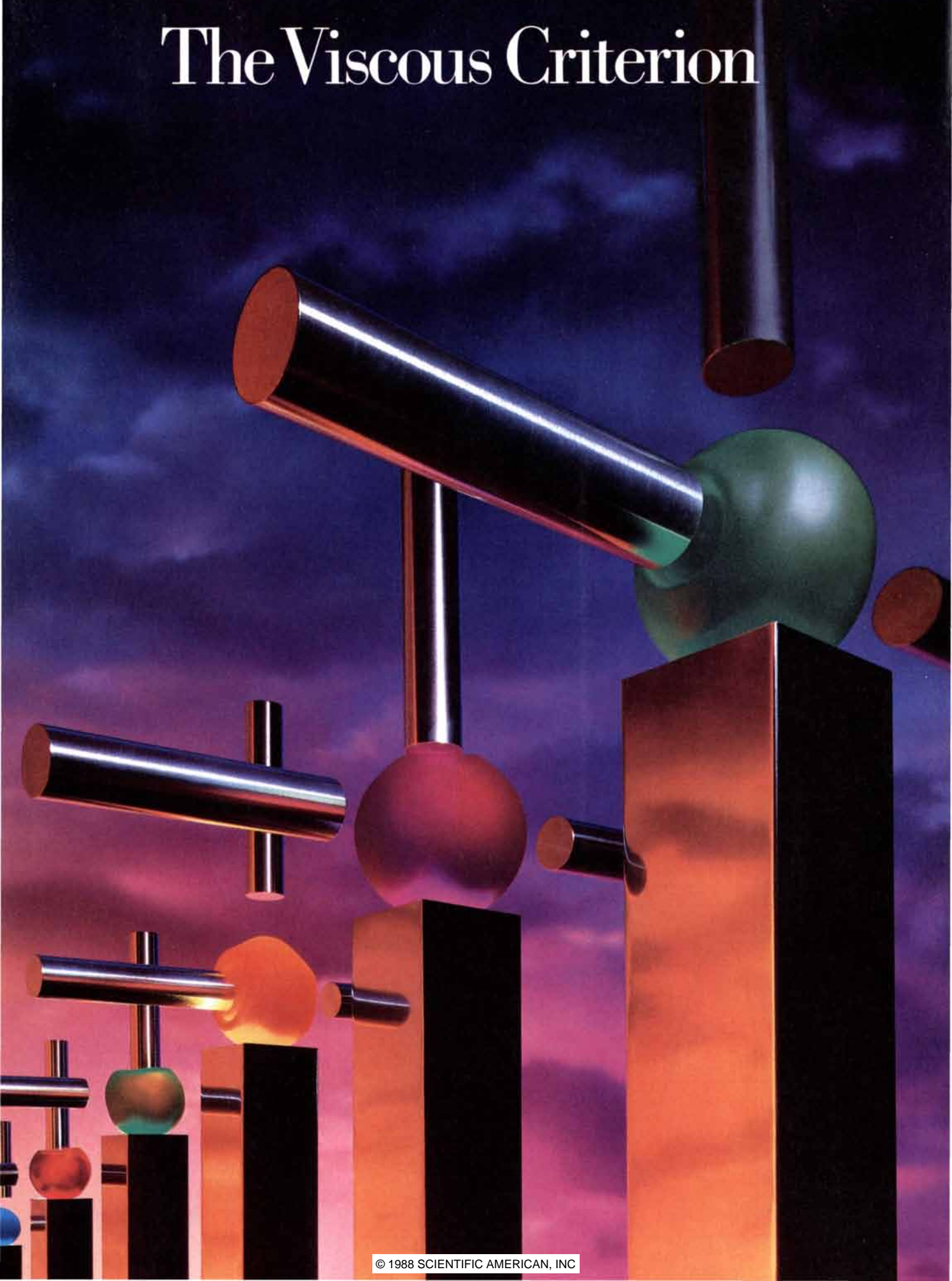
William O. Baker, a former chairman of the AT&T Bell Laboratories and an intelligence expert who has served on numerous Federal advisory committees, is considered one of the most

influential science-policy advisers to the Bush team: he believes science and technology planning "has to penetrate every operational aspect of the executive branch." The 1988 campaign was marked by little high-level effort to promote a Republican agenda for these issues, he says. In October, Graham made an attempt by forming the short-lived National Executive Council of Science and Technology Leaders for Bush/Quayle. The council included several prominent science-policy figures, including Baker and George A. Keyworth, Graham's predecessor in the Reagan Administration. Although the council never met, its members may have influence on the new administration; they have informally suggested to the Bush transition team candidates for a new Assistant to the President for Science and Technology. Solomon J. Buchsbaum of AT&T, the recent chairman of the White House Science Council, was rumored to be a leading candidate for the position in late November. D. Allan Bromley, a Yale University physicist who was a member of the White House Science Council under Reagan and cochaired an influential study of university research facilities, is also said to be held in high regard by Bush.

Further advice to the Bush team has come from the Commission on Science, Technology and Government established by the Carnegie Corporation of New York last April. The commission, a bipartisan body chaired by Joshua Lederberg, president of Rockefeller University, and William T. Golden, president of the New York Academy of Sciences, has sent a draft of its first report to the transition team. Golden says one of the commission's principal recommendations is to strengthen the OSTP; one suggestion is to fill four associate directorships there that have been left vacant in recent years. The known Democratic sympathies of some of the commission's members may, however, mean that the report will be received coolly.

Not to be outdone, the National Academy of Sciences and the Institute of Medicine are preparing, for the first time, a series of "white papers" for the transition team. The presidents of the two bodies and senior staff members have written the papers, which have not undergone the peer scrutiny usual for academy reports and are described as "editorials." The subjects covered

The Viscous Criterion



The Viscous Criterion

Two scientists at the General Motors Research Laboratories have developed a way to predict the probability and severity of impact injuries in the body's soft tissues, including the heart, liver, and the central nervous system. It is an essential step in designing safety systems to reduce such injuries.

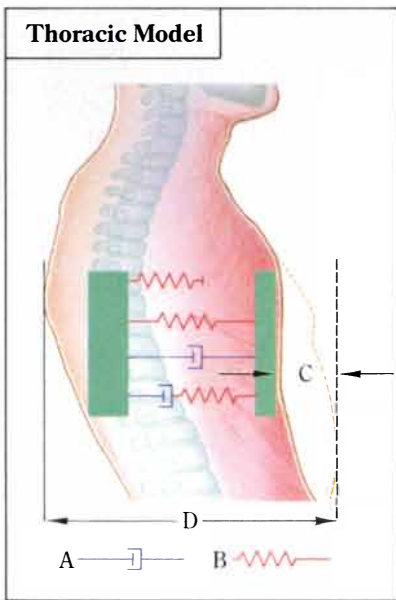


FIGURE 2: Lateral view of a viscoelastic model of the human chest. Dashpot elements (a) and spring elements (b) represent thoracic compliance. Compression (C) is expressed as a percentage of original chest depth (D).

FIGURE 1: A) Plot of viscous response, $VC(t)$, during impact (black). $C(t)$ =normalized compression (red). $V(t)$ =rate of chest deflection (blue). $[VC]_{max}$ defines the Viscous Criterion. B) Range of validity for Viscous Criterion (yellow).

Designing an automobile to reduce the risk of injury to its occupants in a collision demands an ability to correlate the forces generated by the crash with the biological effects experienced by the people involved.

Military rocket sled experiments in the late 1950s measured man's ability to withstand sudden changes in speed. The resulting Acceleration Criterion was used in setting 60g (60 times the force of gravity) as the maximum spinal acceleration allowable under federal motor vehicle standards in a 30 mph crash test.

This Acceleration Criterion treats the body as a rigid structure. Over the years, however, subsequent research on injury mechanisms indicates that injury criteria based on whole-body acceleration are incomplete predictors of injury risk.

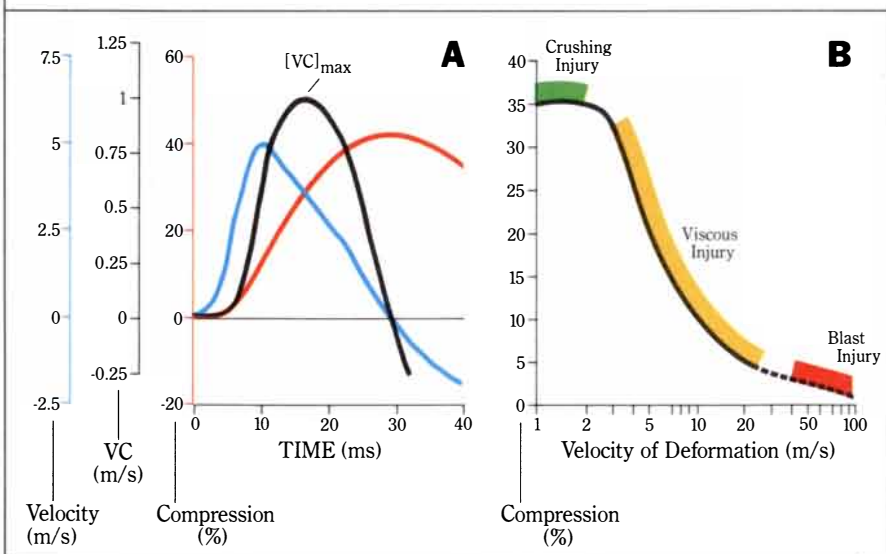
The body is a deformable structure, and injury can be sustained when the chest is compressed in an accident. At low speeds of deformation (less than 3 meters per second), the tolerance to rib cage

damage and the risk of injury correlate closely to the maximum compression of the chest—expressed as a percentage of the original chest depth from sternum to spine. This Compression Criterion, developed in the early '80s by GM researchers in conjunction with the University of California at San Diego, is useful in evaluating injury risk for safety-belted occupants, where tight coupling to vehicle deceleration reduces the amount of chest compression in a collision.

Doctors Ian Lau and David Viano—both members of the Biomedical Science Department of the General Motors Research Laboratories—began in 1981 to evaluate the importance of velocity in assessing the risk of impact injury. They were concerned that the maximum compression tolerance might underestimate chest injury risk at high speeds of chest deformation (greater than 3 m/s)—typical of unrestrained occupants in a frontal crash, or in high-speed side impacts.

The two scientists designed a series of experiments that held maximum compression constant at 16%—well below the tolerance level of 35%—and varied the rate of abdominal compression from 5 m/s to 20 m/s. The experiments verified that severity of soft tissue injury increased as the velocity of compression increased.

These results, plus further analysis of previous experiments, led Viano and Lau to develop a function called the *Viscous Response* to describe the behavior of soft tissue during an impact event. Viscous Response was defined as the instantaneous product of the velocity of deformation and compression, varying over time: $VC(t)$ (Figure 1A).



The mathematical form of the Viscous Response is derived from analysis of a mechanical analog of the viscoelastic response of the human thorax. (Figure 2). The dashpots (2a) represent the behavior of viscous soft tissue, while the springs (2b) correspond to the elastic skeletal response to impact.

In computing impact energy absorbed by the analog, the dominant term is the product of velocity of deformation and compression, with compression defined as chest deflection normalized by the original chest depth, (Fig. 2, D). Therefore, the Viscous Response is related to absorbed energy.

Drs. Viano and Lau suspected that the injury mechanism for soft tissue was also related to absorbed energy, and designed further experiments to verify the predictive abilities of the peak Viscous Response (VC_{max}). In these tests, velocities of deformation ranged from 5 m/s to 22 m/s and maximum chest compressions ranged from 4% to 55%. Analysis of the test data showed that the maximum Viscous Response was an accurate predictor of injury risk for the entire data set. In addition, VC_{max} was the only biomechanical response that adequately defined injury risk for the full range of test conditions, including the extremes of only 4% compression at 22 m/s, as well as 55% compression at only 5 m/s.

Investigation across this range of deformation velocities effectively links together existing knowledge of crushing injuries, high-speed impact injuries, and data available on blast injuries (Figure 1B).

Applying the Viscous Criterion to previously published blunt frontal impact data, Lau and Viano used

statistical analysis to show that VC_{max} was highly correlated with the risk of severe injury. "For velocities of chest deformation above 3.0 m/s," says Dave Viano, " VC_{max} is the principal indicator of injury, whereas for very slow speeds of deflection, the Compression Criterion assesses crushing injury risk. We are, therefore, recommending a viscous tolerance for the chest of VC_{max} equal to 1.00 m/s, and a compression tolerance of C_{max} equal to 35% to minimize the risk of severe injury in an accident."

Ian Lau points out the importance of such risk assessments as targets for automotive designers. "Based on our new awareness of the mechanism of soft tissue injury, General Motors has already designed a self-aligning steering wheel that can be an excellent countermeasure for reducing abdominal injuries in a crash."

The new wheel works in concert with the energy-absorbing steering column, and is available as standard equipment on the 1989 Chevrolet Cavalier. Says Lau, "This is an excellent example of engineering and medical science working together. And because GM is the only auto maker with a biomedical research facility and a dedicated staff of research professionals, it can only happen here."

General Motors



MARK OF EXCELLENCE

THE MEN BEHIND THE WORK:



David C. Viano and Ian V. Lau are both members of the Biomedical Science Department at the GM Research Laboratories.

Dr. Viano (right) is a Principal Research Scientist, leading the department's Safety Research Program. Dave received his BS in Electrical Engineering from the University of Santa Clara; he holds both an MS and a Ph.D. in Applied Mechanics from the California Institute of Technology. Dr. Viano joined GM in 1974 following postdoctoral work in Biomechanics at the Swiss Institute of Technology. His interests include technologies to improve occupant protection, the biomechanics of trauma and disability, transportation safety, and public health approaches to injury control.

Dr. Ian Lau came to the Research Laboratories in 1978, and is now a Senior Staff Research Engineer. Ian has a BS in Electrical Engineering from Lowell University. He holds a Ph.D. in Biomedical Engineering from the School of Medicine of the Johns Hopkins University. Ian was also a Postdoctoral Fellow of the American Heart Association at the Hopkins School of Hygiene and Public Health. His other research interests include traumatic cardiac arrhythmias, and occupant interaction with the steering and supplemental restraint systems.

are AIDS, global climate change, science advice to the president and space policy. The National Academy of Engineering is also submitting a paper on technology and competitiveness. Congress recently authorized a new post of Under Secretary for Technology in the Department of Commerce that, it is hoped, may improve technology-transfer programs administered by the department.

Science policy is hardly expected to be the first priority of a new administration. Nevertheless, the deluge of advice on the subject would seem to have some chance of being heard: according to his longtime associate Vetter, Bush has no need to be convinced of the growing importance of technology for the economy and is likely to be attentive. —Tim Beardsley

Back to the Bases

Watson slips into the driver's seat of the genome project

Last fall James D. Watson, codiscoverer of the structure of DNA in 1953 and director of the Cold Spring Harbor Laboratory since 1968, took on a new challenge. He is directing the contribution of the National

Institutes of Health to a \$3-billion international effort to analyze our genetic heritage: the complete sequence of bases in human DNA. Watson now spends two days a week at the NIH's campus in Bethesda, Md., as associate director for human-genome research.

The challenge is bureaucratic as well as technical and scientific. The new position gives Watson a minuscule staff and no formal spending power, and he has to dovetail the NIH's efforts with those of the Department of Energy, which is jockeying with the NIH for leadership of the U.S. effort. The multifaceted project requires creating new data bases and software, developing new laboratory techniques and, hardest of all, engineering unprecedented cooperation between turf-jealous Federal bureaucracies and investigators not known for their humility.

In addition he must contend with the controversial nature of the project itself. Critics argue that the billions of dollars needed to obtain the complete sequence of the three billion pairs of bases in the genome would be better spent targeted at specific research problems. Watson himself was chary when the Energy Department first initiated discussion of an organized genome project a few years ago, believing that the department was not the

proper venue for it. Two years ago, however, Congress awarded the NIH \$18 million it had not even requested to begin groundwork on the project. This demonstration that the genome project could win new money on its own and so not pauperize other research was enough to mollify doubters. This year the NIH has \$27.6 million earmarked specifically for genome research; the Energy Department is close behind with \$17.9 million, and France, Japan and the Soviet Union, among other countries, are also supporting genome studies.

Last October, on his first day in a bare Bethesda office, Watson spoke with *SCIENTIFIC AMERICAN* about the rationale for the project and its prospects. He answers the criticism that funds would be better spent on identified problems by arguing that the wealth of information gained from the project will make other research more efficient. And he points out that the effort as it is now envisioned will not merely start sequencing at one end of the genome and push on blindly; rather, it will attack on several fronts, producing maps based on genetic landmarks as well as physical maps based on identified fragments of DNA. Those maps will serve as guides to the most interesting stretches of DNA, which

The facts on deterministic fractals. By the expert in the field.

FRACTALS EVERYWHERE

by

Michael Barnsley

Michael Barnsley is inventing the future of deterministic fractal geometry. You have read about his exciting work in publications such as *IEEE Spectrum*, *Byte*, and *Scientific American*. Now you can read his new authoritative book on the subject of fractals, with revolutionary applications in:

- ▲ image compression
- ▲ satellite reconnaissance and imagery capabilities
- ▲ computer-aided design
- ▲ flight simulation
- ▲ cinematic special effects

Handsomely illustrated, including 32 full-color plates, this innovative work is your key to mastering fractal geometry.

1988, 424 pages, \$39.95/ISBN: 0-12-079062-9



ACADEMIC PRESS Harcourt Brace Jovanovich, Publishers
Book Marketing Department #27128, 1250 Sixth Avenue, San Diego, CA 92101

Credit Card Customers Call Toll Free 1-800-321-5068



will be sequenced first. A complete sequence of the genome—the ultimate goal of the project—is thought likely to take 15 years and will require advances in automation to keep the cost within reasonable bounds.

Watson thinks a high-resolution genetic map of the entire genome could be completed within five years, if not two or three; several research groups in the U.S. are already mapping regions of interest. Watson also believes early sequencing efforts will be crucial, because they will lead to economies of scale: "Until you have to sequence you won't develop methods to make it cheaper."

Watson—although "not a computer jock" himself—sees as his first priority establishing an overall structure for the computerized data bases that will store, organize and make accessible the new information. Maintaining the data bases will, he says, require training experts in biological data management. He is not worried about his lack of formal authority to authorize such expenditures, seeming confident that the National Institute of General Medical Sciences (the conduit for NIH genome funds) will support his spending recommendations. The central managerial problem will, he says, be to run the project in such a way that the "inherently boring" nature of sequencing large regions of DNA does not stultify investigators. Nor does the tussle between the NIH and the Energy Department for leadership of the project perturb Watson. The question, he says, has not yet been resolved, and it "will not be until it has to be." In the meantime the two agencies have signed a memorandum of understanding that establishes formal coordination by way of a joint advisory committee.

Charles Cantor was recently appointed director of one of the two Energy Department genome research centers, at the Lawrence Berkeley Laboratory (the other is at Los Alamos). Cantor will also be the first chairman of the department's genome steering committee, which makes him in effect Watson's counterpart. The Energy Department centers, unlike the research groups supported by the NIH, have staffs of dozens that include physicists and instrument builders; they will enable the department to develop new research tools even as research proceeds. "The NIH does not have in its employ many physicists and full-time computer scientists," Cantor points out. Yet he maintains a cooperative stance, saying the department will play a role complementary to that of the NIH. Cantor hopes to make recom-

mendations soon on computer software for the project.

International cooperation may be even harder to bring about than national coordination: there is likely to be competition to map regions of the genome associated with important genetic diseases, Watson says. Victor A. McKusick of the Johns Hopkins University School of Medicine is president of a new institution, the Human Genome Organization, that aims to coordinate the effort worldwide. The scale of the international effort will be close to \$200 million per year within three or four years, with the U.S. contributing more than half, according to Cantor. Has support from Congress been adequate up to this point? "I think we're lucky to have got what we've got," is Watson's agile reply. —T.M.B.

PHYSICAL SCIENCES

Lone-Star Science

The supercollider will be built in Texas—if it is built at all

Ten-gallon hats and yee-haws no doubt filled the skies over Texas after the U.S. Department of Energy announced that it intended to build the world's largest particle accelerator, the Superconducting Supercollider, on a site some 25 miles south of Dallas. But Texas' victory over six other states that had been vying for the prize may prove to be hollow.

The supercollider appeared to have widespread support two years ago, when the Reagan White House finally gave its official approval to the machine. It would be 53 miles around, big enough to encircle most of New York City, and would smash protons into each other with an energy totaling 40 trillion electron volts (TeV). That is more than 20 times the energy attained by what is now the world's most powerful accelerator, the Tevatron, at the Fermi National Accelerator Laboratory in Batavia, Ill.

Other machines more powerful than the Tevatron may be operating by the mid-1990's. The Soviet Union is constructing a 3-TeV proton collider, and the European laboratory for particle physics (CERN) in Switzerland is considering building an 18-TeV collider. These accelerators, however, may not be sufficiently powerful to discover the latest grail of particle physics: the Higgs boson. Finding this particle could yield insights into the origin of mass and perhaps lead to a single

Announcing publication of the new SCIENTIFIC AMERICAN CUMULATIVE INDEX 1978-1988



AN IMPORTANT REFERENCE GUIDE TO EVERY ARTICLE, AUTHOR & DEPARTMENT IN MORE THAN TEN CONSECUTIVE YEARS OF SCIENTIFIC AMERICAN

Covers the subject matter of more than 125 issues of Scientific American—and transforms them into a source book of contemporary science and technology.

A must for scientists, engineers, educators, researchers and all who need scientific and technological information.

Includes these indexes — • Key Word Index to Topics • Authors • Titles • Tables of Contents • Book Reviews • The Amateur Scientist • Computer Recreations.

PUBLISHING DATE — MARCH '89
LIST PRICE \$24.95
SPECIAL LIMITED-TIME PRE-PUBLICATION PRICE \$17.95
RESERVE YOUR COPY NOW!

SCIENTIFIC AMERICAN
415 Madison Avenue • New York, NY 10017

Reserve _____ copies of the new SCIENTIFIC AMERICAN CUMULATIVE INDEX 1978-1988 at the Special Pre-Publication price of \$17.95, and ship immediately upon publication in March, 1989. Add \$1.50 shipping and handling for each copy ordered.*

Name _____

Address _____ Apt _____

City _____ State _____ Zip _____

My check/money order is enclosed for \$ _____

Charge my VISA MasterCard

Access Eurocard

Card # _____ Exp. Date _____

Signature _____

*Add applicable sales tax for CA, ILL., MASS., MICH., OH., NY. OUTSIDE THE U.S. Remit \$17.95 in U.S. funds drawn on a U.S. bank (\$21.95 Canadian funds) or by credit card. Add \$2.50 shipping and handling per copy ordered.

01/89

The Superconducting Supercollider is aptly named: it dwarfs any other accelerator existing or planned



MANHATTAN ISLAND and the outlying boroughs of New York City would be largely encompassed by the proposed Superconducting Supercollider. The accelerator, which the Department of Energy actually hopes to build not in New York but in Waxahachie, Tex., about 25 miles south of Dallas, calls for a tunnel 53 miles in circumference. Some 10,000 superconducting magnets would accelerate protons nearly to the speed of light and smash them into each other with a violence rivaling that of the big bang.

theory unifying all the forces of nature. Proponents of the supercollider say it alone would be powerful enough to be a sure bet: it would either find the Higgs boson or prove that it does not exist, thereby pointing the way to new theories.

It is also a sure bet that the supercollider would yield economic benefits to Texas. The Energy Department proposes spending more than \$5 billion over the next eight years to build the accelerator; when it is completed, it would employ 2,500 scientists and technical staff, as well as 500 visiting scientists, and consume a \$270-million annual budget. Little wonder, then, that 25 states responded promptly to the Energy Department's request for bids in April, 1987. Later

that year the department named the seven finalists: Arizona, Colorado, Illinois, Michigan, North Carolina and Tennessee in addition to Texas.

Inevitably, as the number of potential beneficiaries has dwindled, political opposition has grown. The Energy Department had asked for a 1989 budget of \$363 million—about 10 times more than the previous year's budget—in order to begin construction. Last summer, after extensive hearings, Congress provided only \$100 million for continued design studies and ordered that construction not begin for at least another year. The delay will allow the new administration to reconsider the project.

President-elect Bush is expected to support the Energy Department's

eight-year budget, but even with his backing the supercollider may be further delayed or even killed by Congress. Some legislators have been swayed by the increasingly vocal opposition of such prominent physicists as Arno A. Penzias of the AT&T Bell Laboratories and Philip W. Anderson of Princeton University, who argue that the supercollider will hurt physics by draining funds and manpower from valuable smaller-scale research. Critics have even challenged particle physicists' claim to the throne of science, pointing out that no imaginable extrapolation of particle physics can explain the behavior of such large-scale systems as, say, bacteria.

Congress has other more worldly concerns. One involves the supercollider's powerful superconducting magnets. About 10,000 would be needed to accelerate and focus the protons as they hurtle around the oval-shaped tunnel. The Energy Department maintains the magnets can be manufactured for \$100,000 each. Early prototypes performed poorly, however, and although the problems have recently been overcome, some experts think the fixes may make the magnets more expensive to produce.

The threat of cost overruns could be fatal. Congress is determined to reduce the budget deficit, and the supercollider is competing for funds with a number of other large research projects, such as the Space Station and the National Aerospace Plane. These projects will arguably have a more tangible payoff; moreover, they are not based in a single state and so may engender wider political support. The Energy Department hopes to reduce the cost of the supercollider to the U.S. by soliciting contributions from other countries, such as Canada, Great Britain and Japan, but these countries are not likely to make a firm commitment until they see the U.S. do so.

The bitterness of the states that lost the bidding war could compound the Energy Department's troubles. Officials from Illinois (which had been widely considered to be the front-runner) and other losing states have demanded an investigation of the site-selection process. The Energy Department says Texas offered the best combination of geology and "regional resources" such as utilities, universities and industrial facilities; moreover, the state offered to put up \$1 billion of its own, more than any other contender. Some observers also think Texas was chosen because, alone of all the final contenders, it may have enough political clout to get the supercollider built.


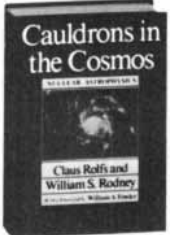
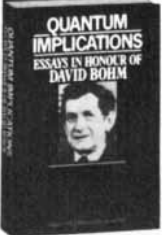



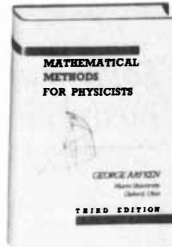
3 BOOKS FOR ONLY \$1 EACH

as your introduction to the **LIBRARY OF SCIENCE.** You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.


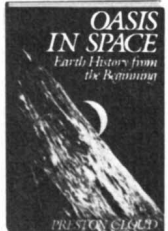


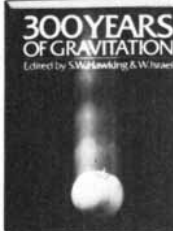

Values to \$87.40

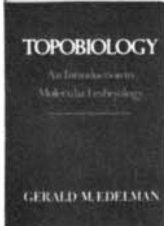
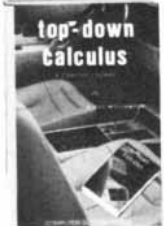


(Publishers' prices shown.)

						
63850 \$14.95	85621 \$18.95	87701 \$24.95	65657 \$24.95	76946 \$19.95	86670 \$17.95	48750 \$17.95

						
36186 \$24.95	37361-3 \$74.95 (Counts as 3 choices)	73177-2 \$49.95 (Counts as 2 choices)	55851 \$19.95	48834-2 \$49.95 (Counts as 2 choices)	36801 \$18.95	61310-2 \$49.95 (Counts as 2 choices)

						
79752-2 \$44.50 (Counts as 2 choices)	40040-2 \$39.95 (Counts as 2 choices)	63856 \$19.95	73283 \$29.95	53888 \$14.95	36508 \$19.95	32500 \$22.95

					
48840-2 \$39.95 (Counts as 2 choices)	65180 \$29.95	84319 \$17.95	56716 \$18.95	84261-3 \$69.60 (Counts as 3 choices)	52373 \$21.95

			
84410 \$21.95	84407 \$25.95	70350 \$17.95	60755 \$27.50

MEMBERSHIP BENEFITS • In addition to getting 3 books for only \$1.00 each when you join, you keep saving substantially on the books you buy. • Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of 65% off the publishers' prices. • At 3-4 week intervals (16 times per year), you will receive the Library of Science News, describing the coming Main Selection and Alternate Selections, together with a dated reply card. • If you want the Main Selection, do nothing, and it will be sent to you automatically. • If you prefer another selection, or no book at all, simply indicate your choice on the card and return it by the date specified. • You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card is missing, please write to The Library of Science, Dept. 2-CL4, Riverside, NJ 08075, for membership information and an application.

Scientific American 1/89

After all, James C. Wright, Speaker of the House of Representatives, Lloyd M. Bentsen, who in spite of his failed vice-presidential bid remains a powerful force in the U.S. Senate, and President-elect Bush all claim Texas as their home.

—John Horgan

Greenhouse America

A global warming may destroy U.S. forests and wetlands

In the next century people living on the coasts of North America will be desperately staving off rising seas with dikes and bulwarks. Many coastal marshes and estuaries will have vanished. Throughout the interior of the continent forests will be dying. Hundreds of species of plants and ani-

mals, traumatized by these changes, will face extinction.

This is the unsettling prognosis of the most comprehensive examination to date of how a "greenhouse" warming could affect the U.S. The Environmental Protection Agency spent almost two years and \$2.5 million on the study, which is still in draft form and is scheduled to be delivered to Congress soon.

The EPA began where a 1987 study by the National Academy of Sciences left off. The academy predicted that the burning of fossil fuels, slash-and-burn farming and other human activities could double atmospheric levels of greenhouse gases such as carbon dioxide and methane within 40 years. As a consequence, the academy concluded, the temperature of the earth's atmosphere could rise by from two to

five degrees Celsius; the level of the oceans, fed by melted snow and ice, could rise by as much as two meters. These dramatic predictions spurred the Senate Committee on Environment and Public Works to ask the EPA: What might happen to the U.S.?

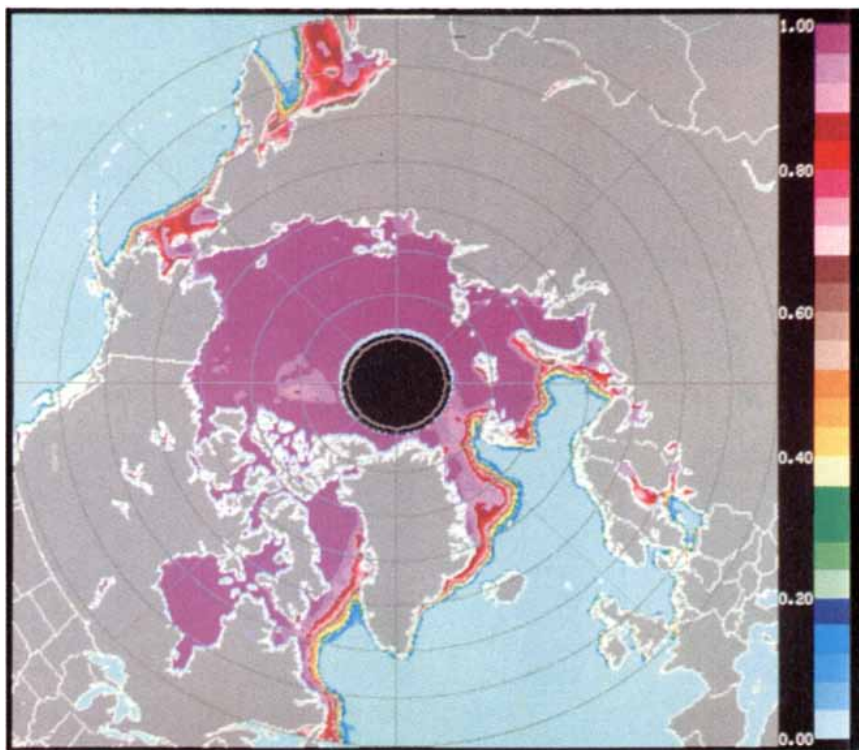
The EPA workers note that it is easier to predict what will happen to the entire world, which is essentially a closed system, than to make forecasts for specific regions. Nevertheless, by plugging the NAS predictions into climate models of the U.S., the EPA workers arrived at various scenarios for the future. Natural ecosystems will apparently suffer the most damage. Up to two-thirds of the nation's wetlands may be destroyed, either submerged or artificially protected from the encroachment of the sea, according to the EPA. Birds, fish and other animals that depend on access to wetlands— notably shrimp that breed along the Gulf Coast—may then perish.

Perhaps the most surprising finding, according to Joel B. Smith, a co-editor of the EPA report, was that forests are extremely sensitive to temperature changes. Even a warming of one degree C. could cause the southern borders of forests throughout the country to move northward. Within 70 years forests in southern states such as Mississippi and Georgia could disappear altogether and be replaced by grasslands. In northern states such as Minnesota and Michigan, boreal trees such as balsam fir and birch may die out, leaving behind only a few scattered oaks.

The news is not all bad for all regions. The EPA suggests that some northern areas may have extended growing seasons and lower heating bills. The construction of dikes could save coastal real estate, and better irrigation and the development of heat-resistant crops could keep southern farmland productive. The dredging of harbors in the Great Lakes, which may sink by as much as two meters owing to higher rates of evaporation, may keep ships from running aground. "Adaptation can occur," Smith notes, "but it will be very expensive."

The EPA is now working on a follow-up report that will recommend ways to reduce or stabilize the emission of greenhouse gases. Congress, impatient, is already considering several bills that have this goal. "There's no question that this will be the big issue of the next decade," says Ron Cooper, a staff member of the Senate Committee on Environment and Public

Early signs of a "greenhouse" warming that could push sea levels higher may already be evident



SEA ICE may already be showing the influence of a global warming, according to Per Gloersen of the National Aeronautics and Space Administration's Goddard Space Flight Center and William J. Campbell of the U.S. Geological Survey. After analyzing 15 years of data from the Nimbus 5 and Nimbus 7 satellites, Gloersen and Campbell concluded that the perimeter of the sea ice around the earth's poles has been contracting steadily. This image, based on data collected by Nimbus 7 in February, 1987, shows variations in the coverage of Arctic sea ice, ranging from zero (pale blue) to 100 percent (purple). The gray areas represent landmasses and the black circle represents a region outside the satellite's view.

Works. "The questions will be: How fast do we move, and do we bring the rest of the world along or do we act on our own?" —J.H.

Time after Time

Once more the black-hole time machine

The journal *Physical Review Letters* is not generally considered light reading, but a recent paper by Michael S. Morris, Kip S. Thorne and Ulvi Yurtsever of the California Institute of Technology seems designed to produce a few smiles. The authors propose to use a black hole as a time machine.

The idea is actually not new. Since the early decades of this century it has been known that a black hole is one end of a "wormhole" that connects two different regions of spacetime. Unfortunately such passageways do not make a useful rapid-transit system. In a nonrotating black hole the wormhole contains a singularity, or point of infinite gravitational and tidal forces, that destroys any would-be commuter before he or she gets through the turnstile.

In both rotating and electrically charged black holes the wormhole does not encounter a singularity. This fact led to much speculation in the 1960's that a cosmonaut might travel through the black hole and emerge in a different region of spacetime. Within a decade, however, investigators had shown that the wormhole in a charged or rotating black hole is unstable; any attempt to traverse or send a signal through it causes the passage to collapse into a singularity with the above-mentioned results.

The Caltech report is a clever attempt to solve the problem for would-be spacetime travelers. The authors suggest some advanced civilization might first "plausibly" extract a wormhole from quantum foam—a state of spacetime that might exist where the curvature is so high that relativity becomes wedded to quantum mechanics and wormholes are created. Such regions would have dimensions of about 10^{-33} centimeter. Once having extracted the wormhole by techniques still to be developed, the civilization would enlarge it to a size suitable for transportation of people.

The key to the Caltech proposal is to place one perfectly conducting charged electrical plate on each side of the wormhole throat. By a phenom-

FREE SCIENCE BOOKS CATALOG

Send for your free catalog of new and recent publications from W. H. Freeman and Company, the book publishing arm of *Scientific American* for more than 20 years. Discounts and bonus books offered.

- Yes, I'd like a free copy of your latest science book catalog.
 Please send literature about the Scientific American Library, too.

Name _____

Address _____

City, State, Zip _____



Mail to: Gail Harleston
 W. H. Freeman and Company, 41 Madison Avenue, New York, NY 10010

Exercise More with Less

■ **More Effective** By duplicating the motion of cross country skiing, the world's best exercise, NordicTrack provides the ideal aerobic workout.

■ **More Complete**

Unlike bikes and other sitdown exercisers, NordicTrack exercises all the body's major muscles for a total body workout.

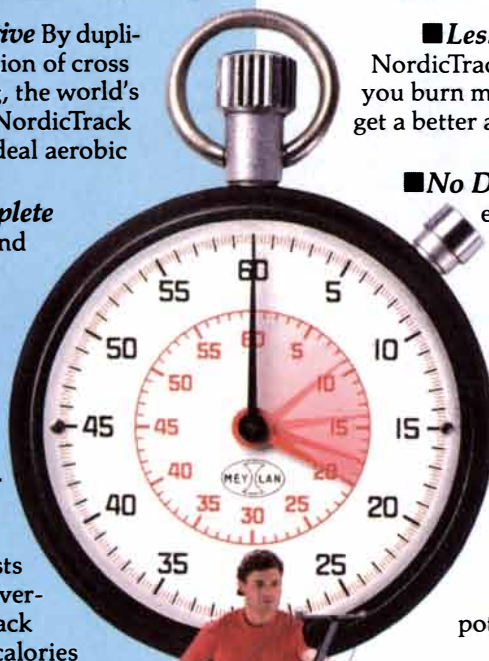
■ **More Calories Burned**

In tests at a major university, NordicTrack burned more calories than an exercise bike and a rowing machine.*

■ **More Convenient** With

NordicTrack, you can exercise in the comfort of your home. NordicTrack easily folds, requiring storage space of only 17" x 23".

*Scientific test results included in NordicTrack brochure.



■ **Less Time** Because NordicTrack is so efficient, you burn more calories and get a better aerobic workout in less time.

■ **No Dieting** No other exercise machine

burns more calories than NordicTrack... So you can lose weight faster without dieting.

■ **No Impact**

Running and some aerobic workouts can cause painful and potentially harmful jarring. A NordicTrack workout is completely jarless.

■ **No Skiing Experience Required**

Even if you've never skied, in a few minutes you'll soon be "tracking" your way to better health.

NordicTrack
 THE BEST WAY TO FITNESS

A CML COMPANY © 1988 NordicTrack

Bill Koch
 Olympic
 Silver Medalist

FREE BROCHURE AND VIDEO

Call Toll Free Or Write:
1-800-328-5888

In Canada 1-800-433-9582
 141 Jonathan Blvd. N., Chaska, MN 55318

Please send free brochure
 Also free video tape VHS BETA

Name _____

Street _____

City _____ State _____ Zip _____

Phone () _____ 320A9

enon known as the Casimir effect, the plates cause a violation of the "weak energy condition." All normal matter obeys the weak energy condition, which can be thought of as the requirement that the average energy density of the material be nonnegative—that is, zero or positive. It is the weak energy condition that causes singularities to form in black holes. If it is violated, a singularity need not be inevitable.

Once the wormhole is in place, it can then be converted into a time machine as follows: one end is accelerated away from the other "gravitationally or electrically" to nearly the speed of light and brought back again. According to relativity, the moving mouth "ages" less than the stationary mouth; it is easy to show that a voyager traversing the wormhole from the moving end to the stationary end will consequently travel backward in time.

The Caltech time machine apparently has an advantage over its predecessors in that the wormhole seems to be stable. Nevertheless, the authors appear reluctant to state this categorically. Moreover, in order for the Casimir effect to work, the plates might have to be placed closer together than the radius of an electron, and this could be forbidden. In any case, it will be a few years before the machine appears in the Hammacher Schlemmer catalogue. —Tony Rothman

TECHNOLOGY

Hive Technology

An SDI researcher invents "killer bee" detectors

Howard T. Kerr is an engineer at the Oak Ridge National Laboratory; he has designed components for nuclear reactors and for the Strategic Defense Initiative. Kerr is also an apiarist; he tends 140 colonies of honeybees and heads a group called the Beekeepers of Tennessee. More than a decade ago, worried that so-called Africanized, or "killer," bees migrating north from South America posed a threat to American beekeeping, Kerr decided to apply his engineering talent to stopping or at least understanding them.

First he tried to develop a way to find and track swarms in the wild. Easy, he thought: just set out bee bait spiked with radioactive isotopes and then wander the countryside with a Geiger counter. On second thought



MINIATURE TRANSMITTER glued to the back of a bee might enable investigators to track its flights and thereby determine its foraging and mating behavior. The device is powered by an array of tiny solar cells and transmits infrared pulses that can be detected at a range of more than 1,000 meters. Diane D. Falter, Kelly Falter, Kenneth Valentine and Gary T. Alley of the Oak Ridge National Laboratory developed the device at the suggestion of investigators studying Africanized bees; they plan to test a prototype this spring.

Kerr realized this approach "might have some problems." He then considered using an infrared camera to spot swarms. But how could Africanized bees be distinguished from other heat-emitting animals, including other types of bees?

In the early 1980's Kerr attended a lecture by Anita M. Collins of the U.S. Department of Agriculture, who mentioned that Africanized bees "sound different" from honeybees native to North America. "A light went on in my head," Kerr recalls. Working with Collins, he found that Africanized bees beat their wings faster and therefore buzz at a higher pitch than domestic bees. He then designed a hand-held acoustic analyzer that flashes a red light when it detects Africanized bees and a green light when it detects domestic bees. "It's a Mickey Mouse device," Kerr notes, "but it's very useful."

Recently Kerr and others at Oak Ridge who have come to share his preoccupation have sought ways to track not just swarms but individual bees. Kerr's co-workers are designing a solar-powered infrared transmitter that could be mounted on the back of a bee and monitored at ranges of up to a mile. Kerr has also outfitted bees with miniature reflectors that send a laser beam directly back to its source. With these devices Africanized-bee investigators could track individual

queens and drones, establish mating patterns and thereby devise strategies for blocking reproduction.

Kerr and his colleagues have even borrowed a tracking device from the SDI arsenal. Called a Doppler laser, it can detect the motion of distant objects—such as nuclear warheads—with high precision. Theoretically the laser could spot a single bee and determine the frequency of its buzzing from thousands of meters away. "The joke," Kerr says, "is that we'd point this thing across the Rio Grande, and when we detected an Africanized bee, we'd just turn up the power and burn it up." He stresses that such a defense would be prohibitively expensive. "H'm," he muses. "That sounds familiar, doesn't it?" —J.H.

Hostile Takeovers

How can a computer network welcome only friendly users?

Most of the world was taken by surprise in early November when a small group of programs—a "virus" or "worm"—designed by a young computer hacker temporarily shut down the Internet, a network connecting thousands of research computers across the U.S. The programs, playing a kind of electronic leapfrog, spread from machine to machine, running multiple copies of themselves and crowding out other tasks. National security was never at issue (the Pentagon's involvement in the Internet is limited to funding connections at universities and think tanks), but the disruption did raise troubling questions about the security of other, more vital networks, such as those that transfer close to \$700 billion every day among banks.

For network designers, says Charles Wilk of the congressional Office of Technology Assessment, the question has always been: "How can you be user-friendly to friendly users and not to unfriendly ones?"

The architects of the Internet "made a conscious decision to go for high functionality, accepting that there would be some security cost," according to Randall L. Frank, who is in charge of computer systems at the University of Michigan. Indeed, the loopholes exploited by the Internet worm were well known to network cognoscenti, but only a few network sites had invested the time to close them.

The combination of features that made the Internet so easy to shut

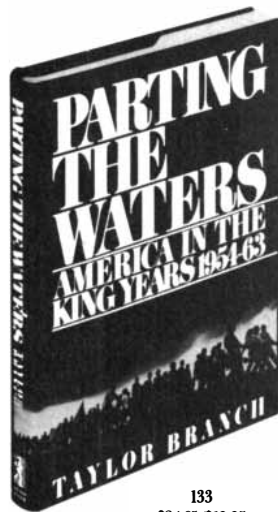
THE EPIC STORY OF AMERICA'S MOST PASSIONATE ERA

Now you may include *Parting the Waters* in your choice of **ANY 4, ALL FOR \$2.**

You simply agree to buy just 1 more book in the next six months.
(First price under each book is Publisher's List. **Boldface** shows Club Price.)

In *Parting the Waters*, the first volume of his *America in the King Years*, journalist-historian Taylor Branch begins the story of how Martin Luther King changed the course of American history. Branch chronicles the civil-rights struggle and lets us see the rise to greatness of a unique leader. Along the way, we experience the arrest of Rosa Parks, the Montgomery bus boycott, the lunch counter sit-ins, the bloody freedom rides, the siege of Birmingham, the murder of Medgar Evers. And we meet a gallery of characters: John and Robert Kennedy, J. Edgar Hoover, politicians and judges, movie stars and students, government agents, journalists and ordinary citizens whose lives were altered forever.

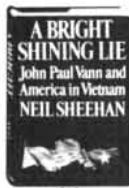
Monumental in scope and impact, *Parting the Waters* is a masterpiece of reporting and leaves no doubt why we celebrate a holiday in King's name. Destined to become a classic, it can be yours now as part of this very special Book-of-the-Month Club offer.



133
\$24.95/\$19.95



808
\$29.95/\$24.95



810
\$24.95/\$19.95



067
\$19.95/\$17.95



484
\$16.95/\$14.95



809
\$22.50/\$19.95



621
\$19.95/\$17.95



609
\$19.95/\$17.95



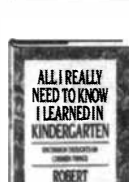
147
\$18.95/\$16.95



632
\$22.50/\$18.95



758
\$15.95/\$13.95



807
\$15.95/\$13.95



812
\$10.95/\$9.95



759
\$27.50/\$21.95



836
\$19.95/\$17.95



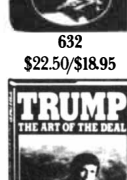
767
\$18.95/\$16.95



496
\$19.95/\$17.95



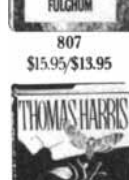
704
\$22.95/\$18.95



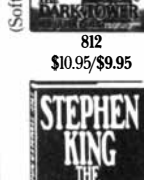
666
\$19.95/\$17.95



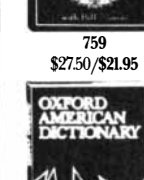
814
\$22.95/\$19.95



779
\$18.95/\$16.95



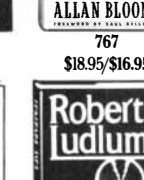
337
\$19.95/\$16.95



624
\$16.95/\$14.95



301
\$17.95



189
\$19.95/\$16.95



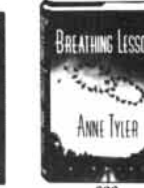
028
\$17.95/\$14.95



466
\$29.95/\$21.95



713
\$24.95/\$19.95



623
\$18.95/\$16.95



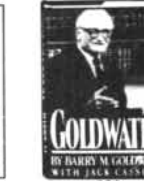
848
\$29.95/\$23.95



799
\$19.95/\$17.95



829
\$22.95/\$18.95



830
\$21.95/\$18.95



034
\$16.95/\$14.95



051
\$18.95/\$16.50

Benefits of Membership. As a member you will receive the *Book-of-the-Month Club News*® 15 times a year (about every 3½ weeks). Every issue reviews a Selection and more than 150 other books, which are carefully chosen by our editors. If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more other books—or no books at all—indicate your decision on the Reply Form and return it by the specified date. A shipping and handling charge is added to each shipment. *Return Privilege:* If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit. *Cancellations:* Membership may be discontinued, either by you or by the Club, at any time after you have bought one book.

Please enroll me as a member of Book-of-the-Month Club and send me the 4 books I've listed below, billing me \$2, plus shipping and handling charges. I agree to buy 1 more book during the next six months. A shipping and handling charge is added to each shipment.

Indicate by number the 4 books you want: **9-51**

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------

50% money-saving option. I may choose a 5th book now at half the Club price, plus shipping and handling, with no obligation to buy another book.

9-79

Book-of-the-Month Club, Inc.
P.O. Box 8803, Camp Hill, PA 17011-8803

Name _____ (Please print plainly) **A170-1**

Address _____ Apt. _____

City _____

State _____ Zip _____

© 1988 BOMC Prices generally higher in Canada.
All orders are subject to approval.

BOOK-OF-THE-MONTH CLUB®

SCORPIO

SCORPIO

SCORPIO

SCORPIO

Buckle up—together we can save lives.



IT RUNS IN THE BLACK FOREST.

As well-mannered as they are on the road, the performance sedans of Germany sometimes seem to forget their manners in dealing with humans. Putting drivers in seats that are sternly unyielding. And even depriving passengers of decent legroom.

Fortunately, Scorpio shatters the myth that owning a precision-engineered

German car requires sacrificing personal comfort.

Ease into Scorpio's driver's seat and feel the soft available leather gently support your back. Ask your companions how they're enjoying the power-reclining rear seats and over three feet of rear legroom.

Then, for the most comfortable feeling of all, turn the ignition. And put your-

self in command of Scorpio's 2.9-liter, autobahn-bred, fuel-injected V-6 and standard anti-lock brakes (ABS).


Scorpio keeps owners comfortable in other ways, too: with the Scorpio Guaranteed Resale Value Program. Membership in Ford Auto Club's Roadcare Plan. And with the Scorpio Free Loaner Program at participating dealers. Some

SCORPIO

SCORPIO

SCORPIO

SCORPIO

LINCOLN-MERCURY DIVISION 



BUT IT'S ALSO NICE TO PEOPLE.

restrictions apply, so ask to see a copy of these programs at your dealer. Scorpio. It's one German touring sedan that treats drivers with utmost respect. Yet still treats passengers with uncommon kindness. For more information, call 1-800-822-9292.



SCORPIO. GERMAN PERFORMANCE YOU CAN BE COMFORTABLE WITH.
Imported from Germany for select Lincoln-Mercury dealers.

SCIENTIFIC AMERICAN

In Other Languages

LE SCIENZE

L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:

Le Scienze S.p.A., Via G. De Alessandri, 11
20144 Milano, Italy

Advertising correspondence:

Publietas, S.p.A., Via Cino de Duca, 5,
20122 Milano, Italy

サイエンス

Y950/copy Y10,440/year Y14,000/(abroad)

Editorial, subscription, advertising correspondence:

Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y

CIENCIA

450 Ptas/copy 4950 Ptas/year \$35/(abroad)

Editorial, subscription, advertising correspondence:

Prensa Científica S.A.,
Calabria, 235-239
08029 Barcelona, Spain

SCIENCE

27FF/copy 265FF/year 315FF/year (abroad)

Editorial, subscription, advertising correspondence:

Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum

9.80 DM/copy 99 DM/year 112.20 DM/(abroad)

Editorial, subscription correspondence:

Spektrum der Wissenschaft GmbH & Co.
Moenchhofstrasse, 15
D-6900 Heidelberg,
Federal Republic of Germany

Advertising correspondence:

Gesellschaft Für Wirtschaftspublizistik
Kaserenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

科学

1.40RMB/copy 16RMB/year \$24/(abroad)

Editorial, subscription correspondence:

ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

B MIPEHAYKI

2R/copy 24R/year \$70/(abroad)

Editorial correspondence:

MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.

Subscription correspondence:

Victor Kamkin, Inc.
12224 Parklawn Drive,
Rockville, MD 20852, USA

TUDOMÁNY

98Ft/copy 1,176Ft/year 2,100Ft/(abroad)

Editorial correspondence:

TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary

Subscription correspondence:

"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

العلوم

1KD/copy 10KD/year \$40/(abroad)

Editorial, subscription, advertising correspondence:

MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 - Kuwait

Advertising correspondence all editions:

SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0550 Telex: 236115

down was its accessibility to all kinds of users—from undergraduate biologists to corporate microchip fabricators—its very high speed and its relatively open design. These same features, of course, give the network its utility. The ability to run a program on any machine in the network from any other machine, for example, underlies distributed file systems, the network's directory services, electronic mail forwarding and many other features. Local area networks (LAN's), which also combine high data rates with easy access and low security (but not nationwide links), could well be the next target for some malevolent "cracker," notes Eugene H. Spafford of Purdue University. Although a few LAN's encrypt at least the passwords that travel over them, most can be tapped by anyone with a personal computer and a coaxial-cable connector.

Security generally bears an inverse relation to flexibility and performance. A network known as uucp (for UNIX-to-UNIX copy, so named because it connects computers running the UNIX operating system) is also quite accessible, but it transmits data by way of standard telephone lines, which carry a few thousand bits per second, in contrast to the 50,000 to three million bits per second of the specialized Internet connections. Although a similar virus could attack uucp, Spafford comments, it could spread to only a handful of machines in the day or so that might pass before it was discovered. The Fedwire network, which transfers almost \$700 million a day among U.S. financial institutions, is still more secure, but computer-security consultant Robert Courtney describes it as little more than "a high-speed Western Union." And even the Fedwire suffers from fraudulent transactions costing its users several hundred million dollars each year before recoveries.

Users stand to lose much more by disconnecting from the outside world than they stand to gain in security. How, then, can a network be made more secure? The answer, if there is one, is a military secret, but encryption techniques are a big part of it. Not only the stored passwords but also the network transmissions themselves must be in code. Such encoding, however, reduces network performance, requires additional hardware and expense and costs the time and trouble of managing and distributing keys to authorized users.

Even encryption may not be foolproof. Users can reveal keys, and computer attacks are becoming more sophisticated as well. The Internet worm

rooted out passwords stored in encrypted form by running through its own encrypted list of commonly used passwords. Several computer scientists have estimated that encrypting an entire 50,000-word dictionary and comparing passwords against it could take less than an hour. (Nonword passwords and an inaccessible password file are fixes being made on the Internet even now.)

Nevertheless, for some networks encryption is probably worth the effort. Courtney cites the IBM Corporation's internal network: in Europe it runs through lines controlled by state-owned telephone companies, and partially or wholly state-owned computer companies are among IBM's European competition. Without encryption, he said, no one would dare to use the network. The same is certainly true of Milnet, the Department of Defense's network for classified information.

For the Internet and other nets like it, however, radical new security measures are not likely to be proposed or implemented. No one trusts truly sensitive data to these channels, Courtney says. And disconnecting from a net would cost users much more than they might gain. In any case, what many network users fear most is not malice but chance—incidents such as the fire that last year destroyed an Illinois Bell switching computer, costing telephone customers and local authorities more than \$500 million, by some estimates.

Courtney thinks the Internet incident was "blown way out of proportion." Computer users should be accustomed to having their machines fail now and then, for one reason or another, and plan accordingly, he said. Unfortunately the myth of invulnerability persists. Courtney recalled in particular the contingency plan drawn up for a Defense Department computer system serving 3,500 people: "The third question was, 'Is your system subject to acts of God?' and the answer was no."
—Paul Wallich

BIOLOGICAL SCIENCES

Preemptive Strike

The body can be made to see foreign cells as "self"

One long-sought goal in medical research is to switch off part of the immune response while leaving the rest of the immune system unaffected. At present organ-trans-

plant patients are usually given immunosuppressive drugs such as cyclosporine A to prevent their immune systems from rejecting the foreign tissue. But treatment usually must be continued for life, and the drugs inhibit the immune response to infection as well as to the new organ. What if the immune system could be taught to view the new tissue as "self" and ignore it, just as it ignores the body's own tissue? Early clinical trials of the approach have produced favorable results.

Jerzy W. Kupiec-Weglinsky and his co-workers at the Harvard Medical School and the Brigham and Women's Hospital in Boston employ monoclonal antibodies that seek targets found on a set of immune-system cells only when those cells have been activated by a foreign substance. The cells are *T* cells, a class of white blood cells that play a major role in transplant rejection. The antibodies' targets on the activated *T* cells are receptors for interleukin-2, a substance that spurs the cells to proliferate.

The strategy assumes that by blocking the interleukin-2 receptors on activated *T* cells, the monoclonal antibodies will prevent those *T* cells from proliferating. *T* cells not actively engaged in an immune response—ones that failed to recognize the new organ as foreign—should be spared for future service. In essence the strategy attempts to mimic the process by which an immune system sometimes learns naturally to recognize a new organ as self.

The approach works in experimental animals: rats treated with antibodies to the interleukin-2 receptor fail to reject transplanted heart tissue. Recent experiments confirm that activated *T* cells fail to proliferate in rats given the antibody treatment. "Suppressor" *T* cells, on the other hand, which inhibit the immune response, become commoner. Moreover, when the antibodies are administered together with cyclosporine A, they augment the effect of the drug in suppressing transplant rejection.

Jean-Paul Soulillou and Y. Jacques at the French National Institute of Health and Medical Research (INSERM) in Nantes pioneered the experimental use of the technique in human beings in 1986. They report that rejection episodes in kidney-transplant patients given monoclonal antibodies to the interleukin-2 receptor as well as immunosuppressive drugs are much less common than they are in patients given only the drugs. Kupiec-Weglinsky and his collaborators are now

also finding that antibody treatment in conjunction with drug therapy delays rejection episodes in transplant patients and makes such episodes less frequent.

Instead of monoclonal antibodies, Samuel Strober and his colleagues at the Stanford University School of Medicine are employing radiation to disable *T* cells selectively before the transplant operation. In their technique, called total lymphoid irradiation, radiation is directed at sites where mature *T* cells are found, principally lymph nodes. The bone marrow, which contains stem cells that continually regenerate the immune system, is protected. The aim is to knock out *T* cells that might react to the transplant without permanently destroying the immune system. Once a transplanted organ has been in place for a few weeks, the immune system accepts it as self. Strober says that three kidney-transplant recipients who received the treatment have long-surviving grafts and now need no immunosuppressive drugs; other patients who have undergone similar therapy need lower doses of immunosuppressive drugs than they would ordinarily.

Other work suggests that suppression of specific immune responses might one day be achieved by gene therapy. Joren C. Madsen, Kathryn J. Wood and their co-workers at the John Radcliffe Hospital in Oxford reported last year that specific tolerance of a heart-tissue graft could be induced in mice by pretreating them with cells that had previously been removed and genetically manipulated. The cells—fibroblasts, or connective-tissue precursors—had been altered to carry histocompatibility genes taken from the donor mouse. Histocompatibility genes specify proteins that appear on cell surfaces and are important markers for the immune system, helping it to distinguish self from "nonself."

The Oxford workers suggest that during the pretreatment with the genetically engineered cells, the immune system of the recipient mice learned to recognize the new histocompatibility proteins as self. The mice therefore did not later react to the same histocompatibility proteins on the graft. Surprisingly, one particular donor gene on its own could forestall rejection, even though the cells in the graft carried a host of other histocompatibility proteins. The workers propose a hierarchy of influence among the markers: tolerance to one "dominant" marker may result in tolerance to a variety of related proteins on the same cell. Wood says the team hopes

to identify dominant markers on human cells and eventually try the technique on human patients. —T.M.B.

Memory in a Neuron

A nerve cell changes shape after conditioned learning

When Pavlov's dog learned to salivate at the clang of a bell, what changed in its brain? Behavioral psychologists may once have considered the question irrelevant, but for modern neuroscientists it is a burning issue. A number of studies suggest that specific molecular and electrophysiological changes take place in neurons, or nerve cells, when an animal undergoes associative conditioning—when it learns, as Pavlov's dog did, to associate a stimulus with a previously unrelated response. At the annual meeting in November of the Society for Neuroscience, Daniel L. Alkon of the National Institute of Neurological and Communicative Disorders and Stroke presented a significant new finding: a neuron can change its structure in the course of associative learning. Structural change in neurons is usually seen only in embryos and young animals in conjunction with normal development. "It has never before been seen in associative learning and memory," Alkon notes.

Alkon's group studies *Hermisenda*, a dainty, alabaster sea slug streaked with neon blue and orange racing stripes. The sea slug instinctively swims toward light, a response that draws it to the ocean surface, where it feeds. In Alkon's laboratory the animal is trained to acquire a different response: every time it sees a flash of light the chamber in which it is housed is whirled on a turntable, which causes the animal to grip the chamber wall more tightly by clenching its foot muscle. Eventually it clenches its foot whenever it sees the light.

One group of animals was thus trained and a second group was left untrained to serve as a control. A third group was subjected to the light flashes and rotation in a random rather than associated order to ensure that any changes observed in the trained group were related specifically to the learning of the new response—that they were not direct responses to the stimulus itself. The investigators then examined a particular neuron called a type *B* photoreceptor, one of the five light-detecting neurons in the animal's eye. The cell's axon (the fiberlike process that carries nerve impulses

HELPING COMPUTER PROGRAMMERS MAKE FAST AND STEADY PROGRESS.

Until now, manual skills were a big part of writing complicated computer programs. They required a major investment in time, effort and money — and days of boredom for the applications designer.

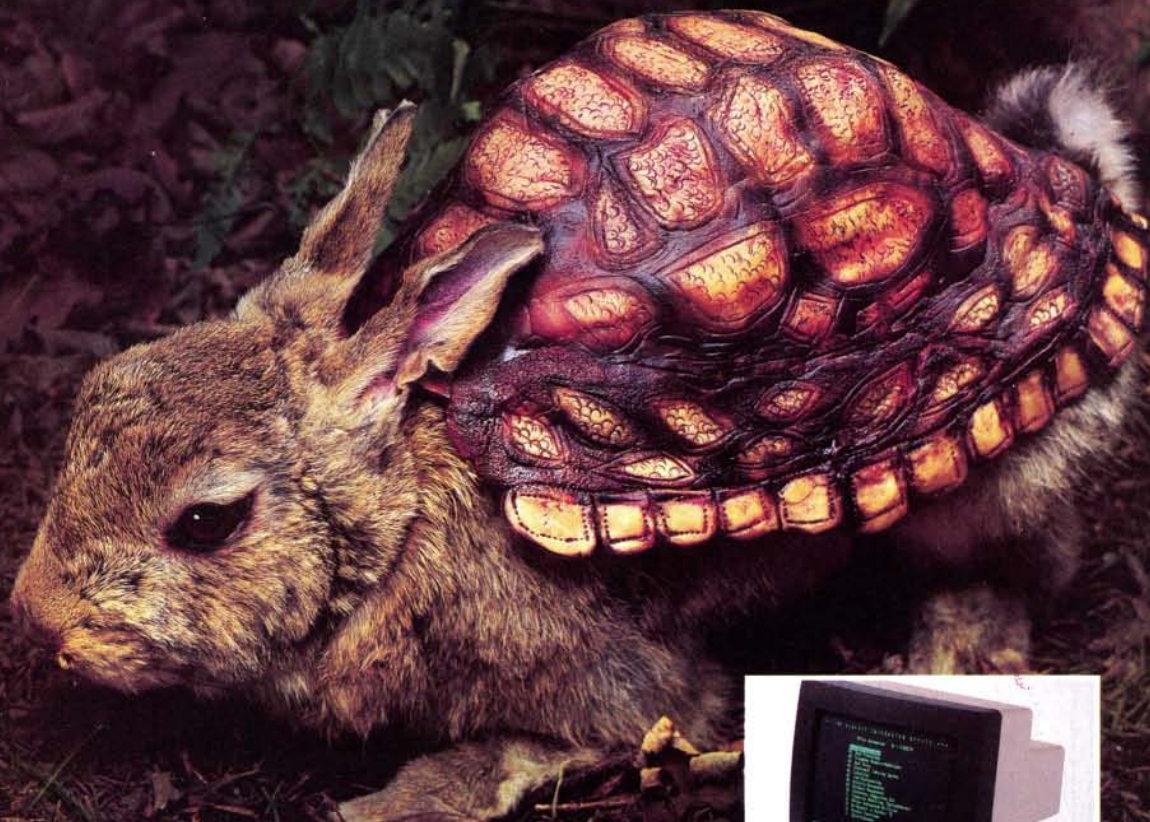
McDonnell Douglas information systems specialists bypass the dull, turtle-speed tedium formerly necessary for good programs. Their Pro-IV® software maps out the best program path and selects routine computer instructions for them.

The result is a working program that

emerges with hare-like speed, ready to use, allowing application systems designers to fully concentrate on the special characteristics of the project.

Designers like Pro-IV software because it lets them try out ideas in prototype programs, quickly and easily, selecting only the best. Managements like it because it's faster, cheaper and yields better results.

For more information, write: Software, McDonnell Douglas, Box 14526, St. Louis, MO 63178



MCDONNELL DOUGLAS

INFORMATION SYSTEMS

MILITARY & COMMERCIAL AIRCRAFT

SPACECRAFT & MISSILES

TRAVEL MANAGEMENT

HELICOPTERS

FINANCING

WHY YOU SHOULD CONSIDER 386 SYSTEMS, DESPITE THEIR

Our new 386-based systems are priced about 35% less than comparable systems—like Compaq's. Which may make you wonder if we've left something important out. Like high performance.

Well we haven't.

In fact, these are among the fastest 386-based systems available. With more advanced features than you'd get in systems that list for up to \$3000 more.

Like Compaq's.

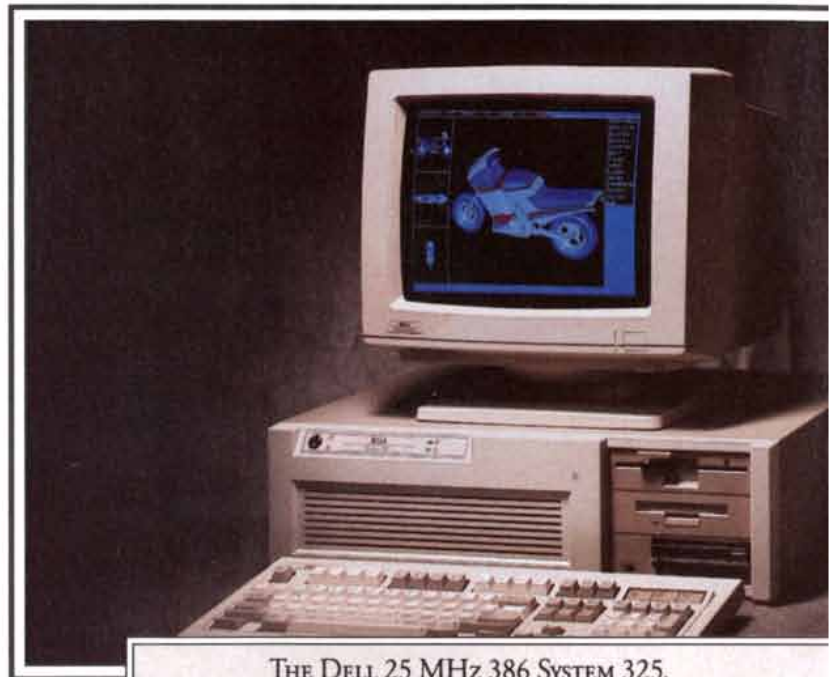
For instance, our 20 MHz System 310 offers you the most extraordinary value available in any 386-based system. It's the machine that PC Magazine (6/14/88) described as "fast enough to burn the sand off a desert floor."

AND IF THAT SOUNDS FAST, WAIT TILL YOU SEE OUR NEW 25 MHz 386-BASED SYSTEM.

At 25 MHz, our new System 325 offers you the highest possible performance in a 386.

Like the System 310, it utilizes the very latest technology, including the Intel 82385 Cache Memory Controller, advanced 32-bit architecture and high performance drives. And of course, both systems are fully IBM PC compatible.

But speed isn't the only reason to buy from Dell. Or even the best.



THE DELL 25 MHz 386 SYSTEM 325.

STANDARD FEATURES: • Intel 80386 microprocessor running at 25 MHz. • 1 MB of RAM* expandable to 16 MB using a dedicated high speed 32-bit memory slot. • Advanced Intel 82385 Cache Memory Controller with 32 KB of high speed static RAM cache. • Page mode interleaved memory architecture. • VGA systems include a high performance 16-bit video adapter. • Socket for 25 MHz Intel 80387 or 25 MHz WEITEK 3167 math coprocessor. • 5.25" 1.2 MB or 3.5" 1.44 MB diskette drive. • Enhanced 101-key keyboard. • 1 parallel and 2 serial ports. • 200-watt power supply. • 8 industry standard expansion slots.

**Lease for as low as \$252/Month.

The Dell System 325 is an FCC Class A device, intended for business use only.

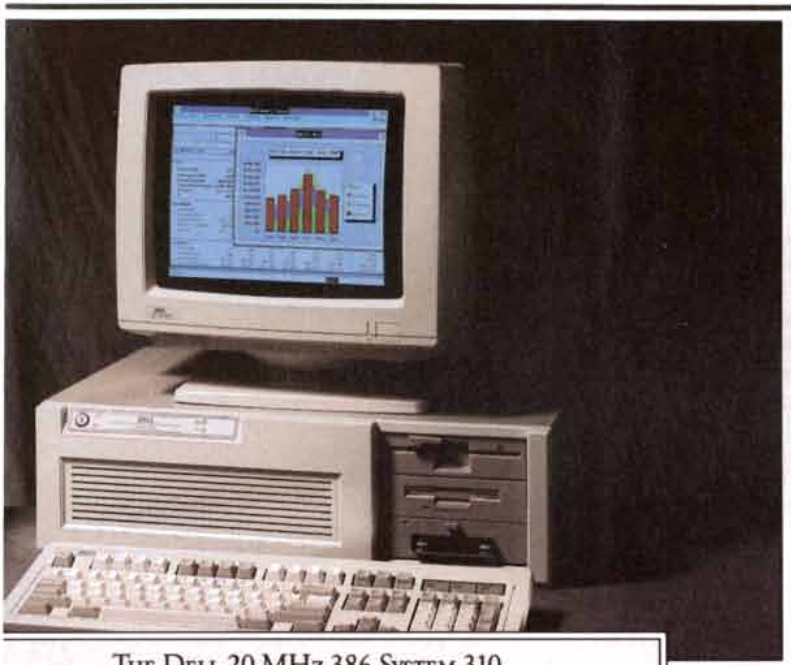
SYSTEM 325 Hard Disk Drives	WITH MONITOR & ADAPTER	
	VGA Mono	VGA Color Plus
150 MB-18 ms ESDI	\$6,999	\$7,299
322 MB-18 ms ESDI	\$8,999	\$9,299

SYSTEM 325 AND 310 OPTIONS: • Intel 80387 math coprocessor: 25 MHz for 325; 20 MHz for 310. • 1 MB or 4 MB memory upgrade kit. • 2 MB or 8 MB memory expansion board kit. • Dell Enhanced Microsoft® MS-DOS® 3.3. • Dell Enhanced Microsoft MS-DOS 4.0. • Both

THE FIRST PERSONAL COMPUTER THAT'S TRULY PERSONAL.

When you order from Dell, we custom configure a system to your exact personal specifications. After evaluating your busi-

CONSIDER THE NEW DELL SUSPICIOUSLY LOW PRICES.



THE DELL 20 MHz 386 SYSTEM 310.

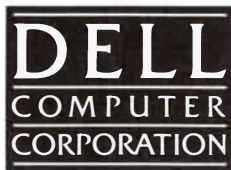
STANDARD FEATURES: • Intel 80386 microprocessor running at 20 MHz. • 1 MB of RAM* expandable to 16 MB using a dedicated high speed 32-bit memory expansion slot. • Advanced Intel 32385 Cache Memory Controller with 32 KB of high speed static RAM cache. • Page mode interleaved memory architecture. • VGA systems include a high performance 16-bit video adapter. • Socket for 20 MHz Intel 80387 or 20 MHz WEITEK 3167 math coprocessor. • 5.25" 1.2 MB or 3.5" 1.44 MB diskette drive. • Enhanced 101-key keyboard. • 1 parallel and 2 serial ports. • 200-watt power supply. • 8 industry standard expansion slots.

**Lease for as low as \$148/Month.

SYSTEM 310 Hard Disk Drives	WITH MONITOR & ADAPTER	
	VGA Mono	VGA Color Plus
40 MB-28ms	\$4,099	\$4,399
90MB-18ms ESDI	\$4,899	\$5,199
150MB-18ms ESDI	\$5,399	\$5,699
322MB-18ms ESDI	\$7,399	\$7,699

MS-DOS versions with disk cache and other utilities. • Dell Enhanced MS® OS/2. *640 KB is available for programs and data. The remaining 384 KB is reserved for use by the system to enhance performance.

business needs, we will help you select the features that are right for you. After your system unit is custom built, we burn-in everything to make sure the entire system works perfectly.



TOLL-FREE SUPPORT AND ON-SITE SERVICE INCLUDED IN THE PRICE.

Every Dell system includes a complete set of diagnostic tools. So troubleshooting is easy. In fact, most problems can be resolved over our toll-free support line. It's staffed by Dell's own expert technicians from 7 AM to 7 PM (CT) every business day.

TO ORDER, PLEASE CALL

800-426-5150

IN CANADA, CALL 800-387-5752

And if your system requires hands-on service, a technician will be at your location the next business day. At no cost to you.◊

Included in the price of your system is a full year of on-site service.

But that's not all. You're also protected by our 30-day money-back guarantee. And our one-year limited warranty on parts and workmanship.△

AND IF YOU STILL THINK YOU GET WHAT YOU PAY FOR, CONSIDER THIS.

When you buy from Dell, you buy directly from our manufacturing facility in Austin, Texas. Which means we eliminate dealer mark-ups, allowing us to give you a lot more 386 for less. We can even design a custom lease plan for your business, which gives you another way to save.

So go beyond your suspicions. Call us at (800) 426-5150 and order the system that's right for you.

away from the cell body) branches to form synaptic contacts with a variety of other neurons, including vestibular cells that relay nerve signals to motor neurons controlling the foot muscle.

When the photoreceptor cell was stained and examined under a microscope, the investigators could see the axon terminals fanning out like a river delta. They examined the same neuron in all the animals and discovered that in the trained animals the axon fanned out over a much smaller region (about 50 percent smaller) than the axons from the two control groups. As Alkon puts it, "It's as though they are focusing." What is more, the animals that learned to respond most strongly had the most tightly focused axon terminals. "We know the cell is making connections to many other cells," Alkon says. "We hypothesize that in focusing it is reducing certain connections and amplifying signals to a smaller number of cells."

What makes the new finding particularly intriguing is that Alkon's group had previously observed electrophysiological and molecular changes in the same photoreceptor cell. In trained animals the cell's membrane becomes better able to conduct electrical signals, which play an important role in regulating the transmission of mes-

sages to other cells. In addition an enzyme called protein kinase C appears to migrate from the interior of the cell to its membrane. Alkon and his colleagues recently proved that the enzyme acts on the ion channels that modulate the electrical signals generated by the cells. They now find that the cells with the most tightly focused axon terminals are the ones best able to transmit the signals.

It remains to be seen whether the observed structural change does in fact reinforce the link between the photoreceptor and the foot muscle. One way to tell would be to inject a different dye into each neuron contacted by the photoreceptor. It should be possible thereby to establish that the photoreceptor's synaptic interaction with certain cells has either increased or decreased.

Alkon thinks his findings in *Hermisenda* may shed light on learning in higher animals and ultimately in human beings. His group has found that protein kinase C migrates to the membrane in hippocampal cells from rabbits after the animals are associatively conditioned. The hippocampus is a part of the mammalian brain that is thought to mediate the acquisition of new memories. "We see the same biochemical and molecular changes in the rabbit hippocampus as we've seen in *Hermisenda*," Alkon says. "We'll be looking for the same structural changes as well." —June Kinoshita

i Teach Yourself Spanish!

You can . . . with Audio-Forum's famous self-study audio-cassette course. Choose your own time, pace, and place to learn the language, and save hundreds of dollars compared with the cost of a private language school.

Our course was developed by the U.S. State Department to train foreign-service personnel. It consists of twelve 90-minute cassettes plus accompanying text and manual. You'll learn naturally by listening to the cassettes (while driving, jogging, or whatever) and repeating during the pauses on the tape. By the end of the course, you'll be learning and speaking entirely in Spanish!

This State Department course comes in 2 volumes, each shipped in a handsome library binder. Order either, or save 10% by ordering both:

Vol. I: Basic Spanish. 12 cassettes, manual, and 464-p. text \$181.50 postpaid.

Vol. II: Intermediate Spanish. 8 cassettes, manual, and 614-p. text, \$151.50 postpaid.

(CT residents add sales tax.)

Full 3-week money-back guarantee if not completely satisfied.

Phone orders call toll-free 1-800-243-1234

Other Spanish-language materials available, as well as courses in 55 other languages. Write for free catalog.

AUDIO-FORUM
THE LANGUAGE SOURCE

Room C135, 96 Broad St.
Guilford, CT 06437 (203) 453-9794

Schizophrenic Results

A link and a nonlink to schizophrenia are found

Because schizophrenia—which is often characterized by hallucinations, delusions, disordered thoughts and a loss of affect—often afflicts several generations of a single family, many psychiatrists have long suspected that it has a genetic component. They have had little proof. Now a transatlantic team has found the first strong evidence that a genetic abnormality can indeed contribute to the debilitating disorder.

Hugh Gurling of University College and the Middlesex School of Medicine in London and his British, Icelandic and American colleagues analyzed chromosomes from at least three generations of seven British and Icelandic families known to have a high incidence of schizophrenia. In particular they traced the inheritance of two genetic markers known as restriction fragment length polymorphisms, or

RFLP's. The markers lie in a region of chromosome 5 known as 5q11-13.

RFLP's are detected by digesting the chromosomes with a set of restriction enzymes, which cut DNA at predictable sites, and then separating the resulting fragments by length. A fragment that often varies in length among different people is an RFLP; it defines a genetically variable part of the chromosome. If a particular form of an RFLP—that is, a particular fragment length—is consistently inherited along with a disease, the dual inheritance is strong evidence that an aberrant gene (or group of genes) within the RFLP or nearby has a role in the disease. The RFLP data collected by Gurling's group, reported in *Nature*, demonstrated an incontrovertible link between the target region of chromosome 5 and schizophrenia in two of the families, and a weaker but still positive link in the other families.

Even though the findings confirm that schizophrenia can have a genetic root, the results cannot be generalized to say that chromosome 5 has a role in all cases of schizophrenia. In fact, a similar study published in the same issue of *Nature* found no connection between the 5q11-13 region and schizophrenia in a large Swedish family with a history of the disorder. That study was done by Kenneth K. Kidd of the Yale University School of Medicine and nine other workers at Yale, the Stanford University School of Medicine and the Karolinska Institute in Stockholm.

Together the two papers support the growing conviction that schizophrenia is a heterogeneous disorder that can have many different causes, all producing similar effects. Sometimes a gene on chromosome 5 may be important; at other times another gene or some combination of genes may play a role. It is also possible, Gurling says, that sometimes schizophrenia has no genetic component.

An abnormality on chromosome 5 may turn out to be only a rare cause of schizophrenia. Nevertheless, as Eric S. Lander of Harvard University notes in a comment accompanying the two reports, the British work remains a milestone in schizophrenia research. It represents, he says, the "first step in using genetics to subdivide patients into more homogeneous groups." The discovery of an assortment of genetic and related biochemical abnormalities that can contribute to schizophrenia should eventually enable physicians to tailor their treatments to each case depending on the underlying physiological disorder. —Ricki Rusting

The BBC Language Course for Children Only Seven Years Old*... and She's Already Speaking French!



Give Your Child
That Critical
Early Advantage!

It's a scientific fact...and one of Nature's marvels. During the early years of childhood, the human mind is best programmed for learning a language — any language.

That's why children learn so much more easily than adults, even before being able to read. They learn the same way they learned English — naturally — by listening, seeing, and imitating. In the international world our children will compete in — where so many Europeans and Asians start a foreign language early — a second language will be essential. Vital for competing with polished and accomplished peers.

Sample ages for beginning a second language**			
Japan	Age 8	France	Pre-School
Sweden	Age 7	Spain	Pre-School
Austria	Age 8	Canada	Pre-School

** Ages represent top schools and programs; compulsory language education usually begins several years later

From the BBC, World Leaders in Language Education.

For the first time ever in the U.S.A., your child can learn French or Spanish using the most successful Language Course for Children ever created!

Muzzy, a unique video learning program, is produced by the BBC — the world's foremost teachers of language. Specifically designed for children (pre-school through age 12), Muzzy uses color animation, enchanting songs, and charming, involving characters (including Muzzy himself), and teaches children to absorb a new language the same way they learned English.

It's so easy and so much fun. In fact, most kids love to watch or listen over and over again, just like their favorite TV shows!

Complete Language Learning Course!

Everything needed for a child to master beginning French or Spanish is included. Four video cassettes. Two audio cassettes. An activity book and an excellent parent's instruction guide plus answer book. All in attractive, durable storage cases.

Through *listen-and-learn* and *see-and-learn*, your child will begin speaking a foreign language from the very first day! He or she can learn alone, or you can help and learn the language, too!

No Risk Guarantee!

Here is perhaps the greatest gift you will ever give your child... a second language. And at an astonishingly affordable price of just \$145†, payable in four credit card installments. And there's no risk! If you and your child are not absolutely delighted, you may return the course within 30 days for a full refund. Order today from Early Advantage, 47 Richards Avenue, Norwalk, Conn. 06857.

† Plus \$4.75 shipping/handling per course.



By exclusive arrangement with the British Broadcasting Co. A program proven with thousands of European youngsters. And the whole family can learn the language, too!

* Proven results for pre-school through age 12. © 1988 MBI

The BBC Language Course for Children

Early Advantage
47 Richards Avenue
Norwalk, Conn. 06857

Satisfaction
Guaranteed.

For Fastest Service – Call Toll-Free: 1-800-367-4534
In CT, AK, HI Call 1-203-855-8717

Yes! Please send me *The BBC Language Course(s) for Children* I have indicated. I understand only VHS format is available.

(Please check appropriate items.) Language: FRENCH SPANISH

Name _____ PLEASE PRINT CLEARLY
Address _____
City/State/Zip _____
Signature _____
(All orders subject to acceptance.)

Charge each of four equal monthly \$37.44* installments to my credit card:
 VISA MasterCard Diners Club American Express

Credit Card No. _____ Exp. Date _____

I prefer not to use my credit card and will pay by check. Enclosed is my deposit of \$50* for each course. I will pay the balance of \$99.75* as billed in three equal monthly installments.

*Includes one-time shipping/handling charge of \$4.75. Connecticut residents add 7 1/2% sales tax. Allow 6 to 10 weeks for shipment. 8434

On Science Advice to the President

Right after Sputnik, the White House created the best science-advisory system in its history. The president needs good science advice more than ever; it is time to resurrect the long defunct system

by Jerome B. Wiesner

During last year's presidential election campaign, a front-section story in the *Wall Street Journal* declared, "Science is big in this election year. Big politics, big applause and big headaches for whoever is elected." The story's list of major scientific issues ranged from the Strategic Defense Initiative (SDI) to a mission to send human beings to Mars. Curiously, it hardly noted the most urgent scientific and technological problems facing the nation: revitalizing U.S. civilian industry, rescuing an ailing planet and creating an educational system capable of meeting these challenges.

What the *Journal* story did reflect accurately was the widespread distortion of priorities in the nation's science policy. This distortion is not simply the result of the last Administration's political agenda. It is also directly attributable to the absence of an effective White House advisory group for science and technology. The SDI proceeded without adequate consideration of the views of independent experts. Efforts to protect the environment have been stalled by the Govern-

ment's reluctance to act on the growing scientific evidence for worldwide environmental degradation. Many Federal technical programs lack careful, knowledgeable supervision and have wasted money or failed outright.

The situation today is in many ways reminiscent of the one President Eisenhower faced at the time of the Sputnik satellite surprise in 1957. The Soviet launching shocked Eisenhower into recognizing how thoroughly isolated he was from Government research-and-development programs. Not only had he been unaware of problems in the U.S. Vanguard satellite program but also the responsible individuals themselves apparently did not recognize how bad it was. Even after Sputnik the president found that these people continued to defend the bungled U.S. program.

Eisenhower responded by creating the post of Special Assistant for Science and Technology and appointing James R. Killian, Jr., then president of the Massachusetts Institute of Technology, to fill it. At Killian's suggestion, Eisenhower moved the Science Advisory Committee that then existed under the Office of Defense Mobilization right into the Executive Office. Together with the committee, Killian developed a system for advising the president that grew and flourished until 1972, when it was abolished by President Nixon. (The committee's scientific judgments had collided repeated-

ly with Nixon's political goals.) Every president since then has tried to bring a source of science advice into the White House but has failed to find a means as effective as the system established by Killian.

President-elect Bush faces a critical decision. Never has there been a greater need for informed national leadership on issues relating to science and technology. But before the president can give direction to the nation's science and technology efforts, he will have to create a better science management system than has existed in the recent past. I believe the best way to achieve this is to resurrect the system initiated by President Eisenhower and nurtured by President Kennedy. It served both presidents well in times of national crisis, and it would serve equally well today.

If the president-elect wants to establish this kind of science advisory system, he will have to approach the task with determination. Many people would prefer that the president not have such a system, and if history is any guide, they will try to persuade him to settle for a modest arrangement. Those who head agencies having large technical programs would prefer that the president get most of his technical advice from them. Other groups having a vested interest in departmental programs feel the same way—defense and space

JEROME B. WIESNER was Special Assistant for Science and Technology under presidents John F. Kennedy and Lyndon B. Johnson. He was president of the Massachusetts Institute of Technology from 1971 to 1980 and is now president emeritus.

contractors, Government laboratories and quite a few members of Congress who have Federally funded programs in their states.

Why is it so difficult to help the president get independent advice on scientific issues or, for that matter, on any complex topic? The fundamental reason goes back to the Constitutional Convention. The nation's founders gave the president heavy responsibilities but did not say much about how he was to get advice and assistance. They obviously did not expect the world to become so complicated or the president's job so burdensome. They probably assumed that cabinet members would be a major source of counsel to the president, but they did not foresee that the departments would grow so numerous or that managing the cabinet officers would

in itself become part of the problem.

Presidents have always depended on hand-picked cabinet officers and assistants for day-to-day advice. Most cabinet officers, however, are so overwhelmed by the job of running their departments that they rarely have the time or energy to understand the president's special problems, particularly those that relate to science. Top people in the Department of Commerce, for example, have little understanding of industrial technology.

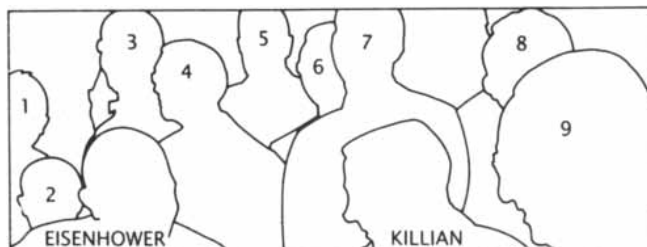
Another difficulty is that the departments have their own fish to fry. The Department of Defense, for example, is an enormous, self-perpetuating bureaucracy that regards its own survival and growth as its top priority. The Secretary of Defense has little control over it. Entrenched bureaucracies may try to thwart the deployment of new

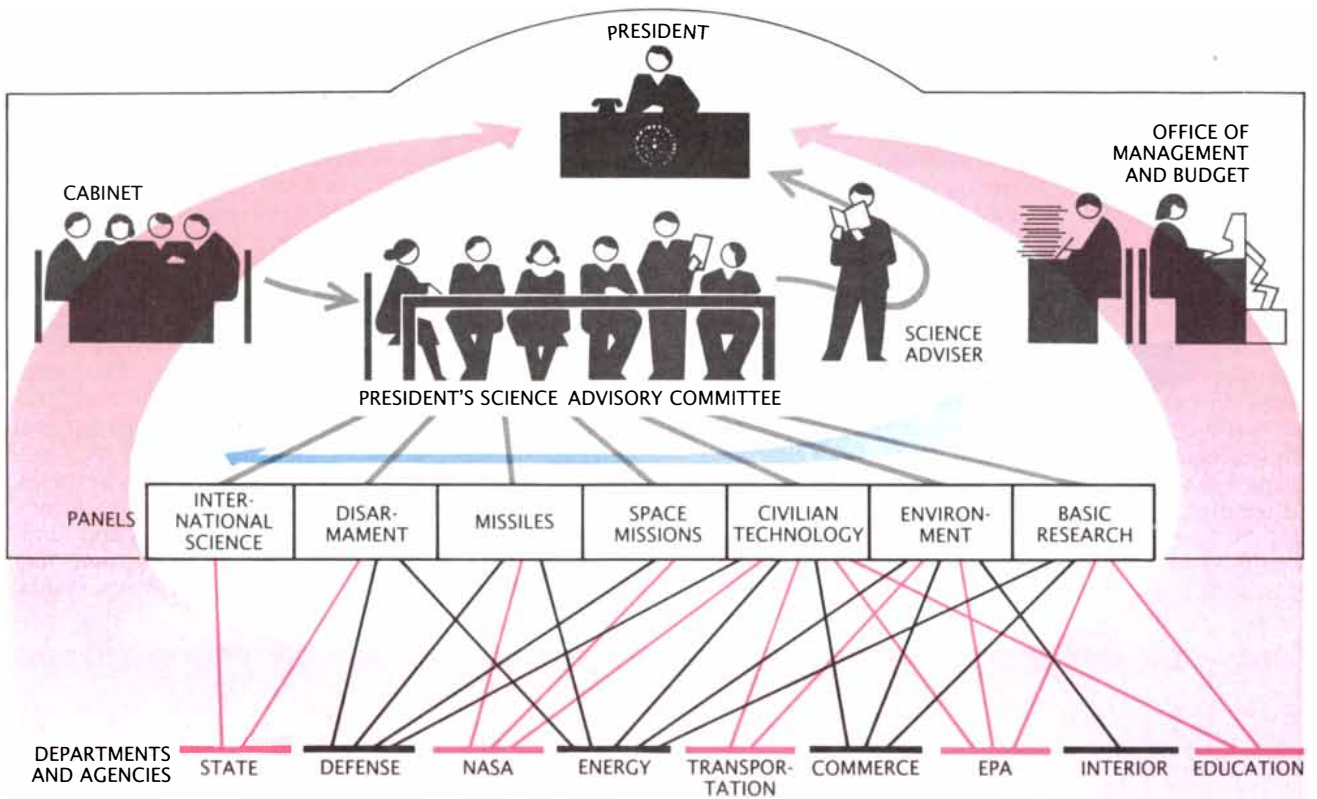
technology because it threatens some political or social status quo. Wider use of nuclear propulsion in ships, for example, might reduce the number of aircraft carriers the Navy must have, in turn reducing the number of battle fleets—and the number of admirals to command the fleets.

The president's dilemma was summarized succinctly by Louis W. Koenig in his book *The Chief Executive*: "Unlike the typical business chief, the President finds no designation in his fundamental charter, the Constitution, as administrative chief." Indeed, "the White House is far from being the command center of the executive branch." Government agencies, although headed by the president's appointees, are authorized and funded by Congress, and programs "are carried out by tenured civil servants,



PRESIDENT DWIGHT D. EISENHOWER shares a light moment with his former science adviser, James R. Killian, Jr., during a meeting with the President's Science Advisory Committee near the end of Eisenhower's term. The committee, created in 1957 after the Soviet Sputnik launch, was notable for its outstanding members. Shown here with Eisenhower and Killian are the author Jerome B. Wiesner (1), science adviser and committee chairman George B. Kistiakowski (2), Harvey Brooks (3), Alvin M. Weinberg (4), Glenn T. Seaborg (5), David Z. Beckler (6), Emanuel R. Piore (7), Wolfgang K. H. Panofsky (8) and I.I. Rabi (9).





PRESIDENT'S SCIENCE ADVISORY COMMITTEE, modeled on the one established in 1957, should operate within the Executive Office and be represented by a science adviser who reports directly to the president. The president would thus have access to technical advice from both the Cabinet and the committee.

With the help of the Office of Management and Budget, the committee would investigate Government programs by appointing panels to study specific topics, which might encompass research programs in several departments. The panel's findings would then be passed on to the committee.

who were on the job before the incumbent president arrived and will remain after he leaves.... the president lacks knowledge of intricate workings of the vast bureaucracy.... Consequently presidents must devote a major share of their term to learning how the executive branch works and why it so often fails to work."

How, then, do presidents cope? The process, as I observed it over five administrations, was remarkably informal, and it varied with the issues of the day and the style of the president. Indeed, I came to think of the president and his staff as the managers of a family business that had grown too big to be operated in the original way but that could not be modernized because it involved a political balance that was understood and accepted by the family. In business, adversity forces modernization. That has not happened, however, in the management of the country.

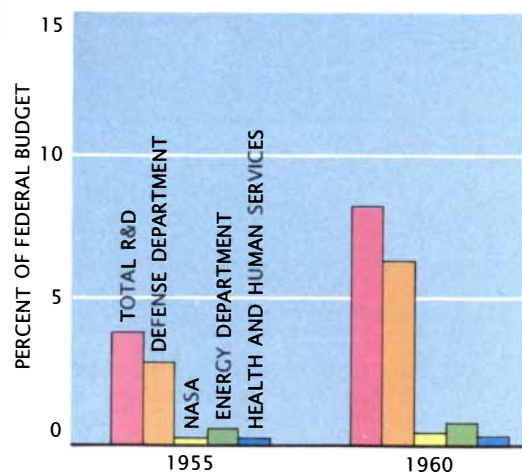
Yet the president must find a way to integrate all aspects of government, set priorities and allocate the chronically inadequate resources. Ideally he should do this all by himself, work-

ing with firsthand knowledge of both broad topics and specific issues. Perhaps Thomas Jefferson could have done it. Today this level of executive involvement is impossible, but this description of the ideal elucidates the role of the president's advisers: to act as his eyes and ears, gathering information, analyzing it and figuring out what his options are. In other words, an effective adviser should be the president's surrogate or sometimes even his alter ego.

In the Eisenhower and Kennedy administrations this goal was achieved to a high degree. Kennedy had comfortable relations with his assistants, which enabled them to serve as alter egos without interfering in his relations with cabinet officers. Ordinarily the assistants would keep the president, as well as one another, informed in a general way. When a major problem occurred, the appropriate assistants would assemble automatically and consolidate crucial information. Only in such critical moments did Kennedy attempt to master the intricacies details of an issue.

An illuminating aspect of Kennedy's style is that he did not confine himself

to the views of those who shared his way of thinking. When an issue became important to him, he always augmented the viewpoints of his assistants by actively seeking outside opinions. For example, when he was trying to decide whether the U.S. should re-



FEDERAL FUNDING for research and development, as a percentage of the Federal budget, reflects general trends and shifting priorities in the nation's science policy. Shown here are R&D budgets for

sume nuclear tests after the Soviet Union did so in 1962, he asked me whether he had talked to everyone who had a strong view on the subject. I made an effort to find representatives of every viewpoint. As he approached a decision, he continued to ask me if he had talked to everyone he should, and I continued to say no, because he had not been willing to see Edward Teller. Finally he agreed to talk with Teller. He could then honestly say he had heard all sides of the issue.

Since World War II, science and technology have presented a singular challenge for the president. Not only do they call for special technical expertise but also they propel rapid changes that create both new opportunities and new problems. Revolutionary technologies and the Cold War combined to drive an arms race centered both on the development of nuclear weapons and on a host of innovative weapons-delivery systems and technologies for command and control. The pace of change outstripped the ability to gauge the performance of new systems, develop effective deployment strategies and assign responsibility to the various branches of the armed forces.

In the immediate postwar years the situation got so bad that the secretary of defense could no longer manage. Military officers lacked the technical background to help the White House reshape military programs. The service branches competed for roles in deploying missiles and radar systems. With every new technical development the services tried to get the scientists and engineers who had worked on it to

help them understand the technology, shape policy and win control over the new system, but that only added to the confusion.

Seeking to establish order, President Truman asked William T. Golden, a prominent financier who had wartime experience with the Navy's technical programs, to study the situation and suggest how he, the president, could better manage the Federal science establishment. Golden recommended appointing a presidential science adviser and establishing a science advisory committee that would report directly to the president. The plan was opposed by many agency heads, however, and the new committee was placed under the Director of the Office of Defense Mobilization. The compromise weakened the effectiveness of the committee and the adviser.

The watered-down committee continued to operate under Eisenhower's Administration. Then came Sputnik, which prompted Eisenhower to appoint Killian as science adviser and to authorize the full mobilization of the President's Science Advisory Committee. Killian also served as chairman of the committee, and so he alone reported directly to the president.

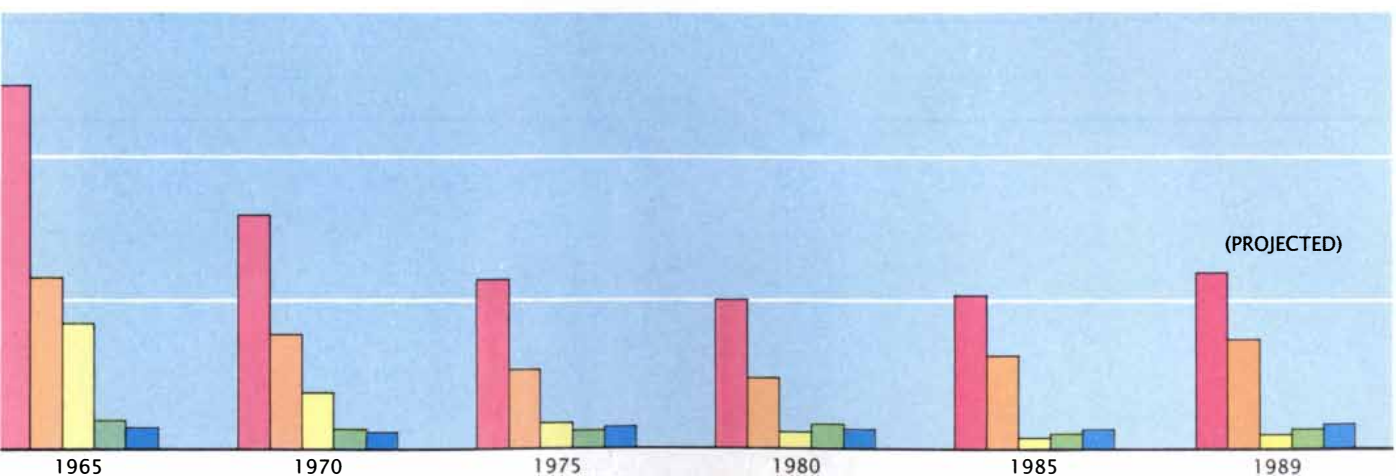
Killian's first task was to identify and evaluate the many problems in the U.S. space program exposed by the Sputnik incident. Killian, his assistants and his colleagues on the committee, particularly Herbert F. York and Edward M. Purcell, then created a single new agency—the future National Aeronautics and Space Administration—out of the disparate programs. They saw to it that the new agency's goals were better defined, and they tried to

ensure that the agency had adequate funding and administrative support. Incidentally, it was Eisenhower's expressed wish to have a civilian space agency that was quite separate from the military space program.

Once the reorganization of the space program was under way, Killian turned the committee's attention to the many other weaknesses in the Government's science and technology programs, particularly in the area of defense. Then, as now, the educational system appeared inadequate for the scientific and technological challenges at hand, and a massive effort was begun to upgrade the quality of science teaching in the nation's schools.

As the science advisory apparatus evolved, it grew into a sophisticated organization that suited the president's many needs. The complex and subtle nature of the system is not widely understood even among individuals who were closely associated with it. Under Killian the Science Advisory Committee developed into a group of about 15 outstanding scientists and engineers, including York, Purcell, I. I. Rabi, James B. Fisk, Edwin H. Land and George B. Kistiakowski (who became science adviser in 1959). Membership on the committee was never, as far as I know, dictated by the political views of the individuals. In addition to the members there was a full-time professional staff that eventually numbered about 30 people.

The committee established permanent panels focused on major science-related concerns such as space, naval warfare, ground warfare, disarmament, environment, basic research,



the four departments and agencies that have the largest technical programs. Funding for defense and space research leaped after the 1957 Sputnik scare. The Apollo lunar missions swelled the National Aeronautics and Space Administration's budget in the mid-1960's. In the late 1960's R&D funding de-

clined with respect to the Federal budget because of growth in other Federal programs. Support for energy research increased after the energy crisis of the mid-1970's. The rise in R&D in the 1980's was driven by large increases in defense research. The data were provided by the National Science Foundation.

medical research and education. The panels were the backbone of the operation. In addition, special panels were created to deal with specific issues. For example, a panel was convened to examine the questions about pesticide pollution raised by the publication of Rachel Carson's *Silent Spring*. The furor that greeted the book illustrates why the president needs independent advice. Pesticide manufacturers and Department of Agriculture staff denounced the book as being entirely inaccurate. Only after the special panel looked into the matter did it become clear that Carson was right.

The committee met as a group once a month, at which time some of the panels would report on their work and the entire committee would discuss it. In this way all the members got a comprehensive view of the nation's technical programs. The system enabled the members to discover problematic trends. For example, in Killian's day the committee saw diverse instances of poor materials technology: steel that was weak and easily corroded, semiconductors that were not pure enough for advanced applications and alloys that could not be used for missile nose cones. The committee responded by calling for stronger Government sponsorship of materials research in the nation's universities.

The membership of the committee and its panels changed slowly. Indeed, a majority of the members who

served under Eisenhower continued to serve under Kennedy, Johnson and Nixon. Turnover on the committee was considerably slower than turnover in the technical staffs of the agencies the committee monitored. Hence committee members often came to be more knowledgeable about individual programs than the agency people in charge of the programs. They also understood how one program related to another better than most program directors did.

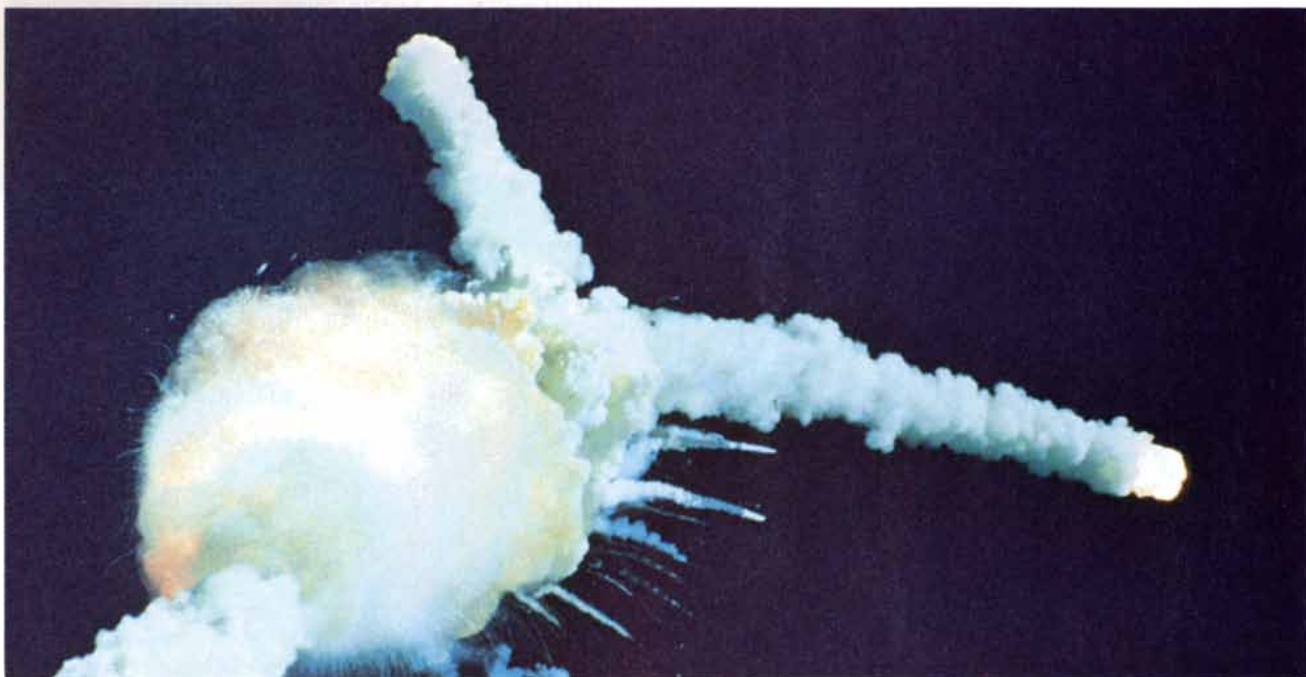
The committee members and their staff had another role, one that is rarely recognized. They served as ombudsmen to whom agency staffers could appeal when they saw serious problems in their agencies and despaired of getting an adequate hearing through normal channels. Many major problems were uncovered in that way. During my tenure the committee heard of problems and delays in a reconnaissance-satellite program. The committee saw that the program was overly ambitious, and so they recommended it be scaled down to a more manageable size. In other cases the committee became aware of redundancy among weapon systems before they were procured and therefore was able to winnow out systems whose performance was not sufficiently improved to justify the cost.

We were helped by the Office of the Budget (now the Office of Management and Budget), which had a much larger

staff capable of monitoring all Government programs in detail. If staff members suspected a problem in some technical program, they would come to us. In recent administrations the science adviser was not part of a well-integrated system, but during Killian's tenure the president made it clear that all technical programs were to be followed by the science advisory committee. Killian had to fight many political battles to get the agency heads to accept this. By the time I came along the system was working smoothly.

If such a committee had existed in more recent times, I believe the *Challenger* shuttle disaster would not have occurred. A committee panel on space programs would have worked continuously with the program staff, attended the launches and studied emerging problems connected with the booster rockets. People within NASA who had complained about safety problems in vain to supervisors would have had an alternative venue in which these problems could be thoroughly investigated. Such a panel would at least have had a chance to catch problems before they caused a catastrophe.

Similarly, a Science Advisory Committee inquiry into the early signs of trouble at the Savannah River plant might well have exposed difficulties there and at other bomb-materials plants and thus prevented the present national crisis that is forcing the plants to be shut down. It might also



CHALLENGER EXPLOSION in January, 1986, might have been prevented if there had existed an effective system for monitoring Federal science programs. Continuous, in-depth scrutiny of

the space-shuttle program by independent scientists and engineers might have made it possible to correct problems with the shuttle's booster rockets before they led to a catastrophe.

have prevented the decades-long exposure to radioactivity of the people living near the plants. (The Savannah River plant is the sole supplier of tritium for the nation's nuclear weapons.)

The best scientists in the nation worked on the President's Science Advisory Committee and its panels. They did this out of loyalty to the president and to the country. So good was the spirit and so great the dedication of the group that its members always put the committee's needs ahead of private responsibilities. They were also committed to confidentiality. No breach of security ever occurred during the 15 years of the committee's existence. It was truly the president's alter ego for science.

Times have changed, but I believe such a science advisory committee, directed by a presidential science adviser, is what the country needs again. Not everyone would agree with me. Although science-policy experts generally acknowledge that the arrangement worked well in its time, many of them say it cannot be re-created.

Some people object to the committee's "elitism." The Science Advisory Committee, however, was elitist only in the sense that it consisted of the nation's top scientists. Should the president have less? Other observers say that laws requiring Government committees to hold open meetings will make it impossible for a contemporary science advisory committee to function effectively. I do not think this would happen. National security matters are discussed in closed meetings anyway. As for other matters, open meetings may require more staff work, but there is no reason they cannot be tried. The Congressional Office of Technology Assessment has been working quite effectively under similar conditions.

There are also those who say it was the special relation between the president and the science adviser that enabled the system to work well. It is true that I was close to Kennedy before he was elected president, but neither Killian nor Kistiakowsky was close to Eisenhower. They just happened to be the right people for him. The important fact was that both presidents understood the value of thorough, honest scientific information. If President-elect Bush wants such help, it will be possible for him to have it.

Granted, some things have changed that will make it more difficult to assemble an effective Science Advisory Committee. Eisenhower and Kennedy could draw on a cadre of scientists

and engineers who had long experience in Government service during World War II and the Cold War. They could provide detailed knowledge on the security-related technical issues that were most important to the president. The wartime generation has been replaced by a new generation of scientists and engineers, and although many of them have government experience, I do not think it has been as intense or as broad-based.

In addition, the problems of today are more difficult than those of 40 years ago. Peace is harder to wage than war. There are no clear-cut ways to reconcile economic growth with the measures needed to curb environmental degradation, stretch dwindling natural resources and solve health and economic problems. The answers will require not only new science and technology but also a mobilization of talent and resources on an unprecedented scale.

Some people propose that what is needed is a new Department of Science, comparable in scale and complexity to the Defense Department. I agree that some kind of new agency should oversee civilian science, but it would be a serious mistake to sweep all Government science projects into a single massive bureaucracy. Individual agencies should continue to manage their own scientific activities, and basic research should continue to be supported by the National Institutes of Health, the National Science Foundation and other nationwide funding agencies.

Given the obstacles, how should President-elect Bush go about establishing a new apparatus for managing science? It cannot be created quickly but rather will have to evolve with time. The president-elect must first select an outstanding science adviser. The individual's ability, judgment and personality will determine the level of support he or she receives from the scientific community, from colleagues in Government and from the president himself. There is much discussion about whether this individual should come from industry or academia. It has been my experience that university people tend to be more thorough because their mission in life is to achieve understanding—but not necessarily to get things done on schedule. Perhaps the science adviser should be someone who can balance both needs, upholding scientific standards while being responsive to the president.

The first act of the science adviser

should be to help resurrect the President's Science Advisory Committee. Members should include the nation's best scientists and engineers. Because few younger scientists and engineers have been deeply involved in government, perhaps the committee should include a few veterans of previous committees along with the best younger scientists, particularly those who have worked with the Defense Department and other agencies in the Government. The committee will need considerable flexibility to face the diversity of tasks.

The new science adviser and committee should begin with many modest projects to explore the problems that must be addressed. Paramount among these are rescuing the planet and creating technologies for both economic growth and coexistence with the environment and with other nations. In order to achieve these goals it will be necessary to develop an entirely new outlook, together with an entirely new infrastructure—new energy sources, civilian industries, modes of transportation and resource management—and an enhanced educational system.

What is more, the president must lead nations and industries from the arena of competition to that of cooperation. Only an international effort will be able to cope with ozone depletion, the greenhouse effect, contaminated oceans and runaway population growth. Surely these problems are the moral equivalent of war; they are a threat to civilization's survival as great as any posed by Hitler, Stalin or the atom bomb. President-elect Bush's choice of priorities will have repercussions well into the next century.

FURTHER READING

A SCIENTIST AT THE WHITE HOUSE. George B. Kistiakowsky. Harvard University Press, 1976.

SPUTNIK, SCIENTISTS, AND EISENHOWER: A MEMOIR OF THE FIRST SPECIAL ASSISTANT TO THE PRESIDENT FOR SCIENCE AND TECHNOLOGY. James R. Killian, Jr. The MIT Press, 1977.

SCIENCE AND TECHNOLOGY ADVICE FOR THE PRESIDENT. Ted Greenwood in *The Presidency and Science Advising*, edited by Kenneth W. Thompson. University Press of America, 1986.

PRESIDENT'S SCIENCE ADVISORY COMMITTEE REVISITED. Edited by William T. Golden in *Science, Technology & Human Values*, Vol. 11, Issue 2, No. 55, pages 5-28; Spring, 1986.

SCIENCE AND TECHNOLOGY ADVICE TO THE PRESIDENT, CONGRESS, AND JUDICIARY. Edited by William T. Golden. Pergamon Press, 1988.



FRUIT-FLY LARVA is stained blue in the anterior midgut (*top*) and the fat body (*bottom*), indicating the activity of an enzyme involved in sugar metabolism. The gene for the enzyme was

activated, or turned on, by a regulatory protein that is ordinarily found in yeast cells. The yeast activator can work in plant and mammalian cells as well as in the cells of insects.

How Gene Activators Work

Much is known about how genes are turned on and off in bacterial cells. Now molecular biologists show that what they have learned is relevant to gene regulation in higher organisms as well

by Mark Ptashne

Every human cell contains about 100,000 genes, but at any given time only a fraction of those genes are working. Many are expressed selectively, during certain stages of development, for example, or in response to environmental signals. How are genes turned on and off? During the past two decades my colleagues and I at Harvard University and elsewhere have studied the switching mechanism that operates in a bacterial virus called bacteriophage lambda. We now find that the ideas we conceived in our studies of that simple case take us quite far indeed in analyzing gene regulation in the cells of humans and other higher forms of life.

To understand the problem recall that a typical gene, a segment of DNA, encodes a protein. The specific sequence of the base pairs that make up that segment of DNA corresponds to the specific sequence of the amino acids that make up the protein. Decoding a gene requires first that it be transcribed into a molecule of messenger RNA (mRNA) and then that the mRNA molecule be translated into protein.

Gene expression can be controlled by regulatory proteins that bind to specific sites on DNA. These proteins are called activators or repressors depending on whether they increase or decrease transcription; some regulators can perform both functions. How do regulatory proteins see specific sites on DNA and turn the genes they control on or off?

Part of the answer involves enzymes called RNA polymerases, which are responsible for transcribing genes into RNA. For the genes I am concerned with here, an RNA polymerase attaches to DNA near the beginning of a gene and then moves along it, transcribing the sequence of DNA into a molecule of mRNA. Now the question can be refined: How does an activator help RNA polymerase to transcribe a gene?

How does a repressor hinder transcription by the enzyme?

In our studies of phage lambda we have learned a great deal about the interactions among regulators, DNA and RNA polymerase that account for gene regulation in that simple organism. But there were many reasons to suspect our findings would not be relevant to higher organisms. For one thing, RNA polymerases exist in forms that at first glance seem quite different in higher and lower organisms. And whereas the regulators we studied in lambda bind very close to the gene they control, in other instances regulator binding sites are found on the DNA at very great distances from the gene—hundreds or even thousands of base pairs away. Moreover, in eukaryotes (higher organisms), but not in prokaryotes such as bacteria, DNA is sequestered in a cell nucleus and is wrapped around proteins called histones much like thread on a spool. Do fundamentally different mechanisms obtain in these cases?

The developments of the past two years suggest that a few simple principles may be common to gene regulation in these evidently disparate situations, whether in bacteria, yeast, plants, fruit flies or humans. I shall state these ideas as they have been developed by the study of phage lambda, and I shall then show how my colleagues and I apply them to the study of gene regulation in higher organisms. I shall consider in particular the problem of gene activation.

Phage lambda caught the attention of scientists more than 30 years ago, when they realized that the virus's life cycle is a dramatic manifestation of gene regulation. When the viral DNA is injected into the bacterium, it follows one of two paths: either most of the viral genes are expressed and the virus replicates and lyses, or bursts, the host cell, or al-

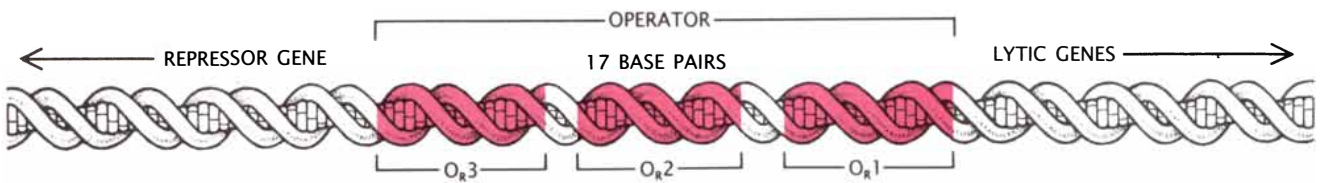
most all the viral genes are turned off and the inert viral DNA becomes a passive part of the bacterial chromosome. The silent genes are switched on, however, when the bacterium encounters a carcinogen such as ultraviolet light.

A regulatory protein encoded by the virus plays a key role in controlling expression of the viral genes. This protein, known as lambda repressor, binds to a few specific sites on the viral chromosome and turns off most of the viral genes. Carcinogens induce the expression of the genes for viral replication and cell lysis (the "lytic" genes) by destroying repressor. But in addition to being a repressor of gene expression, as its name implies, the protein also *activates* transcription: it greatly increases the rate of transcription of its own gene.

Knowledge of the structure of repressor has helped us to understand how it works. The fundamental unit, or monomer, of repressor is folded into two blobs (called domains) of roughly equal size. Two monomers associate to make a dimer, the form of repressor that binds to DNA. Two dimers bind to adjacent sites on lambda's DNA called O_R1 and O_R2 ; there is an additional site called O_R3 nearby, but it is not relevant to this discussion [see illustrations on next page].

The presence of these two DNA-bound dimers affects transcription in two ways, one negative and one positive. First, the repressors prevent RNA polymerase from binding to the DNA and copying the lytic genes (which by convention are said to lie to the right). Second, one of the repressors helps RNA polymerase to bind and begin

MARK PTASHNE is professor of biochemistry and molecular biology at Harvard University. He has written three previous articles on gene regulation for SCIENTIFIC AMERICAN, the first in 1970 and the most recent in 1982.



LAMBDA OPERATOR contains the sites at which regulatory proteins bind in order to control the genes of a bacterial virus known as phage lambda. In this diagram the gene for a regulatory protein called repressor lies to the left of the operator; to

the operator's right lie the so-called lytic genes that mediate replication of the virus and the lysis, or bursting, of the host bacterium. Within the operator there are three binding sites called O_R1 , O_R2 and O_R3 , each made up of 17 base pairs of DNA.

transcription of the gene that encodes repressor itself (which is said to lie to the left of the bound repressors).

The sites to which repressor binds are called operators and the ones to which RNA polymerase binds are called promoters. The promoter to the right of the lambda operator is strong and the one to the left is weak. The operator sites to which repressor binds overlap the strong rightward promoter but lie adjacent to the weaker leftward one. In the absence of repressor, RNA polymerase attaches to the strong rightward promoter and begins transcription. In contrast, only if it is aided by the adjacent bound repressor (in its role as an activator) does polymerase attach to the weak leftward promoter and transcribe the repressor gene. A small difference in the sequences of the two promoters accounts for the fact that one promoter attracts polymerase only weakly and so requires an activator to function, whereas the other functions well as long as polymerase has access to it.

There is a good biological reason for the complexity of this picture—the two repressor dimers binding simultaneously and the dual positive-negative effect of repressor. These and other factors enable the virus to switch efficiently from one mode of growth to

another depending on the environment [see "A Genetic Switch in a Bacterial Virus," by Mark Ptashne, Alexander D. Johnson and Carl O. Pabo; SCIENTIFIC AMERICAN, November, 1982]. But here I return to our original, narrower question: Precisely how does repressor bind to its specific operator sites and turn on transcription of its own gene?

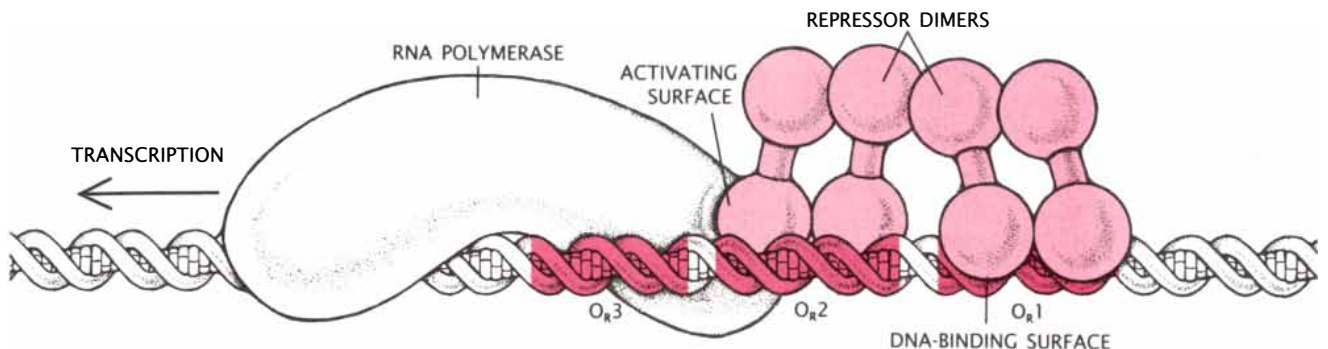
X-ray crystallography and a large number of biochemical experiments have provided a rather detailed view of how the lambda repressor and related regulatory proteins bind selectively to their operator sites on DNA. The DNA-binding surface of each repressor monomer bears a protruding structure called an alpha helix. Alpha helices are found in many proteins, and many different sequences of amino acids fold into alpha helices.

The alpha helix protruding from each monomer fits neatly into the major groove that spirals down the DNA double helix. Chemical groups along the outer surface of the alpha helix form a pattern that is determined by the sequences of amino acids along the alpha helix. The chemical groups on the base pairs that are exposed in the major groove also form a pattern

determined by the base-pair sequence. Only when the two patterns match will repressor bind [see top illustration on opposite page].

The fact that repressor binds as a dimer means that each of two identical alpha helices (called, in this context, recognition helices) must find the appropriate match in the DNA sequence. This requirement explains why each operator site contains two identical or nearly identical half-sites, each of which is recognized by one of the alpha helices. The repressor dimer is twofold symmetrical, and so is its operator site.

Lambda repressor can pick out a specific base-pair sequence from a vast sea of DNA, and when it binds, it does not significantly distort the shape of the DNA double helix. Many other proteins in both prokaryotes and eukaryotes recognize DNA by a similar mechanism; their recognition helices differ in sequence from that of lambda repressor, as do the sequences of the operators to which they bind. They too can identify a binding site that is typically less than 20 base pairs long among millions of unrelated sequences. (Incidentally, each of these proteins, repressor included, bears a second protruding alpha helix adjacent to the recognition helix, the pair



REPRESSOR TURNS ON ITS OWN GENE by helping an enzyme called RNA polymerase to bind to the repressor gene. RNA polymerase is the enzyme that transcribes genes into RNA, the first step in the process by which a gene is expressed as a protein. Repressor "dimers" bind to both O_R1 and O_R2 . (A dimer is a molecule made up of two identical subunits.) One part

of the protein, the "activating surface," is thought to touch the enzyme. The dimers also touch one another to facilitate their own binding, a phenomenon referred to as cooperativity. Moreover, the bound dimers block transcription of the lytic genes to the right by preventing the polymerase from binding; that is how the protein earned the name "repressor."

forming a characteristic "bihelical" motif. This second alpha helix lies across the major groove of DNA, helping to orient the recognition helix in the major groove.)

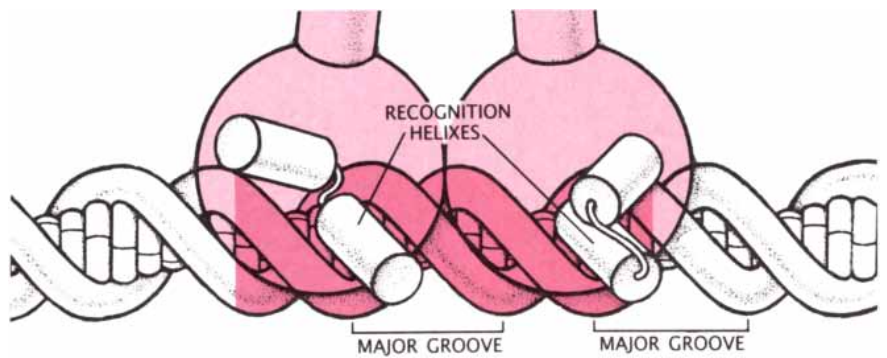
That picture of repressor's interaction with its DNA target has been influential in our thinking about how gene activators may work. It might have been imagined, for example, that when repressor binds it forces the DNA strands apart, causing them to refold in some unusual structure. It is conceivable that such an altered DNA structure could somehow be responsible for activating a gene. But the fact that DNA does not greatly change its shape when lambda repressor binds suggests that DNA binding per se is not what activates transcription.

How then does the bound repressor activate transcription? Part of repressor, distinct from the surface by which it binds to DNA, interacts with RNA polymerase to turn on transcription. This so-called activating surface evidently touches RNA polymerase and helps the enzyme to bind and initiate transcription of the repressor gene. Although little is known about the structure of RNA polymerase, we do know something about repressor's activating surface.

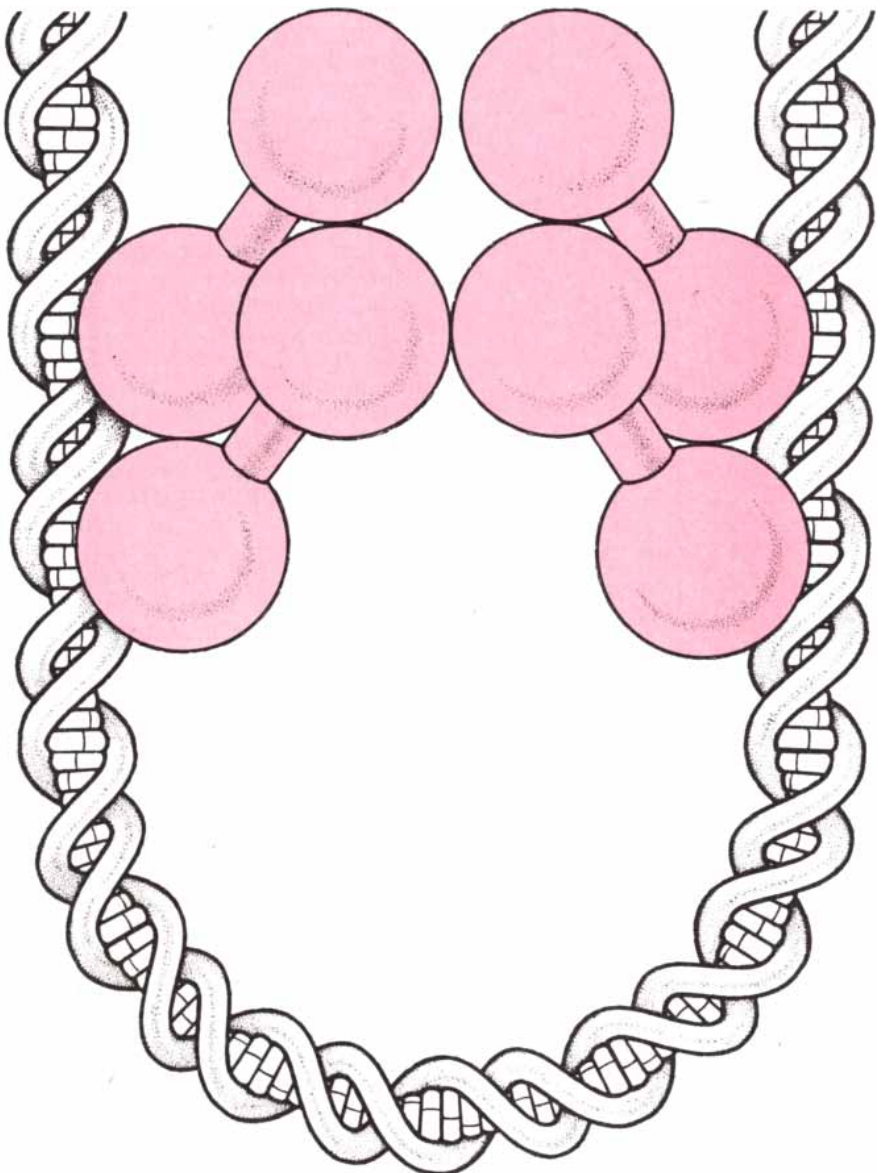
We think an important component of lambda repressor's activating surface is again an alpha helix. Our conclusion follows from analysis of a special class of repressor mutants that bind DNA normally but fail to activate transcription. Such mutants bear changes in or near an alpha helix we might call the activation helix. Inspection of a model of repressor bound next to RNA polymerase reveals that the activation helix is perfectly positioned to touch the polymerase.

The amino acid substitutions in the mutant repressors that bind but fail to activate are of a particular type: they all decrease the amount of negative charge in or near the activation helix. Most of the 20 amino acids that make up proteins are electrically neutral, and in most proteins the few positively or negatively charged amino acids are present in roughly equal numbers. The data from our lambda mutants suggest that negatively charged amino acids are particularly important for activation, an idea that will recur in experiments I discuss below.

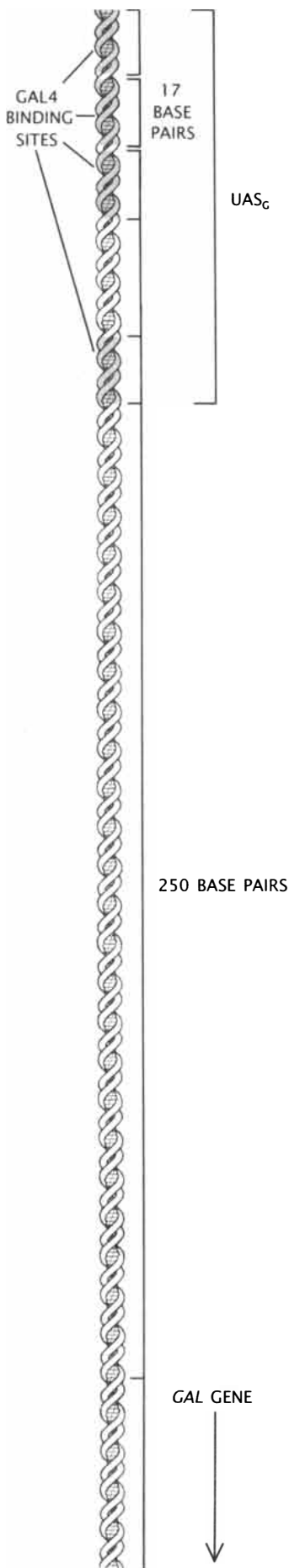
There is one other facet of repressor's mode of action that turns out to have broad relevance. It has to do with the binding of the two repressor dimers required for activation. The di-



DNA-BINDING SURFACE of repressor interacts with DNA by means of protruding alpha helices, structures formed by helical chains of amino acids. Here the helices are indicated by barrels. The "recognition helix" fits neatly into the major groove that runs along the DNA helix like the thread of a screw. The other alpha helix lies across the major groove. The pattern of amino acids in the recognition helices matches the pattern of the base pairs in the operator site to which repressor binds.



COOPERATIVITY AT A DISTANCE occurs when the binding sites for repressor are moved apart. The intervening DNA loops out so that the dimers can touch each other and thereby help each other to bind. The author thinks DNA looping could also facilitate interactions between proteins involved in gene regulation in higher organisms.



mers do not bind independently; rather, the binding of one helps the other one to bind. In order to understand the effect, imagine that a repressor dimer binds first to the rightmost site (O_{R1}), the one known to be the strongest binding site. That bound repressor then helps another dimer to bind to the adjacent site by touching it. The dimers are said to bind to DNA cooperatively.

It is now clear the principle of cooperative binding is widely used by proteins that bind DNA. A complete discussion of the phenomenon is beyond the scope of this article; briefly, we think cooperativity makes the binding of proteins extremely sensitive to small changes in their concentration, thereby enabling genes to switch on and off very efficiently. Cooperativity also helps a protein to distinguish its proper DNA-binding site from among the vast amount of irrelevant DNA. In the case of the lambda repressor the cooperative effect is relatively small: the presence of one bound dimer increases the binding of another by a factor of about 10. But even this effect is critical to repressor action.

One more aspect of cooperativity in lambda has influenced our thinking about how regulatory proteins might affect transcription even when they are bound at a great distance from the genes they control. Just as one repressor can help another to bind when the operator sites are adjacent, so repressor dimers can help each other even if the sites are moved apart along the DNA. The dimers can touch each other because DNA is flexible and loops out to allow the interaction [see bottom illustration on preceding page]. The idea of looping was proposed by Sankar L. Adhya and his co-workers at the National Cancer Institute and by Robert F. Schleif and his colleagues at Brandeis University.

Cooperativity at a distance has been demonstrated or strongly inferred in many instances for prokaryotic regulatory proteins. Indeed, lambda is the exception in that its operator sites, in their ordinary configuration, lie ad-

YEAST GENE is controlled by an upstream activating sequence (UAS_c) located some 250 base pairs away. The gene is called the *GAL* gene because it encodes an enzyme that degrades the sugar galactose; the regulatory protein that activates its transcription is known as GAL4. Each of the four GAL4 binding sites contained in UAS_c is, like a repressor binding site, 17 base pairs long. GAL4 works even when its binding sites are moved 750 base pairs away from the *GAL* gene.

jacent on its DNA. It seems reasonable to assume that the same looping mechanism hypothesized to bring regulatory proteins together can also enable regulatory proteins bound at one site to influence transcription of a distant gene. For example, a DNA-bound activator might work by touching RNA polymerase itself or some accessory protein bound at the start of the gene, the intervening DNA looping out to accommodate the reaction.

How might the principles describing the action of lambda repressor relate to problems of gene regulation in higher organisms? The question has been rather difficult to answer because eukaryotes are harder to manipulate than bacteria. Although the development of the technologies commonly referred to as "recombinant DNA" has made it possible to do new kinds of experiments with the cells of higher organisms, compared with bacterial cells the limitations remain daunting. We chose to study genes in yeast, one of nature's simplest eukaryotes. Yeast grows almost as quickly as bacteria, and it is almost as amenable to genetic manipulation.

Our experiments focus on a group of genes encoding enzymes that degrade the sugar galactose. The *GAL* genes, as they are called, are activated by the protein GAL4. GAL4 binds to a DNA sequence some 250 base pairs away from the beginning of the nearest *GAL* gene. The sequence is called the galactose upstream activating sequence, or UAS_c , the word "upstream" emphasizing the fact that it mediates GAL4 activity at a distance. Indeed, the UAS_c can be positioned in front of and at various distances from other yeast genes, and GAL4 will then activate transcription of these genes as well.

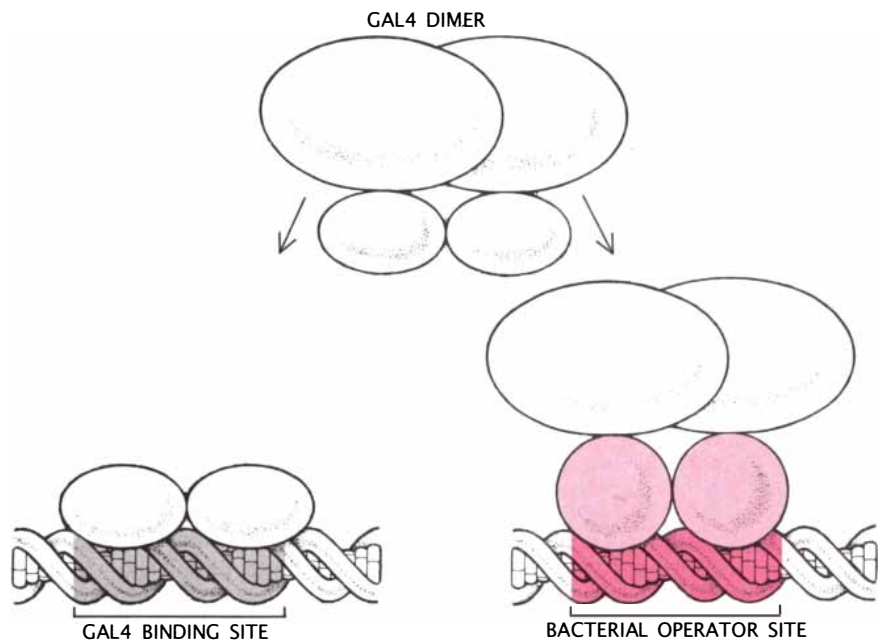
A closer look at UAS_c reveals that it contains four similar sequences that bind GAL4, each 17 base pairs long. Each sequence is, like each lambda operator site, twofold symmetrical (or nearly so), an observation consistent with other experiments suggesting that GAL4, like lambda repressor, binds to each site as a dimer. A GAL4 dimer bound to a single site will activate transcription, but not as efficiently as multiple dimers on multiple sites do.

How does GAL4 work? That is, how does GAL4 bind selectively to UAS_c and then activate transcription of a gene several hundred base pairs away? The following experiments show that, like lambda repressor, GAL4 has a DNA-binding surface and an activating surface. They also show that, unlike the

two surfaces in lambda repressor (which are close to each other on one domain), GAL4's DNA-binding and activating regions are found on different parts of this much larger protein and are easily separated.

The experiment that located the DNA-binding region of GAL4 was done in part by fragmenting the gene that encodes GAL4, putting the fragments back into yeast and then determining which functions, if any, are carried out by the protein fragments produced by the fragments of genes. We found one fragment, the beginning part of the protein, that binds to DNA but fails to activate gene expression. From this experiment we surmised that the activating region of GAL4 must lie in some part of the protein other than its first 100 amino acids.

The experiment that clinched this supposition involved fusing DNA fragments from the GAL4 gene to DNA fragments encoding parts of other proteins, thereby producing protein hybrids. In the key experiment we replaced GAL4's DNA-binding region with the DNA-binding region from a bacterial repressor. Predictably enough, the hybrid protein has no effect on gene expression in ordinary yeast because it does not have a means of recognizing the UAS_G. If, however, an operator sequence known to be recognized by the bacterial repressor is placed in front of a yeast gene, the hybrid protein activates transcription of that gene. The native bacterial repressor cannot activate gene expression in yeast. In other words, to make an activator we need two functions: an activating surface (in this case provided by the GAL4 fragment) and a DNA-binding surface (in



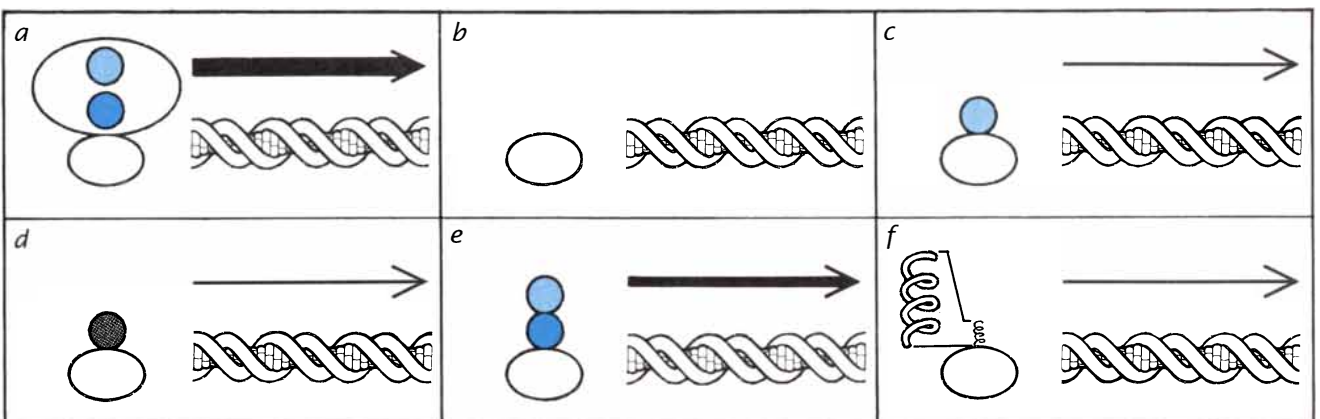
MANIPULATION OF GAL4 shows that its DNA-binding and activating surfaces are in different parts of the protein. Native GAL4 has at least two domains (*top*); a GAL4 mutant protein lacking the large domain can bind to DNA but cannot activate transcription (*left*). That the larger GAL4 domain contains the activating surface is shown by experiments in which GAL4's DNA-binding region is replaced with the DNA-binding region of a bacterial protein related to the lambda repressor (*right*). The hybrid activates a gene when a binding site for the bacterial repressor is inserted into the yeast DNA.

this case provided by the bacterial repressor) to tether the activating fragment to the DNA near the gene.

Incidentally, we do not know precisely how GAL4 recognizes its DNA sites; it does not have a lambda-like bihelical motif. Its sequence suggests that it may bear protruding fingers of amino acids anchored by zinc, structures known as zinc fingers. The experiment I have just described, however, shows that this question, interest-

ing in its own right, is not crucial to our understanding of gene activation.

The bacterial-yeast hybrid experiment prompted us to wonder what the activating surface of GAL4 looks like. The obvious approach to the problem was suggested by the earlier lambda-repressor experiment: isolate mutants of GAL4 that bind DNA but are unable to activate. Assuming that these mutants would be analo-



GAL4 DERIVATIVES helped to identify the parts of the protein involved in activation. In intact GAL4 (*a*) two activating regions (*color*) induce transcription (*arrow*); without these regions transcription does not occur (*b*). Each of the regions can activate transcription independently when it is attached to

GAL4's DNA-binding region (*c, d*). Together they are almost as effective as the intact protein, even though 80 percent of that protein is gone (*e*). A 15-amino-acid chain that presumably forms an alpha helix works nearly as well as one of the activating surfaces when it is attached to the DNA-binding region (*f*).

gous to the repressor mutants, they would bear amino acid changes that would define the activating surface. But, for reasons I shall return to below, our attempts to isolate such mutants produced only protein fragments in which the bulk of the protein was missing. The mutants could not help us to zero in on the surface itself.

The next approach we tried was inspired by Keith R. Yamamoto and his colleagues at the University of California at San Francisco, who were dissecting a gene for a human regulatory protein called the glucocorticoid receptor. We decided to chop up the GAL4 gene further and then attach the part of the gene that encodes the DNA-binding surface to fragments of the remainder. We were trying to isolate a simplified form of GAL4, in which the activating surface of GAL4 was attached directly to its DNA-binding fragment. Actually we found two parts of the protein, each about 100 amino acids long, that can independently activate gene expression when they are attached to a DNA-binding fragment. When both of the fragments are attached, the protein activates gene expression nearly as well as GAL4, even though about 80 percent of the protein has been deleted.

There are two surprising aspects to these results. The first is that the experiment worked at all. Molecular biol-

ogists are used to thinking of proteins as rather precisely defined three-dimensional structures that cannot be tinkered with so easily. It is one thing (or so we thought) to attach parts of one protein to another and expect the hybrid to work; it is quite another to delete large parts of a protein and expect it to retain its activity.

Perhaps we should not have been so surprised. Over the past few years it has become increasingly clear that the domains in proteins can consist of independent functional units. The structure of GAL4 has not yet been determined, but the fact that the activating regions and the DNA-binding function are so readily separated and recombined indicates each is on a separate domain.

The second surprising outcome of our analysis of the GAL4 activating regions comes from comparing their sequences. Very often parts of proteins that carry out similar functions have similar sequences and similar three-dimensional structures. In this case, however, we noted only one common element: the presence of a large excess of negatively charged amino acids. Kevin Struhl and his colleagues at the Harvard Medical School had shown that the activating region of another yeast transcriptional activator called GCN4 bears an excess of negative charge; otherwise its sequence ex-

hibits little obvious relation to either activating region of GAL4.

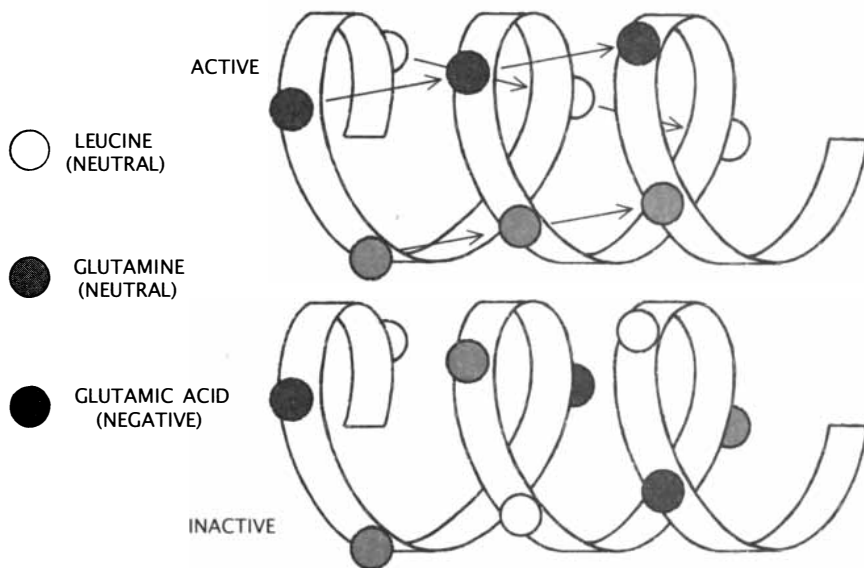
It is remarkably easy to find protein sequences that function as activating sequences when they are attached to a DNA-binding protein fragment. We attached random bits of bacterial DNA to the part of the GAL4 gene that encodes the DNA-binding region. We then introduced these hybrid genes into yeast and examined the yeast to see whether the hybrid proteins could activate a test gene bearing a GAL4 binding site nearby.

A high percentage of the newly generated proteins activated gene expression. Moreover, their sequences have no conspicuous similarities except for the fact that they all bear an excess of negatively charged amino acids. Some of the new activators, whose activating regions are only 50 amino acids long, work nearly as efficiently as the intact GAL4 protein.

The presence of excess negative charge on so many activators suggests that this feature is not coincidental, and an additional genetic experiment reinforces the idea. I mentioned above that attempts to define the GAL4 activating region by isolating mutants specifically deficient in the activation function yielded only fragments of the protein. If there are two activating regions in GAL4 and they can work independently, then in order to lose the activating function a mutant would have to acquire changes in at least two parts of the protein, an event that might be prohibitively rare. Only if most of the protein was lost would both activating regions be absent.

Starting with a simplified version of GAL4, however, mutant proteins can be isolated that activate either more or less efficiently than the native protein. It turns out that there is a good (but not perfect) correlation between charge and activity: mutants with increased activity usually bear amino acid substitutions that increase the negative charge, whereas mutants with reduced activity generally bear substitutions that decrease the negative charge.

A few exceptions suggest, however, that some structural aspect is important in addition to the amount of negative charge the activating surface bears. Inspired by the example of lambda repressor's activating surface, we decided to test whether an alpha helix with negative charges on one of its surfaces could activate. We designed a piece of DNA encoding a short stretch of protein that could in theory fold into such a helix and



EXPERIMENTAL AMINO ACID CHAINS were affixed to the GAL4 DNA-binding region in order to explore the characteristics needed for an activating surface. The same amino acids appear in both chains, albeit in a different order. The amino acids were placed so that if the chains form alpha helices (as is shown here), one chain would have the negatively charged amino acids aligned on one surface (*top*), whereas the negative charges on the other chain would be scattered (*bottom*). The chain with the aligned charges activates transcription but the chain with the scattered charges does not. Hence distribution of charge, not just charge alone, affects the ability to activate.

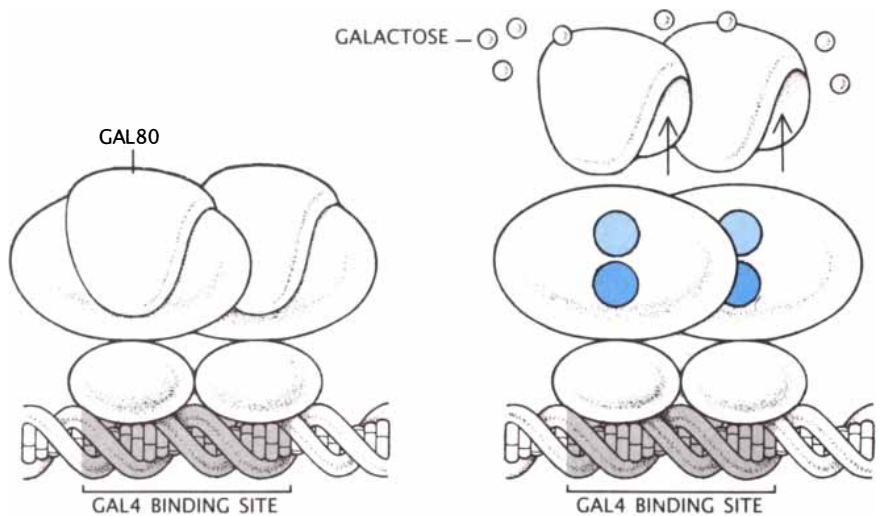
attached it to the DNA encoding the DNA-binding part of GAL4. If the alpha helix does indeed form, it should bear not only a negatively charged surface but also a hydrophobic (oily) surface that might help it pack onto the surface of the GAL4 fragment.

We learned that this hybrid protein, with an activating surface made up of only 15 amino acids, could activate genes in yeast about 20 percent as efficiently as the intact GAL4 protein. If, however, the order of the amino acids was scrambled, the protein fragment did not activate, showing that charge alone does not confer the ability to activate. Perhaps the particular distribution of negative charge along one of the helix's surfaces is indeed required to create an activating region.

The preceding experiments set the stage for a further generalization that many of us did not expect: GAL4 and the various derivative fragments of GAL4 that activate transcription in yeast also do so in the cells of mammals, in cells of the fruit fly *Drosophila* and even in tobacco-plant cells. The experiments in which these facts were demonstrated share a common strategy. In each case the GAL4 gene or one of its derivatives is introduced into a foreign cell, which synthesizes the protein. We then insert a binding site for GAL4 near one of the organism's native genes. In each case—plant, insect and mammal—the yeast activator increases transcription of the native gene to high levels.

We observe no activation if the GAL4 binding site is omitted or if the DNA-binding region of GAL4 is expressed without an activating region. Thus a molecule bearing an activating region attached to a DNA-binding fragment will activate gene expression in many, perhaps even all, eukaryotes.

Given that an activating region retains its function in foreign cells, does it also retain its ability to act over large distances in foreign cells? Remote binding sites called enhancers are known to be prevalent in the genes of higher organisms. Can our activators turn on transcription even when they are bound to DNA more than 1,000 base pairs away from the gene? The answer is yes, provided the activator bears a particularly potent activating region. Such a region is found in a regulatory protein called VP16 that is synthesized by the herpes simplex virus. When the acidic part of the protein is fused to the DNA-binding fragment of GAL4 and the hybrid is deployed in mammalian cells, it activates transcription of a gene that is more



GALACTOSE EXERTS CONTROL over GAL4 by means of a protein called GAL80. When galactose is not present in a yeast cell's environment, GAL80 shields GAL4's activating surfaces, thereby preventing the protein from activating transcription. When galactose is present, the sugar or one of its metabolic derivatives is thought to release GAL80, so that GAL4 is free to activate the genes for enzymes that degrade galactose.

than 1,000 base pairs upstream or downstream from a UAS_G.

We imagine that this segment of VP16 has just the right combination of structure and charge to interact particularly strongly with its target protein. GAL4 itself works only at smaller distances along the DNA, presumably because its activating region interacts somewhat less strongly with the target protein.

My discussion of gene regulation has thus far skirted the issue of how regulatory proteins themselves are controlled in a cell so that they can mediate an appropriate response to environmental conditions. The answer to that question is at least partially known for GAL4. Since the protein turns on the genes that degrade the sugar galactose, it is important that GAL4 work when galactose is present in a yeast cell's environment and not when the sugar is absent. It turns out that an inhibitor protein called GAL80 ordinarily shields the activating surfaces of GAL4; when galactose is present, the sugar or one of its metabolic products releases the inhibitor, exposing the activating surfaces.

The studies and findings I have presented in this article also point to several unsolved problems. First, if we assume that activating regions interact with some other protein involved in transcription, then what is the other protein? It could be RNA polymerase itself, but we suspect, rather, that it is another protein that acts as an intermediary between the regulatory protein

and the polymerase. Evidently this target protein is present in a similar form in many different (and perhaps all) eukaryotes.

Second, what is the nature of the interaction between an activating region and its target? Molecular biologists are used to thinking of protein-protein interactions as being determined by specific structures of the interacting molecules, but here molecules having a variety of negatively charged sequences work efficiently. Finally, it is important to note that the activators we have studied may represent just one class of activating proteins; there could be other classes.

FURTHER READING

- AN OPERATOR AT -280 BASE PAIRS THAT IS REQUIRED FOR REPRESSION OF ARABAD OPERON PROMOTER: ADDITION OF DNA HELICAL TURNS BETWEEN THE OPERATOR AND PROMOTER CYCLICALLY HINDERS REPRESSION. Teresa M. Dunn, Steven Hahn, Sharon Ogden and Robert F. Schleif in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 81, No. 16, pages 5017-5020; August, 1984.
- FUNCTIONAL DISSECTION OF A EUKARYOTIC TRANSCRIPTIONAL ACTIVATOR PROTEIN, GCN4 OF YEAST. Ian A. Hope and Kevin Struhl in *Cell*, Vol. 46, No. 6, pages 885-894; September 12, 1986.
- A GENETIC SWITCH: GENE CONTROL AND PHAGE λ. Mark Ptashne. Cell Press & Blackwell Scientific Publications, 1987.
- HOW EUKARYOTIC TRANSCRIPTIONAL ACTIVATORS WORK. Mark Ptashne in *Nature*, Vol. 335, No. 6192, pages 683-699; October 20, 1988.

Deep Earthquakes

They have posed a fruitful puzzle since their discovery 60 years ago. How can rock fail at the temperatures and pressures that prevail hundreds of kilometers down?

by Cliff Frohlich

In most earthquakes the earth's crust cracks like porcelain. Stress builds up until a fracture forms at a depth of a few kilometers and slip relieves the stress. Some earthquakes, however, take place where the earth cannot fracture. They occur hundreds of kilometers down in the earth's mantle, where high pressure is thought to prevent rock from cracking even at stresses high enough to deform it like putty. How can there be earthquakes at such depths?

These mysterious deep events are common enough. Since 1964 the International Seismological Centre (ISC) in London has catalogued more than 60,000 earthquakes at depths greater than 70 kilometers—22 percent of all the earthquakes located during the period. A few of them have even been destructive. Although almost all catastrophic earthquakes are shallow, occurring at a depth of 50 kilometers or less, a tremor centered at a depth of 100 kilometers devastated Bucharest, Romania, on March 4, 1977.

For the most part, however, deep events have had their greatest impact in geophysics. Their geographic pattern provided evidence for the great unifying theory of modern geophysics, plate tectonics. They have also proved to be good energy sources for

seismic studies of the earth's deep interior, which attempt to decipher structure from the behavior of earthquake waves traveling through the earth. Now deep earthquakes themselves may be yielding their secrets. Seismological observations together with laboratory studies of rock behavior at high pressures have led to plausible accounts of how they may occur.

The depths of all earthquakes had been controversial in the decades before 1927, when the Japanese seismologist Kiyoo Wadati convincingly demonstrated the existence of deep events. Noting that in some cases intense shaking is confined to a small area, certain workers had argued that earthquake sources must lie within a few kilometers of the surface. Other workers arrived at much greater focal depths, down to 1,200 kilometers, when they tried to determine earthquake directions from the deflections their waves produced on the simple seismographs of the day.

The controversy sharpened as more was learned about seismic waves and how they travel. Investigators studying seismograms learned to recognize several kinds of body waves (earthquake waves that travel through the earth, in contrast to surface waves, which follow the surface). First to arrive are *P* waves, or primary waves. Also called compressional waves, they are waves of sound in the earth, propagating as alternating zones of high and low pressure. They are followed by *S* waves—secondary or shear waves—which propagate through the earth as a side-to-side shaking. Comparisons of times at which *P* and *S* waves from a given earthquake arrived at different stations showed that their travel time depends on both distance and the internal structure of the earth.

Given a model of the earth's structure and a set of arrival times recorded at a variety of locations, then, one could in theory determine the location

of the event itself. In 1922 H. H. Turner, who directed the clearinghouse of seismological data that later became the ISC, applied this method in a stimulating and controversial paper. Based on an analysis of data from stations around the world, Turner proposed that earthquakes occur in three depth ranges. "High focus" events have sources near the surface, but normal earthquakes, the most plentiful kind, take place roughly 150 kilometers down. "Deep focus" events have focuses at depths of as much as 650 kilometers.

Turner's approach was sound, but his data were sketchy by modern standards, existing knowledge of the earth's deep structure was incomplete and the inaccurate clocks of the day made the timing of wave arrivals inaccurate by seconds or even minutes. Turner's argument convinced few of his contemporaries. S. K. Banerji of the Bombay Observatory pointed out that if the bulk of earthquakes have focuses at a depth of 150 kilometers or more, few events should produce strong surface waves, and yet Turner's own catalogue reported surface waves in abundance. Harold Jeffreys of the University of Cambridge put forward a more fundamental objection: he argued that earthquakes simply could not take place at such depths.

Below a depth of about 50 kilometers, Jeffreys contended, heat and pressure change the mantle rock from a brittle material, capable of fracturing, to a ductile one. In support of his argument Jeffreys pointed out that since the end of the most recent ice age, coastlines in Canada and northern Europe, relieved of the glaciers' weight, have risen as if the underlying mantle was capable of flowing under stress, like an extremely viscous liquid. He also cited laboratory work confirming that at high temperatures and pressures rock deforms gradually in response to stress instead of fracturing suddenly.

Kiyoo Wadati, a 25-year-old employ-

CLIFF FROHLICH has been a research scientist at the Institute for Geophysics of the University of Texas at Austin since 1978. He received his B.A. at Grinnell College in 1969 and his Ph.D. from Cornell University in 1976. For his thesis he studied the structure of the mantle under the southwestern Pacific by analyzing seismic waves from deep earthquakes, and deep earthquakes have remained at the center of his research since then. In addition Frohlich maintains a lively interest in biomechanics and the physics of sports; among his articles on the subject is "The Physics of Somersaulting and Twisting" (SCIENTIFIC AMERICAN, March, 1980).

ee of the Japan Meteorological Agency, did not refute Jeffrey's argument; he simply presented convincing evidence that some earthquakes are very deep. The frequency and destructiveness of earthquakes in Japan had led the Japanese government to establish what was then the world's best network of seismographic stations. Hence Wadati had abundant data, and he applied new methods of determining earthquake depths. Instead of comparing absolute arrival times at different seismographic stations, as Turner had done, he relied on a time difference that could easily be measured at individual stations even if clocks were inaccurate: the interval between the arrival of *P* waves and the arrival of the slower *S* waves. Because each kind of wave travels at a fairly constant speed, the interval increases in proportion to the station's distance from the earthquake focus.

For most earthquakes, Wadati discovered, the interval was quite small near the epicenter, the point of strongest shaking. For a few events, however, the delay was long even at the epicenter. Wadati saw a similar pattern when he analyzed data on the intensity of shaking. Most earthquakes had a small area of intense shaking, which weakened rapidly with increasing distance from the epicenter, but others were characterized by a lower peak intensity, felt over a broader area. Both the *P-S* intervals and the intensity patterns suggested two kinds of earthquakes: shallow events, in which the focus lay just under the epicenter, and deep events, with a focus several hundred kilometers down.

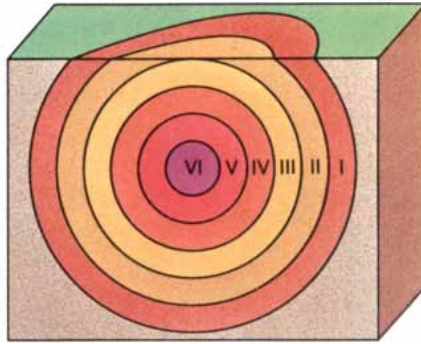
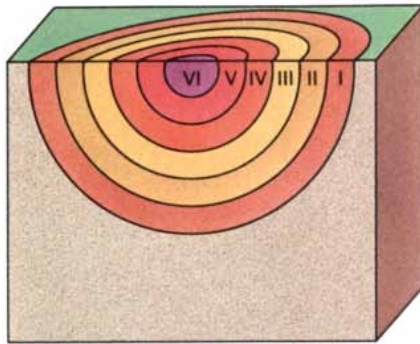
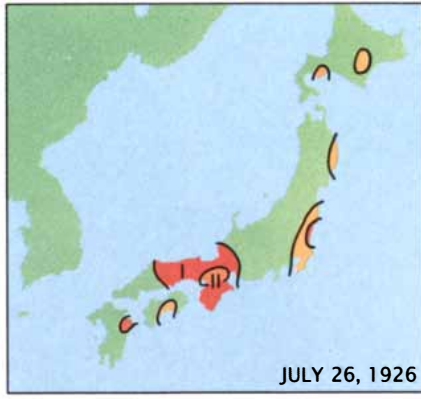
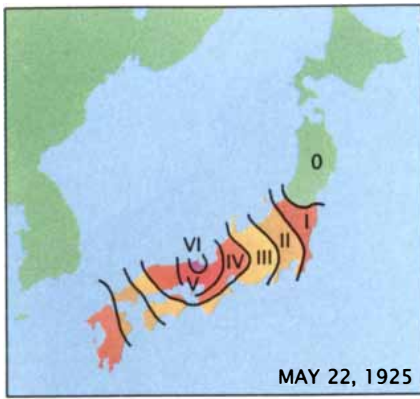
Other workers applied Wadati's techniques to data from earthquakes in other geographic areas and confirmed his results: "normal" earthquakes had a focal depth of 50 kilometers or less, and yet a few events had much deeper origins—as deep as 600 kilometers or more. Turner had been wrong about the depth of normal earthquakes, but deep events did exist. Banerji too had been right: the seismograms of confirmed deep events showed that they produced only weak surface waves.

What of Jeffreys' assertion that mantle rock at a depth of more than 50 kilometers is too ductile to store the stress needed for an earthquake? An observation Wadati made foreshadowed part of the answer: deep earthquakes do not take place in ordinary mantle rock. In 1935 Wadati published a map showing the sites of earthquakes near Japan and their focal depths. He had

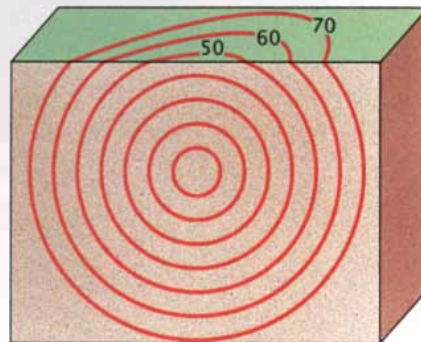
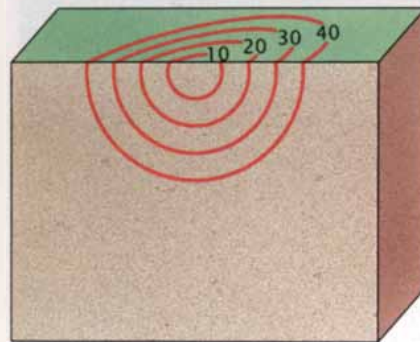
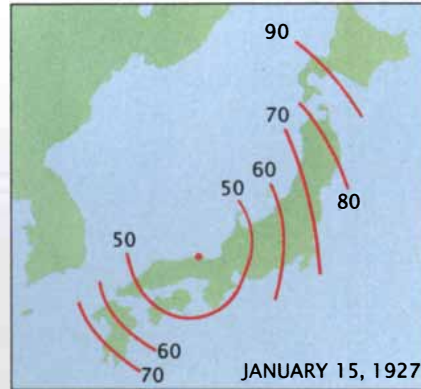
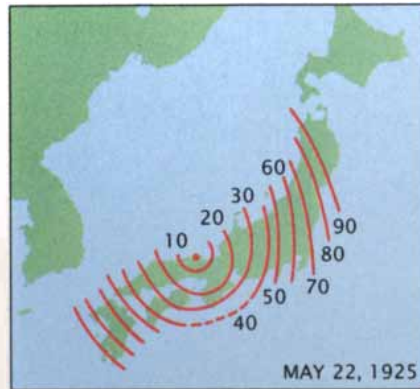


DESTRUCTIVE DEEP EARTHQUAKE is a rare event. The only severe one in recent decades took place on March 4, 1977, 150 kilometers under Bucharest. Columns on the ground floor of this building failed, allowing the corner to slump by one floor. Similar destruction throughout the city killed 1,500 people. The photograph was made by Neculai Mandrescu of the Institute for Physics of the Earth in Bucharest.

INTENSITY OF SHAKING



S-P DIFFERENCE (SECONDS)



SURFACE MANIFESTATIONS distinguish shallow and deep events. The maps compare data on the intensity of shaking and the delay between the arrival of *P* (compressional) and *S* (shear) waves for several earthquakes between 1925 and 1927. The data were collected by the Japanese seismologist Kiyoo Wadati. Shaking was very intense near the epicenter of a 1925 event but fell off rapidly with distance (*top left*); a 1926 earthquake produced less intense shaking that fell off more slowly (*top right*). The *S-P* difference at the epicenter of the first event was less than 10 seconds, but the delay increased rapidly with distance (*bottom left*); the minimum *S-P* difference for a 1927 event was longer—about 40 seconds—but it increased more slowly (*bottom right*). From the data Wadati concluded that the 1925 earthquake had a focus near the surface and the later events had focuses some 400 kilometers down (*cutaway views*).

found that the focuses lay along approximately parallel contour lines, and that the depths of the contours increased steadily from the east coast of Japan westward. He commented: "The possibility of drawing contour lines of the focal depth suggests that there exists in the crust something like a *weak surface*... where the earthquake outburst is liable to occur. This surface extends slopewise in the crust near the Japanese Islands."

Earthquake depths in other parts of the world, Wadati continued, define similar sloping surfaces: "Deep-focus earthquakes are apt to take place on one side nearer to the continent and shallow-focus ones on the other side, [which] is in most cases bordered [by] a very deep sea. This tendency seems to be observable in many volcanic regions in the world."

Almost all deep earthquakes conform to the pattern Wadati described. Wherever they are common—generally at the edge of a deep ocean—they define an inclined zone extending from near the surface to a depth of 600 kilometers or more. Now known as Wadati-Benioff zones, after Wadati and the seismologist Hugo Benioff, who mapped such zones in the 1940's and 1950's, the earthquake patterns provided crucial evidence for the new geophysical paradigm that emerged in the 1960's.

This new view of the earth accounts for many of the earth's major surface features and much of geologic history in terms of a set of moving plates covering the earth's surface. The plates spread apart from mid-ocean ridges; where they collide, generally at the edge of ocean basins, they thrust up mountains and sculpt the margins of continents. This process of plate tectonics is the surface expression of convection, or heat-driven circulation, in the earth's mantle. Hot material rises from within the mantle and circulates horizontally near the earth's surface. The top 50 kilometers or so of the horizontal flow cools to form the rigid plates, which include the earth's crust and some of the underlying mantle.

The cold, downgoing limb of the circulation is found where plates converge. There one plate is subducted: it bends under the other plate and sinks back into the mantle. Deep earthquakes helped to establish the reality of subduction when it was realized that they take place in a descending slab and that the Wadati-Benioff zone traces its shape. The deep trench that is generally found just seaward of the Wadati-Benioff zone—the "deep

sea" of Wadati's description—marks the downward bend of the subducted plate; the line of volcanoes that often forms nearby is fed by molten material rising from the slab. Wadati had been prescient: in his 1935 paper he had speculated that the earthquakes and volcanoes near Japan might be the result of continental drift (a forerunner of plate tectonics), which had been proposed some 20 years earlier by Alfred Wegener.

That geophysical scheme supplies part of the answer to Jeffreys' objection. The deep earthquakes in a Wadati-Benioff zone take place in rock that is hundreds of degrees colder than the surrounding mantle and hence is less ductile and better able to store elastic energy. Yet other factors also seem to affect the distribution of deep earthquakes. Their focuses, for example, are not scattered evenly along the Wadati-Benioff zone. Instead changes in earthquake frequency seem to coincide with depths where the crystal structure of mantle rocks changes to a denser phase as a result of increasing pressure.

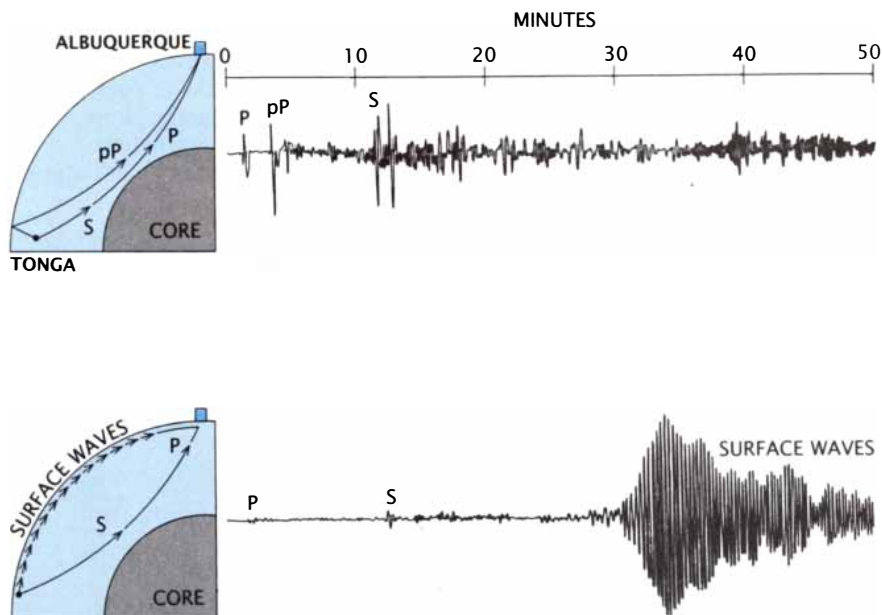
At about 400 kilometers, for example, seismographic studies show a sudden increase in seismic-wave velocity, indicating an increase in rock density. There olivine—a silicate compound with various admixtures of iron and magnesium that is the main constituent of the mantle and the subducted plate—changes to a denser crystal phase known as spinel. At about this depth the number of deep earthquakes falls to a minimum.

Some subduction zones stay quiet below this first transition. Zones that do have earthquakes at these depths, however, show their highest level of activity from below the transition down to another, more mysterious boundary at a depth of about 650 kilometers. Again a sharp increase in seismic-wave velocity marks the boundary, but workers disagree about whether the increase in density it indicates represents a second phase change or a change in composition. In any event earthquake activity drops abruptly to zero near this second boundary.

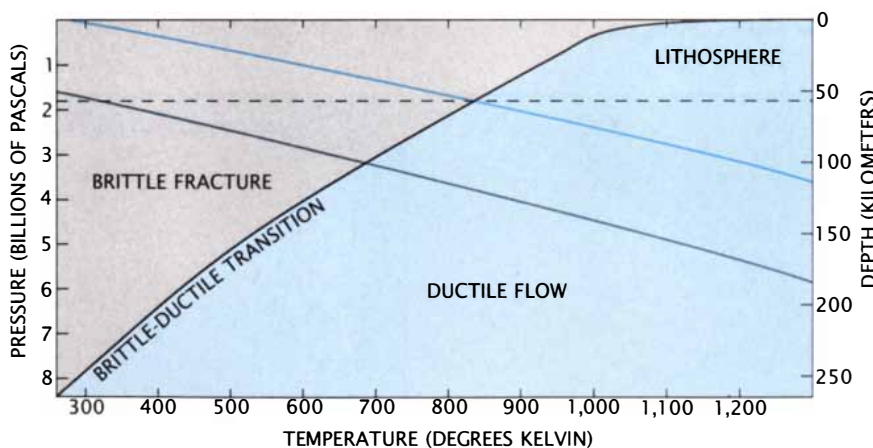
In an effort to determine the maximum depth of deep events Philip B. Stark, then at the University of Texas at Austin, and I applied several means of analysis. Among other things, we examined the intervals between ordinary *P* waves and *pP* waves (pressure waves that travel upward to the surface and are reflected back through

the earth to a distant seismographic station). We found that the deepest recorded events took place at a depth of between 680 and 690 kilometers. Beth A. Rees and Emile A. Okal of Northwestern University did a similar analysis, with comparable results.

The cutoff of earthquake activity is too abrupt to result from a gradual softening of the slab as it is heated by the surrounding mantle. It may mean that subducted slabs cannot penetrate the 650-kilometer boundary. In that case convection may be confined to



SEISMOGRAMS of deep and shallow earthquakes in Tonga were recorded in Albuquerque, N.Mex., nearly a fourth of the way around the world. The deep event (*top*) produced strong *P* and *S* waves, which passed through the earth at different speeds. Some of the *P* waves took the form of *pP* waves, having traveled upward from the focus to the earth's surface and then been reflected back through the earth. Because of its depth—about 625 kilometers—the event produced only a few surface waves. Both the *P* and the *S* waves from the shallow event (*bottom*) were relatively feeble; most of the energy was observed as surface waves, the last form of signal to arrive.



PUZZLE OF DEEP EARTHQUAKES is shown in a chart of the conditions of pressure and temperature under which rock changes from a brittle substance that can fracture when it is stressed, causing an earthquake, to a ductile medium that responds to stress by deforming gradually. Pressure and temperature increase with depth, so that rock ordinarily becomes ductile at a depth of about 60 kilometers (*blue curve*). Rock in deep-earthquake zones is anomalously cool, but even if it is 500 degrees cooler than "normal" mantle, the rock should still be ductile by about 100 kilometers (*gray curve*). Yet deep earthquakes have been recorded at nearly 700 kilometers.

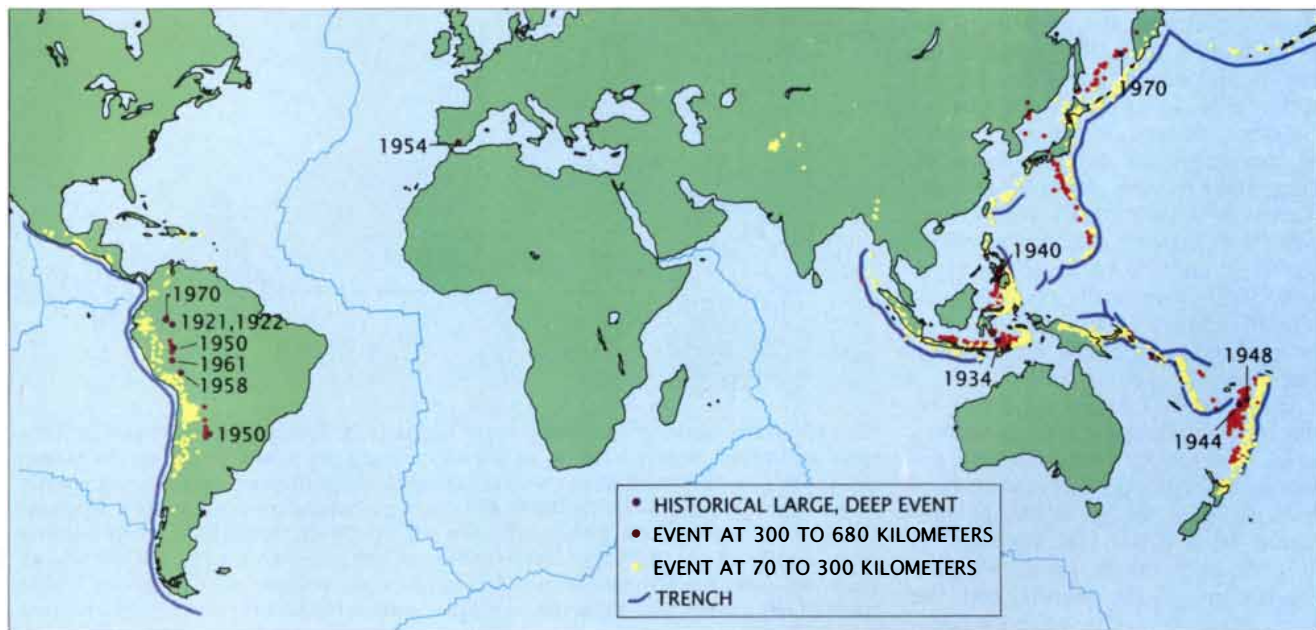
the upper mantle, above the boundary, and material from the upper and lower mantle may never mix. The change in seismic-wave velocity at 650 kilometers would then be likely to mark a change in the composition of the mantle. Alternatively (and this is one of the most hotly debated controversies in solid-earth geophysics), the subducted slab may penetrate the lower mantle. Convection would then involve the entire mantle, and the 650-kilometer boundary would mark a phase change in a compositionally uniform medium. A concomitant change in the rock's

mechanical properties would be responsible for the cutoff of earthquake activity.

The descent of a subducted plate provides several possible sources of the stress released in deep earthquakes. A descending plate is variously bent, stretched and compressed; heating and phase changes could also generate stresses by changing rock volume. How is such stress released? What actually happens at the focus of a deep earthquake?

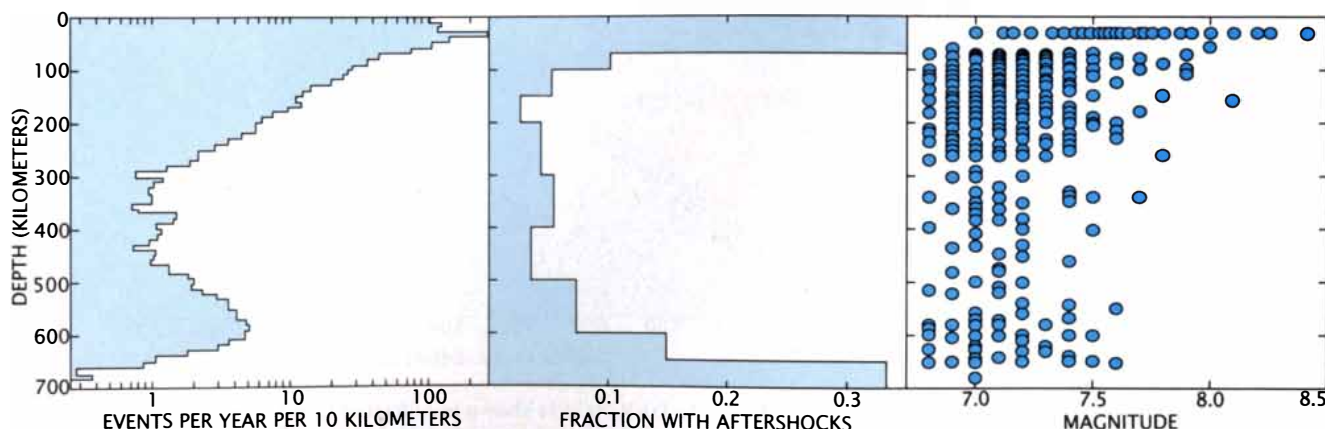
Deep earthquakes can be as large as

all but the very largest shallow earthquakes; the most destructive deep earthquake of recent years, the 1977 Romanian event, had a magnitude of 7.2, and an earthquake with a magnitude of 7.6 took place 650 kilometers under Colombia in 1970. And yet the mechanism by which the energy is released must differ from the brittle fracture that triggers shallow earthquakes. Even though the material in which deep earthquakes occur is much cooler—hence stronger—than Jeffreys had thought, it still should not fracture as rocks do at low pres-



MAP SHOWS DEEP EVENTS recorded over the past 25 years, distinguished by depth; it also shows historical very large earthquakes (those that have a magnitude of more than 7) at extreme depths (more than 630 kilometers) and gives their

year. Nearly all deep earthquakes occur near a deep-sea trench, where one of the rigid plates of lithosphere—the crust and uppermost mantle—that make up the surface of the earth is being drawn into the mantle through the process of subduction.



EARTHQUAKE STATISTICS change with depth. The number of earthquakes with a magnitude of 5 or more in each 10-kilometer interval of depth (left) reaches a minimum at about 400 kilometers but then increases again before abruptly falling to zero at about 650 kilometers. Aftershocks are rare for most deep events of moderate size but become commoner at the

greatest depths (middle). The strongest earthquakes are generally recorded at shallow depths, but at greater depths the size of the largest events remains quite constant until seismic activity stops altogether (right). The data suggest that changes in the crystal phase of mantle rocks, thought to occur at depths of 400 and 650 kilometers, may affect deep earthquakes.

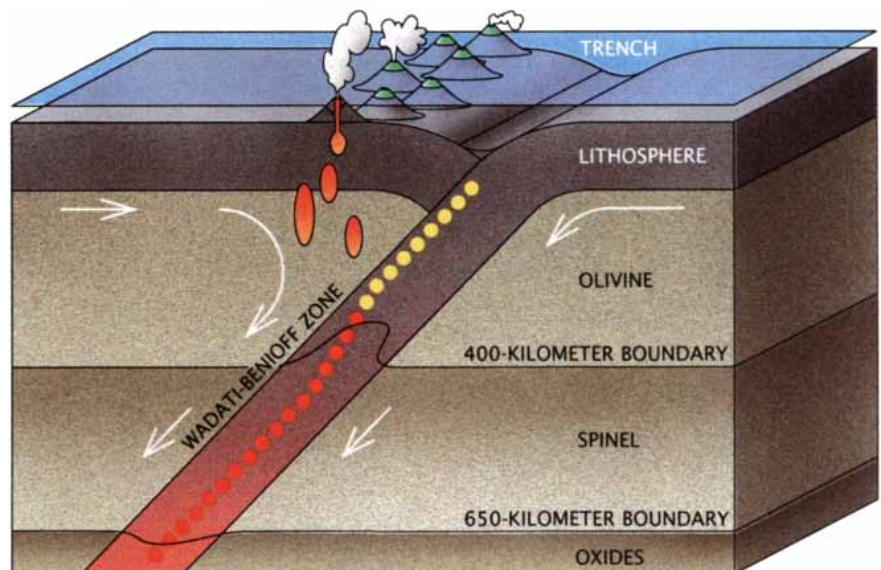
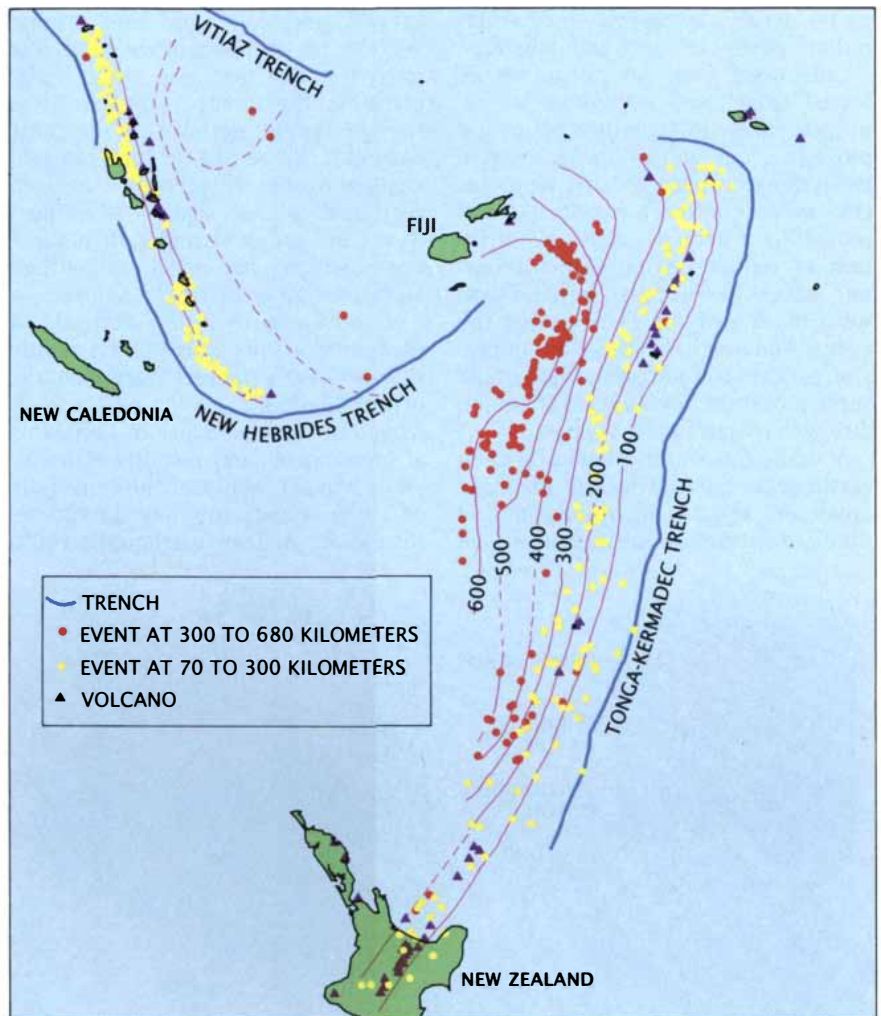
sures. If stress were to open a crack in the slab, the weight of all the rock above it would simply close it again. If the rock deforms at all, it should deform plastically. Jeffreys' objection still holds: conventional earthquakes, in which rock fractures and slips, should not happen in the mantle.

Recent seismographic studies of deep earthquakes support the case against a conventional mechanism. Nearly all large shallow events are accompanied by many smaller tremors known as aftershocks. Aftershocks often occur along the same plane as the initial slip, apparently releasing residual stress along the fracture. Aftershocks are much rarer for deep events. The 1970 deep earthquake under Colombia, which was probably the largest really deep event of the past 25 years, had no aftershocks whatever. Deep earthquakes that do have aftershocks—and they are commonest among the deepest events—generally have only one or a few at most.

The aftershocks that do occur define a spatial pattern quite different from those of shallow earthquakes. Recently Raymond J. Willemann, then at the Los Alamos National Laboratory, and I studied the spatial relations between the initial shocks and the aftershocks of deep earthquakes. Small shallow earthquakes often have aftershocks centered relatively near the main event, which is consistent with the idea that the aftershocks represent continued slip along the same fracture that produced the main tremor. We found, however, that some small deep earthquakes—ones with a magnitude of 5.5 or less—have aftershocks at a distance of 30 kilometers or more from the initial shock. A rupture zone responsible for such a small earthquake is not likely to be 30 kilometers long.

Deep aftershocks, moreover, do not fall along a plane, as shallow ones often do. The existing data suggest, on the contrary, that they are more or less randomly distributed in three-dimensional space around the initial event. Again the pattern suggests that deep earthquakes and shallow ones have fundamentally different mechanisms.

One attractive but incorrect way of accounting for deep earthquakes was proposed almost as soon as they were discovered. It holds that they are the direct result of the transformation of subducted material to a denser phase. Such transformations must take place in the subducted rock, and if they happened fast enough—if the rock in effect imploded



SUBDUCTION ZONE is the setting for nearly all deep earthquakes. The focal depths of earthquakes along the Tonga-Kermadec Trench, a deep-sea trench in the southwestern Pacific bordered by seismic activity and volcanic islands, fall along a series of parallel contour lines of increasing depth—a pattern known as a Wadati-Benioff zone (top). The Wadati-Benioff zone traces the subduction of a lithospheric plate (bottom): the earthquakes take place within the descending slab. The downward bend of the subducting plate is responsible for the trench, and molten material rising from the plate feeds the line of volcanoes. The process is driven by the convective circulation of the mantle; the descending plate is the cold, downgoing limb of the circulation.

as its density increased—they could radiate energy as earthquake waves.

Unfortunately the seismic waves detected from deep earthquakes look nothing like the signature of an implosion. In an implosion nearby material moves inward, toward the focus. One would therefore expect all seismographs (which register the direction as well as the amplitude of seismic waves) to record an initial downward motion of the earth during the event. Moreover, because an implosion produces radial rather than transverse motion, it would generate much stronger *P* waves than *S* waves.

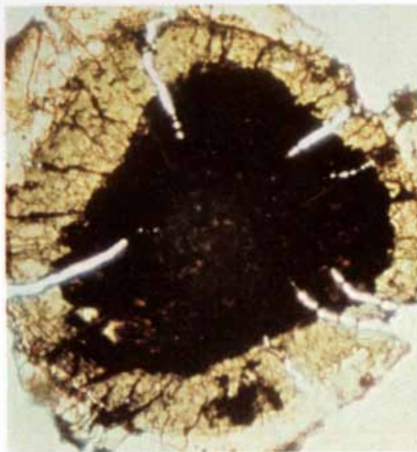
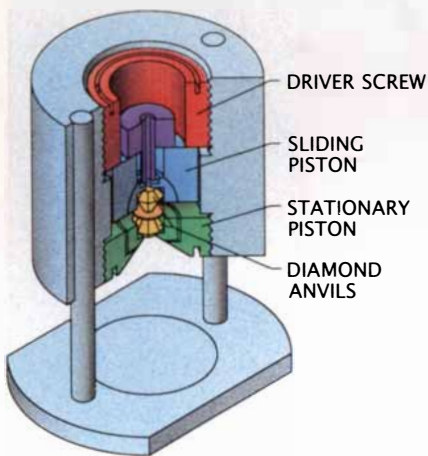
Actually the first motions of a deep earthquake are downward in some areas and upward in others, just as they are in shallow earthquakes. The

upward and downward motions are segregated, as if part of the earth had moved in one direction along a slip plane and the other part had moved in the opposite direction; it is the same pattern observed in seismograms of shallow events. Furthermore, in deep earthquakes as in shallow ones the *S* waves are much stronger than the *P* waves, which points to slip rather than an implosion as being the source.

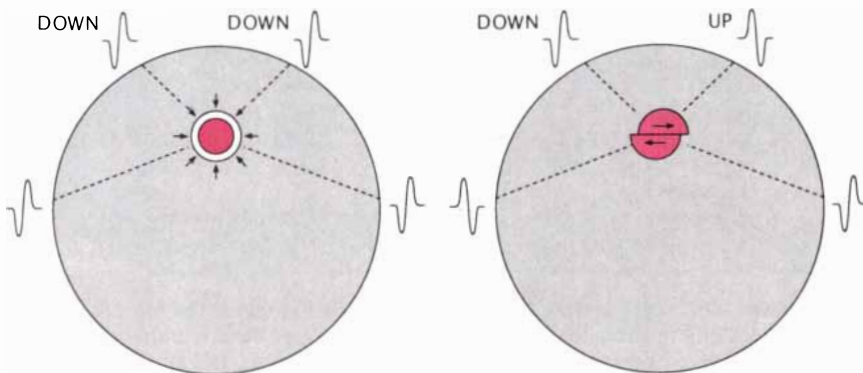
So how can rock slip abruptly, if the enormous pressures of the mantle rule out brittle fracture? One scenario, proposed in the 1960's by David T. Griggs of the University of California at Los Angeles and recently elaborated by Masaki Ogawa of the University of Tokyo, posits runaway ductile deformation. A deep earthquake could

take place when rock deforming under shear stress begins to produce frictional heat faster than the surrounding rock can carry it away. The heat softens the rock and even melts some of it, accelerating the deformation. This feedback process could cause both the temperature and the slip rate to increase explosively and produce an earthquake.

The plausibility of the mechanism depends strongly on the composition and structure of the rock in a Wadati-Benioff zone. It is most strongly favored if rock structure—an existing weak layer, for example—tends to concentrate ductile slip along a plane. It is by no means certain that the layering in subducted material has the right orientation to foster slip in the directions most often observed in deep earthquakes.



PHASE CHANGES in mantle rock, which may have a role in deep earthquakes, can be simulated in a device called the diamond-anvil cell, which compresses rock samples between two diamonds (*left*). The diamonds' transparency allows the samples to be heated with a laser and photographed. In a sample of olivine (a primary mantle constituent) that has been compressed and heated to as much as 300,000 atmospheres and 1,500 degrees Celsius, distinct phases form concentric rings (*right*). A pale outer ring of unaltered olivine gives way to yellower spinel in a transition thought to occur at a depth of 400 kilometers; at the center, where pressure and temperature are highest, is the dark oxide phase into which spinel may change at 650 kilometers. The photograph was provided by William A. Bassett of Cornell University.

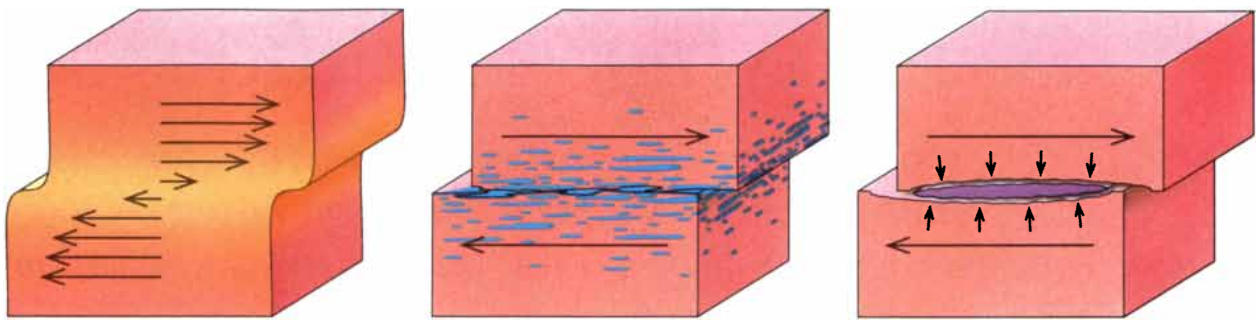


IMPLOSION due to the sudden transformation of subducted rock to a denser phase can be ruled out as a cause of deep earthquakes. An implosion should show itself on seismographs around the world as an initial downward motion (*left*). Instead deep earthquakes generate an initial upward motion at some points and a downward motion at other points (*right*), suggesting their source is lateral slip in deeply buried rock.

A different scenario attributes slip in deep-earthquake zones to the effect of trapped fluids. Laboratory work has shown that at pressures equivalent to shallower depths, fluids trapped in rock pores can counteract the forces binding a potential fracture, allowing it to fail at a lower shear stress than before. In at least one case, at the Rocky Mountain Arsenal near Denver, a sequence of shallow earthquakes occurred after fluid wastes were injected into the earth, apparently lowering confining stresses enough for rock layers to slip.

In 1966 C. B. Raleigh and Mervyn S. Paterson of the Australian National University suggested that pore fluids in deep-earthquake zones might have the same effect, allowing subducted material to fracture like rock at lower pressures. The source of the fluids, Raleigh and Paterson proposed, might be the dehydration of minerals such as serpentine (a form of magnesium silicate) in the subducted material: the release of water incorporated in their crystal structure as they are heated by the surrounding mantle to temperatures above 500 degrees Celsius. Raleigh and Paterson also proposed other sources of fluid: water trapped in sediments in deep-sea trenches and carried down with the crust, and partially molten mantle rock.

To affect the bulk properties of rock, such as its tendency to fracture, a fluid must be able to migrate through it, and it is not certain that mantle rock is porous enough. Moreover, the dehydration of minerals would be expected to occur at specific temperatures and pressures, corresponding to specific depths. If pore fluids do trigger deep earthquakes, those depths might



PROPOSED MECHANISMS OF SLIP acknowledge that at great depths the earth cannot simply fracture. In one scenario slow rock deformation accelerates abruptly as the frictional heat it creates builds up, softening the rock and speeding the deformation in a runaway process (*left*). A second mechanism attributes slip to the influence of fluids (*middle*). Below a certain

depth high pressure might release water bound in the crystal structure of subducted minerals; the water might counteract forces binding potential faults and allow them to fail. A third proposal holds that shear stress could induce a phase change in a layer parallel to the stress (*right*). The sudden change in crystal structure would weaken the rock, allowing it to slip.

show intense concentrations of seismicity. Yet the frequency of earthquakes shows only a moderate variation with depth.

What variation there is (the falloff in earthquake activity at the 400-kilometer olivine-spinel boundary and its revival at greater depths) seems to bear some relation to the depths of phase changes. Stephen H. Kirby of the U.S. Geological Survey has proposed a deep-earthquake mechanism that depends on phase transitions but, in contrast to earlier proposals, results in slip rather than implosions. As a surrogate for actual mantle rock Kirby and his colleagues studied ice and tremolite, a calcium magnesium silicate, both of which change to a denser phase at pressures that can readily be produced in the laboratory.

When the workers compressed each material to a pressure slightly below that of the normal phase transition and subjected it to shear stress, they found that the phase transition was triggered along a thin layer parallel to the stress. The sudden rearrangement of crystal structure along the layer apparently weakened the material, allowing it to slip. Kirby and his colleagues noticed that in the process their samples emitted cracking or snapping noises—laboratory analogues of earthquakes.

Kirby proposes that such premature phase changes also take place in subducted rock as it is stressed, and that the resulting slip accounts for at least some earthquakes in Wadati-Benioff zones. The proposal would not conflict with the occurrence of deep earthquakes over a broad range of depths below the 400-kilometer boundary: several investigators, including William A. Bassett of Cornell University, have found that phase transitions in subducted material can occur at a variety of depths depending

on its precise composition and the speed at which it is descending. Kirby's mechanism would, however, account for the abrupt disappearance of deep earthquakes at depths greater than 680 kilometers. At that depth all the known phase transitions in the mantle have already taken place.

No one has yet shown that shear stress has the same effect on phase transitions in actual mantle rock that it has on transitions in ice and tremolite. But even if Kirby's hypothesis is wrong, phase transitions may play some role in deep earthquakes. Perhaps they simply generate stresses that are abruptly released elsewhere, by some unknown failure mode.

It may soon be possible to choose among the various hypotheses with greater confidence. Raymond Jeanloz and Charles Meade of the University of California at Berkeley are re-creating mantle conditions in the laboratory in order to study proposed deep-earthquake mechanisms. A fist-size press called a diamond-anvil cell generates the needed pressures by squeezing a minute rock sample between the points of two diamonds. The sample can be heated by shining a laser through one diamond; phase transitions and other changes in the rock can be seen through the other diamond, while acoustic sensors detect any "earthquakes." The studies are still quite preliminary, but so far they suggest that at high pressures olivine fails only when it also contains serpentine—a result favoring Raleigh and Paterson's dehydration mechanism.

By respectively demonstrating the reality of deep earthquakes and their "impossibility," Wadati and Jeffreys posed a puzzle that geophysicists are still struggling to solve. (As it happens, both men are still alive more than 60 years later.) In the con-

text of plate tectonics and mantle convection, which deep earthquakes themselves helped to establish, these events have led to new puzzles.

One concerns the earthquake cutoff at a depth of 680 kilometers: does it mark the lower limit of mantle convection or just a change in the mechanical properties of a mantle that is convecting throughout its depth? The occasional deep earthquakes that occur in regions lacking known subduction zones embody another puzzle. Deep earthquakes in Romania and the Hindu Kush, two such regions, may reflect the presence of an old subduction zone obscured by later tectonic activity. That explanation is less plausible for the tremors sometimes recorded under northern Africa and Spain. There the riddle of deep earthquakes is wrapped in a further mystery: the possibility that some deep earthquakes can occur in the complete absence of subduction.

FURTHER READING

- SHALLOW AND DEEP EARTHQUAKES. K. Wadati in *The Geophysical Magazine*, Vol. 1, No. 4, pages 162-202; March, 1928.
- DEEP-FOCUS EARTHQUAKES AND THEIR GEOLOGICAL SIGNIFICANCE. Andrew Leith and J. A. Sharpe in *The Journal of Geology*, Vol. 44, No. 8, pages 877-917; November-December, 1936.
- KIYO WADATI AND EARLY RESEARCH ON DEEP FOCUS EARTHQUAKES: INTRODUCTION TO SPECIAL SECTION ON DEEP AND INTERMEDIATE FOCUS EARTHQUAKES. Cliff Frohlich in *Journal of Geophysical Research*, Vol. 92, No. B13, pages 13777-13788; December 10, 1987.
- LOCALIZED POLYMORPHIC PHASE TRANSFORMATIONS IN HIGH-PRESSURE FAULTS AND APPLICATIONS TO THE PHYSICAL MECHANISM OF DEEP EARTHQUAKES. Stephen H. Kirby in *Journal of Geophysical Research*, Vol. 92, No. B13, pages 13789-13800; December 10, 1987.

The Mixing of Fluids

Viscous fluids flowing in simple, periodic patterns in two dimensions can generate the chaos that leads to efficient mixing. Experiments and computer models reveal the underlying mechanism

by Julio M. Ottino

What do the eruption of Krakatau, the manufacture of puff pastry and the brightness of stars have in common? Each involves some aspect of mixing. Violent mixing of magmas might have triggered the eruption of Krakatau; stretching and folding—the archetypal mixing process—is applied in the making of puff pastry, and mixing in the interior of a star determines the chemical composition and therefore the brightness of the star's surface. Instances of mixing can be found literally throughout the universe, spanning an enormous range of time and length scales. Exhaled gases mix with the ambient air in a matter of seconds, whereas the mixing processes in the mantle of the earth can take several hundred million years or longer.

Mixing also plays a critical role in modern technology. Chemical engineers rely on mixing to ensure that substances react properly, to produce polymer blends that exhibit unique properties and to disperse drag-reducing agents in pipelines. Yet in spite of its ubiquity in nature and industry, mixing is only imperfectly understood. Indeed, investigators cannot even settle on a common terminology: mixing is often referred to as “stirring” by oceanographers and geophysicists, as “blending” by polymer engi-

neers and as “agitation” by process engineers.

Regardless of what the process is called, there is little doubt that it is exceedingly complex and is found in a great variety of systems. In constructing a theory of fluid mixing, for example, one has to take into account fluids that can be miscible or partially miscible and reactive or inert, and flows that are slow and orderly or very fast and turbulent. It is therefore not surprising that no single theory can explain all aspects of mixing in fluids and that straightforward computations usually fail to capture all the important details.

Still, both physical experiments and computer simulations can provide insights into the mixing process. Over the past several years my colleagues and I have taken both approaches in an effort to increase understanding of various aspects of the process—particularly of mixing involving slow flows and viscous fluids such as oils.

Stirring two oil-based paints is a good example of the mixing of viscous fluids. After just a few seconds of stirring one can produce a dizzying pattern of stretched and folded striations. (Bookbinders take advantage of this in the “marbling” that sometimes adorns the covers or endpapers of books.) Yet unless one has stirred purposefully, one will probably find that there are a few “islands” of unmixed paint among the convoluted streaks. Although the mixing of viscous fluids can produce fantastically complex structures, it can also produce patterns that have some regularity and coherence.

My students and I at the University of Massachusetts at Amherst have sought to characterize the flows that produce such patterns by doing experiments and computer simulations reminiscent of mixing two paints. In some of our experiments we inject blobs of dyed glycerine into a body of colorless glycerine in a deep cavity.

When the sides of the cavity are made to move periodically, the shearing forces they exert on the viscous fluid in the cavity can stretch and fold the colored blob in a rather complicated way; the entire cavity soon displays an intricate pattern of folds within folds. At the same time, however, a similar blob in the same receptacle can just as easily experience almost no stretching; the blob may move and rotate, but it regularly returns to its initial position. How do such markedly different patterns arise?

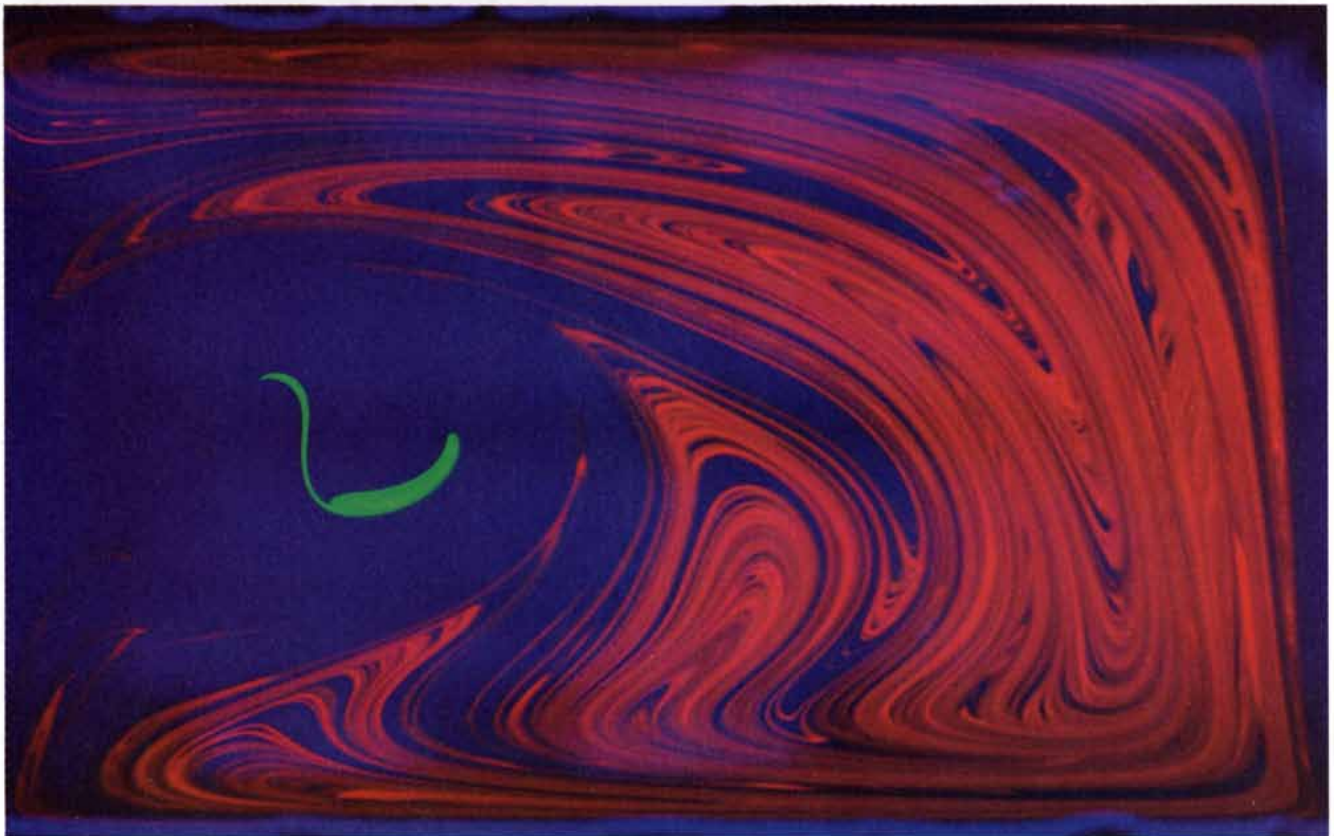
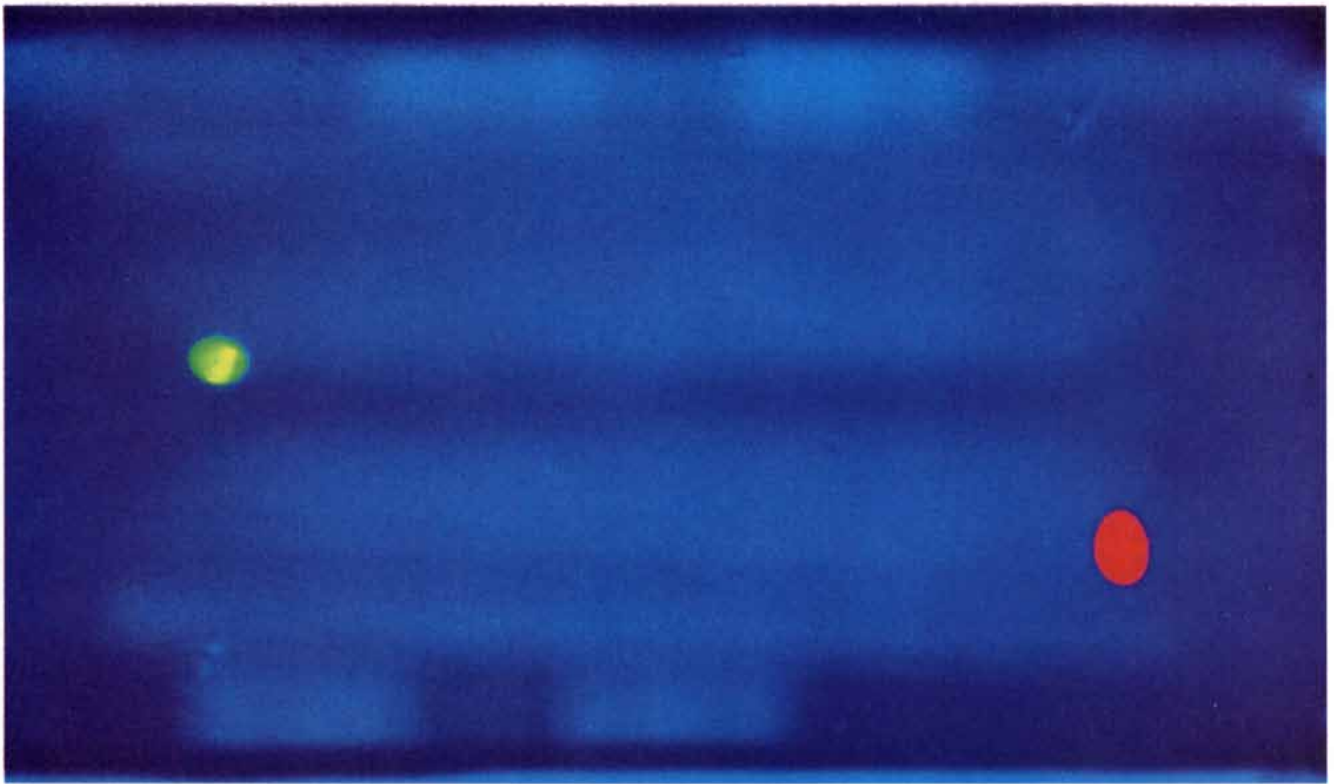
Foundations of Fluid Mechanics

The key to understanding the basic aspects of such mixing lies in the concept of “motion”—an idea that can be traced to the 18th-century Swiss mathematician Leonhard Euler. The motion of a fluid is a mathematical expression that states where each particle of the fluid will be at any future time. If one could know the motion for a particular flow, one would in principle know almost everything there is to know about the mixing it produces. One could, for example, calculate the forces and total energy necessary to achieve a certain degree of mixing in the system.

In the course of the past century the description of flow in terms of a fluid's motion has largely been abandoned in favor of a description based on the fluid's velocity field: an expression that specifies the velocity of the fluid at every point in the flow region at any given time. Yet if one knows the motion, it is an easy matter to compute the velocity field, whereas knowing the velocity field is not enough to calculate the motion explicitly. Because the motion is therefore in some sense a more fundamental description of flow, my co-workers and I prefer to work with what many may consider an antiquated concept.

Underlying the motion is what is known as a point transformation, a

JULIO M. OTTINO is professor of chemical engineering and adjunct professor of polymer science and engineering at the University of Massachusetts at Amherst. He studied at the National University of La Plata in Argentina and at the University of Minnesota, where he earned his Ph.D. in 1979. His experience in mixing colored materials is not limited to the laboratory; it extends to his paintings, which have been featured in an individual art show in Argentina. Ottino has recently finished a book titled *The Kinematics of Mixing: Stretching, Chaos, and Transport*, which will be published by Cambridge University Press.



CHAOTIC AND NONCHAOTIC flows are both evident in an experiment carried out by Kenny Leong and the author in their laboratory at the University of Massachusetts at Amherst. A rectangular cavity is filled with glycerine, and two blobs of tracer that fluoresce respectively in green and in red are injected just below the surface (*top*). Each side of the cavity can slide in a direction parallel to itself independently of the other sides. In this particular run the top and bottom sides were

made to move periodically but discontinuously. The top side moves from left to right for a time and stops, at which point the bottom side moves at the same speed and for the same length of time but from right to left; the pair of movements constitutes one period. After 10 such periods (*bottom*) the red blob has been stretched and folded several times: it was placed in a region of chaotic mixing. The green blob has been stretched only somewhat: it represents an "island" of nonchaotic mixing.

Celebrity. \$11,495.* A great way to teach your kids the value of the dollar.

- Front-drive, four-door, 6-passenger family sedan. ■ Electronic Fuel Injection ■ All-season steel-belted radials. ■ EPA estimated MPG city 23, highway 30.
- 3-year/50,000-mile Bumper to Bumper Plus Warranty.†
- Standard power steering and brakes.

THE *Heartbeat* OF AMERICA TODAY'S CHEVROLET

Show them how much you get in a Celebrity. Like room. A huge trunk. And lots of standards. Which is our way of giving you more for less than you'd expect. Then, show them the bottom line that's less than many family cars. Leaving you more to spend on your family. They'll like that part. Best.



*MSRP, including dealer prep. Tax, license, destination charges and optional equipment additional. †See your Chevrolet dealer for terms of this limited warranty.



Chevrolet and the Chevrolet emblem are registered trademarks of GM Corp. © 1988 GM Corp. All Rights Reserved. Let's get it together... buckle up.



© 1988 SCIENTIFIC AMERICAN, INC

mathematical operation that enables one to identify a particle of fluid and to specify its position at some time in the future. Each fluid particle is "mapped" to a new position by the application of the transformation. Particles initially identified as being separate cannot occupy the same position at the same time, and one particle cannot split into two. Although a point transformation exists in theory for all mixing flows, in only the simplest cases can it be obtained exactly. For this reason much of what is known about mixing is limited to relatively simple fluid flows, such as linear flows in which lines of tracer do not bend. Yet these types of flows cannot possibly capture the processes that lead to efficient mixing, which are inherently nonlinear. To get at least an idea of what is involved in such processes, one has to consider steady flows in two dimensions.

Two-dimensional Flows

All two-dimensional flows consist of the same building blocks: hyperbolic (or saddle) points and elliptic points [see illustration on page 63]. A fluid

moves toward a hyperbolic point in one direction and away from it in another direction, whereas a fluid circulates around an elliptic point. (I should also mention that there is a third type of point called a parabolic point, in which the fluid motion is shear, or tangential. Such points are found, for example, in the fluid flowing along a solid wall. Parabolic points can be neglected in describing the nature of mixing in two-dimensional flows.) As one might expect, mixing in a steady two-dimensional flow is rather inefficient in comparison with mixing in three-dimensional flows—particularly those that change continuously with time. In fact, there are just two main possibilities in a steady, bounded two-dimensional flow: the fluid particles either repeatedly follow the same paths, called streamlines, or they do not move at all.

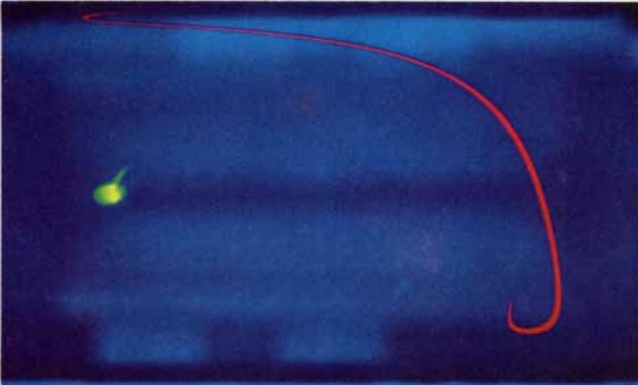
Because streamlines in steady flows are fixed and the trajectories of fluid particles can never cross, the fluid particles have no opportunity to come in contact with one another, that is, to become mixed. Is there a way to get rid of the streamline confinement so that fluid particles can avoid having to re-

peatedly follow the same streamline in the flow? There is, if the flow pattern can be made to change with time so that a streamline in one pattern crosses a streamline in a later pattern.

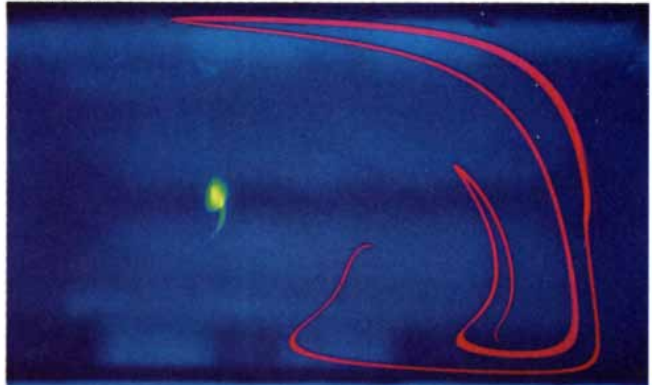
The simplest way of doing so (and the easiest to analyze) is to force the flow to vary with time in a periodic manner. For such a flow to lead to effective mixing, however, it must be capable of stretching and folding a region of fluid and returning it—stretched and folded—to its initial location. The stretching-and-folding operation corresponds to what is called a horseshoe map, as described by Stephen Smale of the University of California at Berkeley.

The fact that in order to mix a material efficiently one must return part of it to its initial location is counterintuitive. Yet if the mixing occurs in a bounded system, there is really no alternative. If one repeatedly throws a dart at a target, some throw will eventually fall arbitrarily close to any other throw, since there is simply a finite amount of area on the target. By the same token, repeated stretching and folding in a closed cavity will invariably place fluid particles very close to

1 PERIOD



3 PERIODS



8¼ PERIODS



8½ PERIODS



STRETCHING AND FOLDING characteristic of chaotic mixing is traced out by the red blob in this set of photographs of the experiment described in the illustration on page 57. After just three periods the basic stretching-and-folding pattern is plain-

ly visible. The green island that marks a region of mostly non-chaotic mixing and the folds that mark a region of chaotic mixing move about the cavity, but they return to the same positions (albeit somewhat deformed) after each period. The

their initial positions at certain times.

If a fluid particle in a periodic flow returns to its exact initial position after a certain amount of time, that particle defines a so-called periodic point. Depending on the number of periods required for the particle to return to its starting position, it is referred to as a periodic point of period one, period two and so on. A periodic point can also be classified as hyperbolic or elliptic, depending on the direction of the flow in its immediate neighborhood.

As a periodic elliptic point traces its cyclic trajectory, the material surrounding the point not only circulates around it (as it would around a fixed elliptic point) but also moves with it. In spite of the material's rotation and translation, however, it does not readily shed matter to the rest of the flow. Such regions of material are seen as "islands" of fluid, and mixing within islands is typically slow. Since material can neither enter nor leave the neighborhood of a periodic elliptic point, such points are obstacles to efficient mixing.

Similarly, as a periodic hyperbolic point traces its cyclic trajectory, sur-

rounding material that moves with the point undergoes contraction in one direction and stretching in another. In doing so the point sheds stretched filaments of fluid in one direction and attracts material in another. (If one assumes that the fluids are incompressible, the stretching and contraction must balance each other.)

The Fingerprint of Chaos

Where does the material shed from a periodic hyperbolic point go? From where does the material approaching the point come? One possibility is that an inflow joins smoothly with an outflow—that the material being shed from a hyperbolic point is attracted to the same or another hyperbolic point. This is in fact what happens in a steady flow (although in that case the hyperbolic points are fixed and not periodic), and as a result the flow does not stretch and fold the material efficiently.

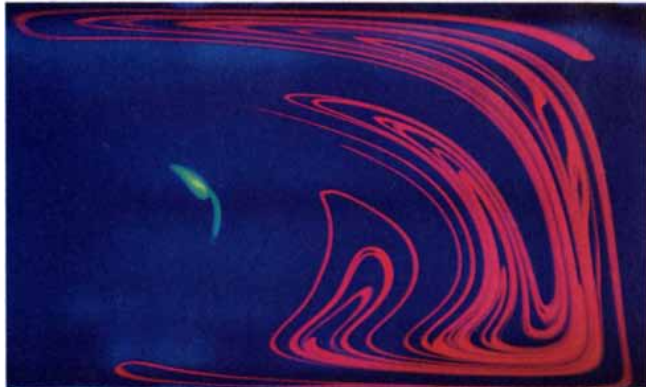
Time-dependent two-dimensional flows can result in efficient stretching-and-folding mechanisms, however, because in such flows it is possible for a region of outflow associated with a

hyperbolic periodic point to cross the region of inflow of the same or another hyperbolic point. A point where inflow and outflow of a single hyperbolic point intersect is called a transverse homoclinic point. If the intersection results from the flows of two different hyperbolic points, it is called a transverse heteroclinic point.

Homoclinic and heteroclinic intersections are the fingerprint of chaos. From a mathematical viewpoint, then, a system that is able to produce horseshoe maps or transverse homoclinic or heteroclinic intersections can be classified as being chaotic. It turns out that a horseshoe map actually implies the existence of transverse homoclinic points; likewise identifying a single such point is sufficient to imply the existence of a horseshoe map.

The fact that a single crossing of inflow and outflow leads invariably to transverse homoclinic points, and that such a crossing can occur even in what appear to be "well-behaved" physical systems described by Newton's laws of motion, was originally discovered by the 19th-century French mathematician Henri Poincaré. Yet the analysis of the astoundingly complex behav-

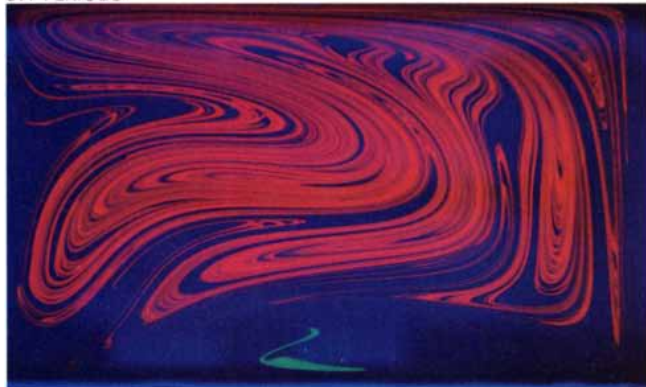
5 PERIODS



8 PERIODS



8 1/4 PERIODS

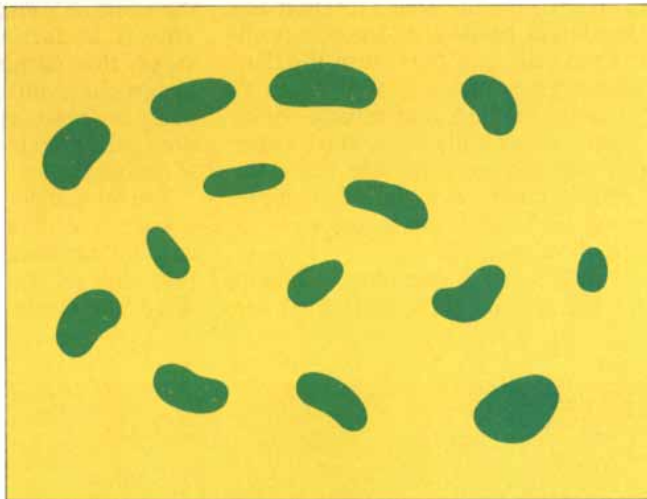
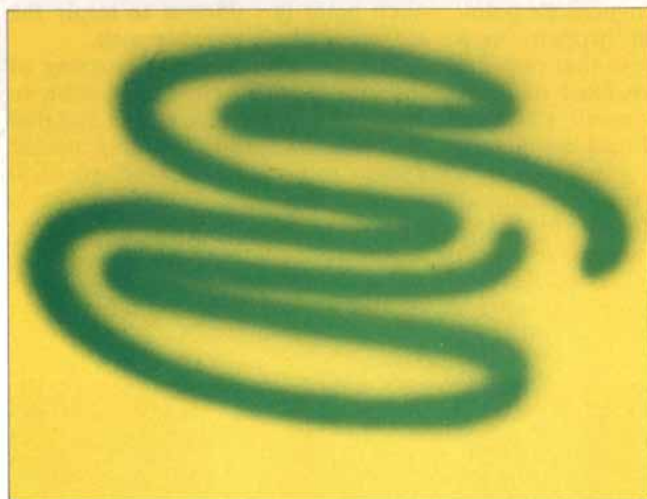
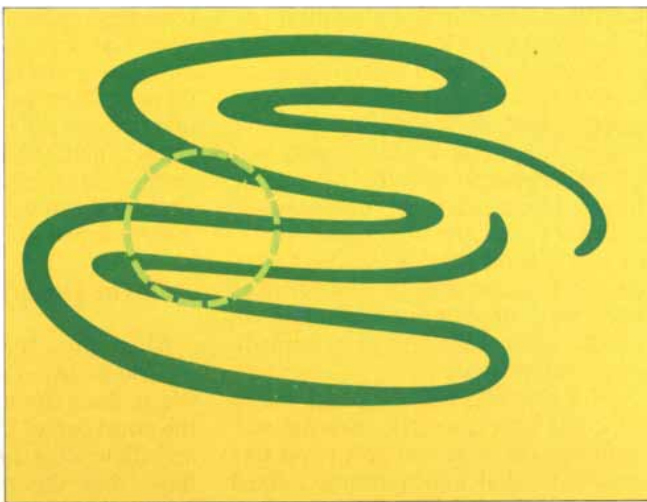
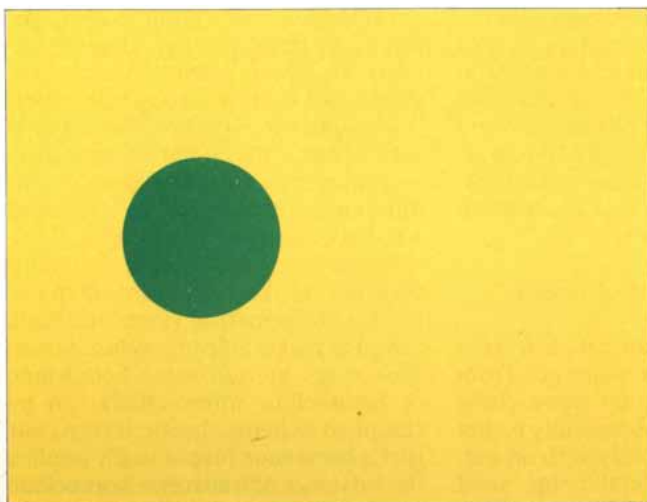


9 PERIODS



wisp that has been drawn from the green blob reveals that the blob undergoes a complete rotation approximately every two periods. If the experiment were run backward, the green blob would return approximately to its initial shape and posi-

tion, since the error in reproducing its motions in reverse grows only linearly. On the other hand, "unmixing" the red blob from the bulk fluid is virtually impossible: the error in reproducing its motions in reverse grows exponentially.



MIXING of fluids in natural or industrial processes generally involves molecular diffusion and breakup as well as stretching and folding. Only in an idealized case can a blob of dye in a body of fluid (*top left*) be stretched and folded indefinitely without diffusing or breaking (*top right*). An interesting feature of such a hypothetical case is that in order to have efficient

mixing, part of the blob must return to its initial position. Molecular diffusion (without which perfect mixing is impossible) normally makes the boundaries between miscible fluids indistinct (*bottom left*). In the case of immiscible fluids the blob, in stretching, can break into droplets that can subsequently coalesce to form several smaller blobs (*bottom right*).

ior that ensues from such crossings (which is today known as chaos) overwhelmed Poincaré, and he decided not to explore the matter further.

To the extent that mixing can be represented by a deterministic point transformation, mixing should be kinematically reversible. In other words, it should be possible to “unmix” fluids (at least if one disregards molecular diffusion). Yet everyday experience suggests that mixing is an irreversible process. Even though the system is deterministic in theory, the motions leading to repeated stretching and folding cannot be undone.

A rather similar situation exists in other physical systems, such as those analyzed by Poincaré, that consist of many particles whose respective motions are described by deterministic equations. (These kinds of systems

are typically referred to as Hamiltonian systems.) One of the most noted 19th-century American physicists, J. Willard Gibbs, recognized that even Hamiltonian systems can have an inherent irreversibility and unpredictability, and the fact that he invoked a thought experiment having to do with mixing to explain it is a measure of his insight. Apparently his observation went unnoticed until the Swedish oceanographer Pierre Wellander pointed it out in a perceptive 1955 journal article.

Capturing Chaos in Flows

The fact that stretching and folding has a prominent role in mixing had been known in chemical engineering since the 1950's as a result of the pioneering work of Robert S. Spencer

and Ralph M. Wiley at the Dow Chemical Company and of William D. Mohr and his co-workers at E. I. du Pont de Nemours & Company, Inc. The consequences of this fact—the existence of horseshoe maps and homoclinic and heteroclinic points—were not recognized until recently.

The Russian mathematician Vladimir I. Arnold appears to have made the first direct connection between chaos and fluid flows. According to Michel Hénon, a French astronomer working at the Nice Observatory, Arnold suggested in 1965 the possibility that fluid-mechanical systems can display chaotic particle trajectories. Hénon pursued Arnold's conjecture, and in a three-page paper that contained only one figure he was able to demonstrate that a three-dimensional, steady flow of a fluid without viscosity can in-

deed give rise to chaotic streamlines.

In 1984 Hassan Aref, then at Brown University, observed that the equations describing the trajectories of fluid particles in a two-dimensional flow are formally identical with those describing a Hamiltonian system. He carried his observation further, proving by means of a computational example that a Hamiltonian system subject to periodic forces can indeed produce effective mixing.

In three dimensions the analogy between mixing and Hamiltonian systems does not work, but in two dimensions the analogy is exact: fluid mixing can be considered to be a visual representation of the behavior of a chaotic Hamiltonian system. Aref's work, coupled with the fact that a two-dimensional flow is much easier to study in the laboratory than three-dimensional flows, inspired me to check for signs of chaos in a cavity-flow experimental system my students and I had constructed at Amherst in 1983.

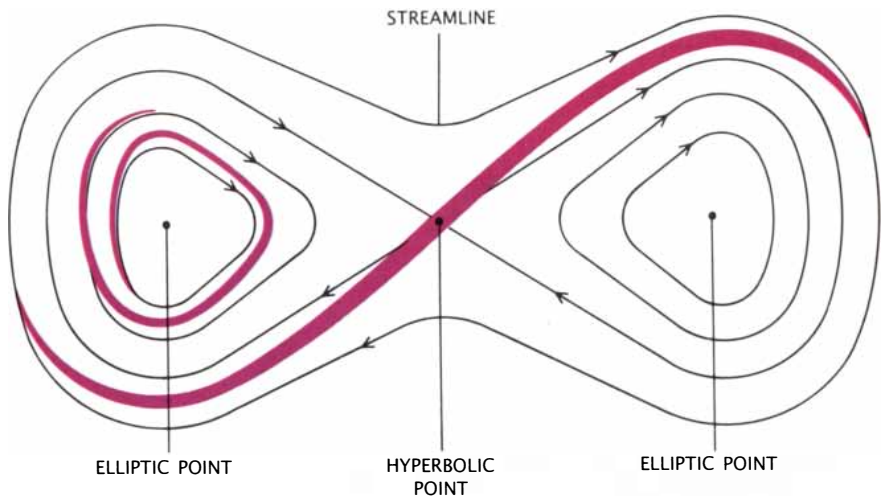
Taking Pictures

Kenny Leong, a graduate student, and I were able to determine the approximate locations of some of the periodic points and large-scale structures in two-dimensional flows by carefully recording stroboscopic images of the system in operation. (Since we are interested in rapid mixing, we concentrated on the behavior of periodic points of low order, say period one, two or three; higher-order points do not take part in the process as often as the lower-order points do.) In a typical experiment we place blobs of fluorescent dye at certain positions in the rectangular cavity, illuminate the cavity with ultraviolet light, set its sides moving in a particular pattern and record the blobs' positions and contortions by taking photographs of the system at regular intervals. If the mixing is effective, the dye particles explore a large region of the system. Conversely, if mixing is poor, the blobs leak dye only slowly into the rest of the fluid or remain close to elliptic periodic points.

In other experiments Paul D. Swanson, another graduate student, and I concentrated on flows that have an exact analytic solution to the equations of motion for the fluid. In this way we can best compare our experimental results with those predicted by theory. Unfortunately the number of systems for which exact analytic solutions are available is rather small, and many are so idealized that they cannot be duplicated by laboratory experi-

ments. One of the systems that does admit exact solution and is amenable to experiments is the flow between two rotating eccentric cylinders. Such a system has also been studied by Aref (who is now at the University of California at San Diego) and by Michael Tabora and Rene Chevray of Columbia University.

Extensive experiments in two-dimensional chaotic flows reveal that the large-scale fluid structures of mixing (such as the positions and shapes of islands and large folds) are quite reproducible; the smallest details of the stretched-and-folded structure are not. The reason is that small deviations in the initial placement of the



ELLIPTIC AND HYPERBOLIC POINTS are typical features of flows in two dimensions. The photograph (*bottom*), made by Leong and the author, shows such a flow, generated as opposite sides of a rectangular cavity filled with glycerine were moved in opposite directions at constant speed. The orange lines (produced by a tracer originally injected along a line extending from the lower left to the upper right corner of the cavity) are nearly aligned with the flow's streamlines, the lines that moving fluid particles follow in steady flows. The flow pattern contains three fixed points: a central hyperbolic point and two elliptic points on each side of it. The flow near each elliptic point (*top*) produces a whorl that rotates clockwise; it increases the length of the tracer linearly with time. Flow in the vicinity of a hyperbolic point approaches the point in one direction and leaves it in another. Because the fluid material cannot cross streamlines, such a steady two-dimensional flow is ineffective in mixing. If the flow is made to vary with time, however, then the stretched filaments of tracer do not have enough time to align themselves with new streamlines and are thereby quickly folded by a change in the direction of the flow.

dye blobs are magnified within the chaotic regions of the flow. That is just as it ought to be: it should be impossible to reproduce any run of our mixing experiments exactly. After all, the objective of mixing is to create randomness. This is precisely what is achieved by the stretching-and-folding mechanisms operating in the experiments.

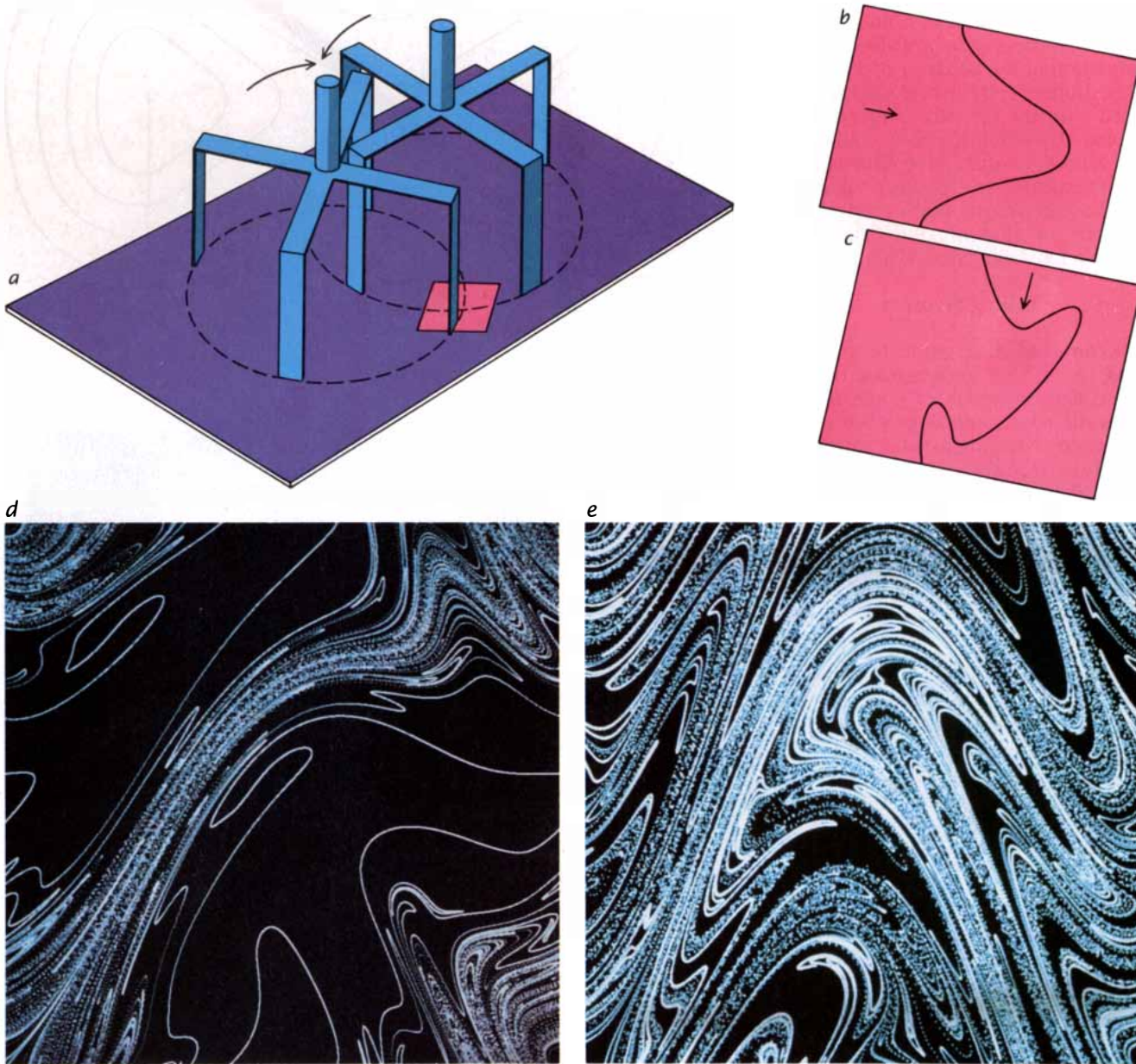
It is also interesting to note how chaos can peacefully coexist with symmetries in the flow, such as those represented by periodic points. In fact,

by systematically eliminating symmetries in a chaotic flow, my co-workers and I have been able to increase the mixing efficiency of the flow.

Experimentation vs. Computation

If the experimental system is rather simple (so that a mathematical expression can be derived for the velocity field), it can easily be simulated on a computer. In a typical program a number of test "particles" are strategically

placed in a simulated motion or velocity field; the computed positions of the particles after 1,000 or so periods then give a good picture of the behavior of the system in general after it has run for a long time. The image generated by this type of simulation is called a Poincaré section, and a Poincaré section that looks complicated is often taken to be computational evidence of chaos [see top of illustration on opposite page]. Computer simulations of mixing also exhibit a form of



EGGBEATER MODEL developed by John G. Franjone and the author serves to illustrate the basic stretching-and-folding process involved in mixing (a). A line drawn on the surface of a fluid cell is stretched and folded as a blade pushes through the fluid first in a direction perpendicular to the line (b) and then in a direction parallel to it (c). The line is stretched without breaking; any parts that extend beyond the top of the cell reenter at the cell's bottom, and parts that extend beyond the

left-hand side reenter the cell at the right-hand side. A computer can generate images of the cell that depend on the number of times the blades have been pushed through the cell. In the images shown a single initial line consisting of 100,000 points has been stretched and folded 16 times under two different mixing conditions. The resultant mixing can be confined to regions of the cell (d) or can cover the entire cell (e), depending on how "vigorously" the blades push through the fluid.

kinematic irreversibility, but in their case it arises from the exponential magnification of errors that are introduced by the computer, which can only handle numbers consisting of a finite number of digits.

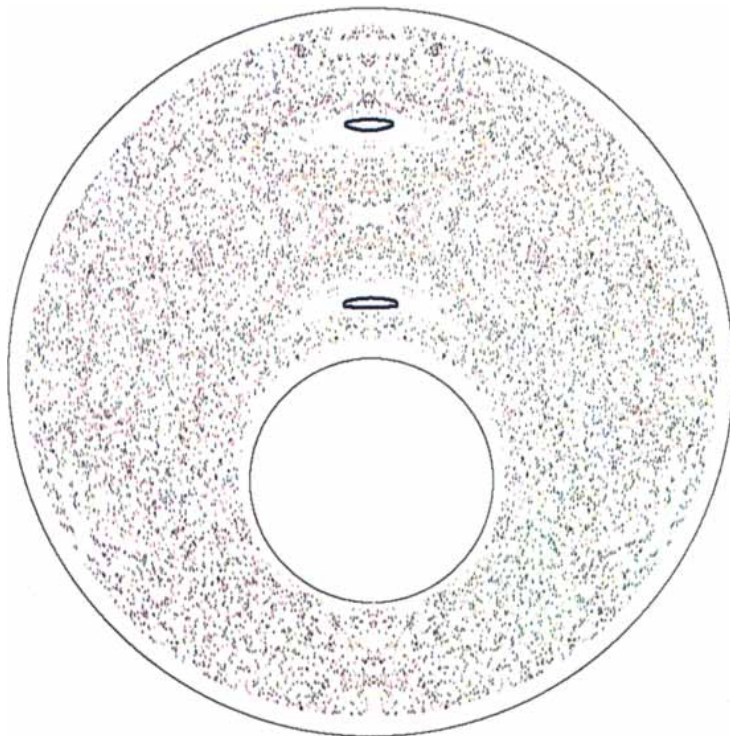
If computer simulations of mixing exist, why bother to do the physical experiments? The first thing that should be kept in mind is that the resolution of the velocity field needs to be much higher for simulations of mixing than for simulations of most other problems in fluid mechanics. Even rather simple velocity fields can produce extremely complicated structures [see illustrations on pages 60 and 61]; in some mixing problems one would like to resolve some of the finer details of the structures.

In a simulation of the rectangular-cavity flow, for example, a conventionally computed velocity field might be too coarse to capture the details of stretched-and-folded striations. It would also be virtually useless for pinpointing the exact positions of the periodic points, which determine the complex behavior of chaotic flows. Moreover, whereas in most fluid-mechanical problems the objective is to obtain an approximation of the velocity field, in mixing the problem starts, rather than ends, with the specification of the velocity field.

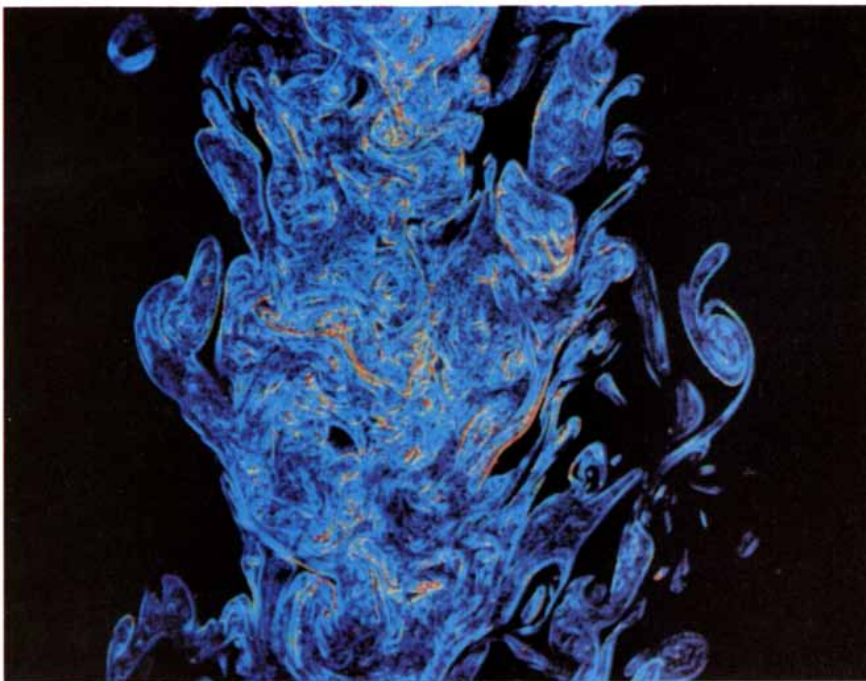
For this reason mixing studies have focused largely on what are in effect caricatures of flows (described by equations that in some cases can be solved exactly) rather than on more realistic problems (whose solution can only be approximated). Indeed, the numerical methods for approximating the solutions of fluid-mechanical equations often introduce spurious effects that do not really exist in real fluid-mixing problems.

Yet even mimicking the simplified flows of our experiments by computational means often results in insurmountable difficulties. A computer treats a fluid as though it consisted of discrete elements. Hundreds of thousands of such elements may constitute a single blob of dye in a simulation, and the number of computations involved in charting its chaotic behavior in a mixing flow can be enormous.

To keep track of all the striations in regions of chaotic mixing in a relatively simple example, such as the one shown in the illustrations on pages 60 and 61, as much as 300 years of computer time in a machine capable of a million floating-point operations per second might be required. To be sure, one might argue that detailed tracking is unnecessary—that it would be bet-



VISCOUS JOURNAL-BEARING FLOW, the flow between two rotating eccentric cylinders, can be modeled on computers. If the cylinders are moved periodically in opposite senses, the flow leads to chaotic mixing, as is seen in the system's Poincaré section for 1,000 periods (*top*) and its stretching map for 10 periods (*bottom*). A Poincaré section is generated by placing a number of colored test "particles" in the simulated flow of a mixing system, calculating their motion for each period and moving them to their new positions. A stretching map depicts the regions where fluid has been stretched in a simulated flow. Most of the stretching takes place within the white regions; little stretching occurs in the colored regions. The stretching map shown is strikingly similar to the structure produced in the actual flow (see cover of *this issue*). The images were made by Paul D. Swanson and the author at Amherst.



TURBULENT FLOW can generate structures that are completely different from those produced in slow, viscous flows. The image, made by K. R. Sreenivasan of Yale University, is a computer reconstruction of a jet of water that has been expelled from a circular nozzle into still water. The flow structures were originally recorded on film by dissolving a fluorescent dye in the expelled water and shining a sheet of laser light along the axis of the nozzle. The intensity of the resulting fluorescence is proportional to the relative concentration gradient of the dye in the water; the image was color-coded to range from deep blue to red according to the magnitude of the concentration gradient. The turbulent flow shown appears to be made up of superposed fractal structures, including several eddies.

ter to account for the stretching in a statistical sense. But would this not be an admission of defeat? If the velocity field (or the motion) can be known exactly, why should one attack the problem statistically?

In sum, new theoretical developments need to be coupled with well-designed experiments, since—more likely than not—brute computational force will not help to answer many questions about chaotic flows. For example, what kinds of motion must the sides of a cavity undergo in order to reduce the size of all the islands in the cavity (including the new ones that might spontaneously appear) below some specified level? The answer to this question might one day make it possible to design a sophisticated pattern-recognition system that can detect the presence of islands in a mixing system and then alter the flow in order to mix the islands with the rest of the fluid.

Limitations and Complications

Before such “smart” mixing machines can be built, however, much more will have to be learned about realis-

tic flows. Although the experiments and computer simulations I have described in this article provide some insights into general problems in mixing (such as how to increase exponentially the area of contact between two fluids), they represent instances of rather specific, idealized problems. The cavity flows described in this article, for example, do not exhibit inertia. In other words, the flow stops as soon as the cavity’s sides stop moving. As a result such flows do not lead to some of the characteristic processes observed in turbulent flows.

To put it in more technical terms, the Reynolds numbers (the ratio of inertial to viscous forces in a fluid) of the flows we have studied in our experiments have been low. Flows characterized by low Reynolds numbers (so-called laminar flows) are orderly and smooth, whereas those characterized by high Reynolds numbers produce rather complex time-varying velocity fields that lead to rapid mixing. An observer at a fixed point in our experimental cavity would see the same simple velocity field repeated periodically rather than seeing the nonperiodic and unpredictable fields

generated in a turbulent flow. Yet it is precisely because of turbulence that it is easier to mix cream in coffee with a spoon (a system that has a relatively high Reynolds number) than to mix two colors of house paint with a spatula (a system that has a low Reynolds number).

Although I have in a sense excluded the most effective mixing flows from this discussion (those that are turbulent), there is reason to believe some of the ideas presented in this article might nonetheless lead to useful concepts for the study of such flows. Slightly more elaborate versions of chaotic two-dimensional flows, for example, display a nonperiodic velocity when they are measured at a fixed point. Clearly, however, much more work is necessary before turbulence can be understood as well as we now understand laminar flows.

In this discussion I have also simplified matters by assuming that diffusion is unimportant in mixing. This in fact is not the case. To take the effect of diffusion in mixing into account, one can apply a simple model stipulating that the rate of diffusion between adjoining striations of two miscible materials is controlled by how quickly the striations are “squeezed” and made thinner, which in turn depends on the component of the flow in a direction perpendicular to the striations. In this way mixing has a double effect that speeds diffusion: it increases the contact area between the fluids while reducing the distance through which the fluids have to diffuse as well as increasing the concentration gradients. Such a model can actually be extended to account for the effect of mixing on such chemical reactions as combustion.

Another common process that—for simplicity—I have ignored is the breakup of droplets in immiscible fluids, which is actually very complex. There are two limiting cases: a low-viscosity fluid dispersed in a body of high-viscosity fluid and a high-viscosity fluid dispersed in a body of low-viscosity fluid. Both cases are difficult to analyze but for different reasons. In the first case the low-viscosity fluid is subject to the bulk of the shear forces, because it cannot effectively transmit the stresses to the droplets of the high-viscosity fluid. Actually a steady shear flow cannot break up a droplet that has a viscosity about four times greater than that of the suspending fluid. Elongational flows are more successful than shear flows in this respect. Yet elongational flows might not be particularly effective in the second

case (in which low-viscosity droplets are dispersed in a high-viscosity fluid), because it is necessary to stretch the droplets considerably before they break up.

My co-workers and I at Amherst have done studies involving the mixing of two fluids of different viscosity in our experimental systems. As might be expected, the amount of breakup is much less within islands than it is in chaotic regions. On the other hand, it could be that too much stirring might

force the droplets to come together; the fluids might sometimes actually become unmixed as a result of their coalescence. Using simple computer models, we have been able to predict the kinetics of such aggregation in simple chaotic flows.

Finally—and most obvious—there is the fact that all our experiments so far have consisted of two-dimensional flows. The real world, of course, is measured in terms of three dimensions. Only recently have my students

and I built the first apparatus capable of producing controlled mixing experiments in three-dimensional flows, and we are beginning to run experiments on it now. There are many fundamental questions regarding mixing in slow three-dimensional flows, and unfortunately some of the intuition we have obtained from our study of two-dimensional flows does not necessarily carry over to flows in three dimensions.

A First Step on a Long Journey

The list of mixing problems does not end here. Mixing of viscoelastic fluids (fluids, such as Silly Putty, that return to their original shape after being deformed) is a formidable problem about which little is known, in spite of the fact that it figures prominently in the processing of polymers having high molecular weight. Mixing of delicate fluids, which are unable to sustain high shear rates without becoming degraded, is important in bioengineering. The mixing of extremely viscous fluids by means of thermal motions is of interest to geophysicists who study the mixing of magmas in the earth's mantle.

Notwithstanding the daunting complexity of mixing processes in nature and industry, there is hope that the understanding can then be fruitfully applied in chemical plants and laboratories. Furthermore, because simple experiments serve as analogues for chaos, they might clarify some fundamental features of chaotic systems in general. Experiments such as those described in this article are a first step in that direction. Only a few of these ideas have been explored so far; there is plenty of room for both basic research and technological exploitation.



STRIATIONS characteristic of mixing in viscous flows are evident in a magmatic rock from the Inyo volcanic chain in eastern California. Indeed, the rock resulted from the mixing of two different magmas, one of which (the one producing the lighter striations) contained minute bubbles of volatile substances. Diffusion across such magma striations is very slow; the time necessary for diffusion to erase striations on the order of a centimeter thick is greater than the age of the earth. The photograph is by Ichiro Sugioka and Bradford Sturtevant of the California Institute of Technology.

FURTHER READING

STIRRING BY CHAOTIC ADVECTION. H. Aref in *Journal of Fluid Mechanics*, Vol. 143, pages 1-21; June, 1984.

FLUID MECHANICAL MIXING—LAMELLAR DESCRIPTION. William E. Ranz in *Mixing of Liquids by Mechanical Agitation*, edited by Jaromir J. Ulbrecht and Gary K. Patterson. Gordon and Breach Science Publishers, 1985.

ANALYSIS OF CHAOTIC MIXING IN TWO MODEL SYSTEMS. D. V. Khakhar, H. Rising and J. M. Ottino in *Journal of Fluid Mechanics*, Vol. 172, pages 419-451; November, 1986.

MORPHOLOGICAL STRUCTURES PRODUCED BY MIXING IN CHAOTIC FLOWS. J. M. Ottino, C. W. Leong, H. Rising and P. D. Swanson in *Nature*, Vol. 333, No. 6172, pages 419-425; June 2, 1988.

Carbohydrates and Depression

Several related behavioral disorders recognized in the past decade are characterized by disturbances of appetite and mood. One of the best-known is seasonal affective disorder, or SAD

by Richard J. Wurtman and Judith J. Wurtman

On May 16, 1898, the intrepid Arctic explorer Frederick A. Cook made the following notation in his journal: "The winter and the darkness have slowly but steadily settled over us... It is not difficult to read on the faces of my companions their thoughts and their moody dispositions... The curtain of blackness which has fallen over the outer world of icy desolation has also descended upon the inner world of our souls. Around the tables... men are sitting about sad and dejected, lost in dreams of melancholy from which, now and then, one arouses with an empty attempt at enthusiasm. For brief moments some try to break the spell by jokes, told perhaps for the fiftieth time. Others grind out a cheerful philosophy; but all efforts to infuse bright hopes fail."

We now know that the members of the Cook expedition were suffering from classic symptoms of winter depression, a condition related to a recently described psychiatric disease known as seasonal affective disorder, or SAD. As the journal entry makes clear, recognition of the association between depression and the onset of winter is not new. But in recent years

there has been growing interest in SAD and in two behavioral disorders, carbohydrate-craving obesity (CCO) and premenstrual syndrome (PMS), that share some of its symptoms. The symptoms include depression, lethargy and an inability to concentrate, combined with episodic bouts of overeating and excessive weight gain; they tend to be cyclic, recurring at characteristic times of the day (usually late afternoon or evening in CCO), month (just prior to menstruation in PMS) or year (generally fall and winter in SAD).

Over the past decade a wealth of information has emerged that casts light not only on the clinical expressions of this group of mood and appetite disorders but also on the disturbed biochemical processes that underlie them. It now appears that these disorders are affected by biochemical disturbances in two distinct biological systems. One system involves the hormone melatonin, which affects mood and subjective energy levels; the other involves the neurotransmitter serotonin, which regulates a person's appetite for carbohydrate-rich foods. Both systems are influenced by photoperiodism, the earth's daily dark-light cycle. Indeed, photoperiodism appears to be the basis for the cyclic patterns of all three disorders.

At high latitudes in the Northern and Southern hemispheres SAD appears in the late fall or early winter and lasts until the following spring. Once expressed, it tends to recur annually unless the patient moves to a place where day length does not decrease significantly in fall and winter. Sufferers complain of episodic bouts of depression combined with profound cravings for carbohydrate-rich foods. They go to sleep early and stay in bed for nine or 10 hours, unlike patients with nonseasonal depression, who have difficulty sleeping. Their sleep, however, is intermittent and not fully refreshing; during the

day they are often drowsy and have trouble concentrating. Once spring arrives SAD patients are full of energy and creativity; they are almost manic in their zest for life. At the same time their craving for carbohydrates lessens and most lose the weight they had gained over the winter.

The following case history typifies many SAD sufferers. Patient M, a 53-year-old teacher, stands five feet four inches tall and weighs 181 pounds. She is unhappy about her weight and over the years has spent a lot of money on short-lived diets. "I know my problem is carbohydrates: when I'm on a diet I stay away from bread, potatoes and sweets and I always lose weight. But when I'm not dieting I get anxious and tense in the midafternoon and I'm unable to concentrate on what I'm doing. I want to eat something to calm myself, so I buy crackers or donuts and nibble on them. At home sometimes I just keep eating until I go to bed." Shortly after Thanksgiving, Patient M experienced two months of feeling tired and depressed. "I told my husband to leave me alone and assigned my pupils problem sets so I wouldn't have to talk to them at school. The house was a mess. I stopped eating except for bread and pasta, but I still gained weight. Finally when spring came I felt better—perhaps because the school year was ending and summer was about to begin."

The symptoms described by Patient M are virtually the same as those associated with CCO and PMS, except that carbohydrate cravers are affected daily, typically in the late afternoon and early evening, and PMS sufferers are affected monthly, during the luteal phase of the ovarian cycle, which lasts for two weeks prior to the onset of each menstrual period.

Interest in seasonal mood disorders was sparked in the early 1980's, when Peter S. Mueller, a psychiatrist at the National Institute of Mental Health,

RICHARD J. WURTMAN and JUDITH J. WURTMAN are respectively professor and research scientist in the department of brain and cognitive sciences at the Massachusetts Institute of Technology. Richard Wurtman, who is also director of M.I.T.'s Clinical Research Center, received his bachelor's degree at the University of Pennsylvania and his M.D. from the Harvard Medical School. His first SCIENTIFIC AMERICAN article, coauthored with Julius Axelrod, appeared in July, 1965. The current article is his sixth for the magazine. Judith Wurtman joined the staff at M.I.T. in 1974. She has a B.A. in zoology from Wellesley College and a Ph.D. in cell biology from George Washington University. The Wurtmans are the editors of a seven-volume series titled *Nutrition and the Brain*.

reviewed data on a 29-year-old woman he had been treating for cyclic bouts of winter depression. Over the course of several years the patient moved to a number of different cities. Mueller maintained contact with her and observed that the farther north she lived, the earlier she became depressed in the fall and the longer she stayed depressed in the spring. On two occasions when the woman traveled to Jamaica in midwinter her depression disappeared within a couple of days of arrival.

Mueller began to speculate that sunlight (or the lack of it) contributed in some way to the woman's depression and decided to experiment with phototherapy (a form of treatment previously shown to be effective in treating jaundiced infants and psoriasis). On consecutive mornings he exposed the patient to 2,500 lux of supplemental, full-spectrum light. (A lux is a unit equivalent to the illumination cast on a surface by one candle one meter away, which is equal to from one-fifth to one-tenth of a foot-candle.) In less than a week the patient had recovered from her depression.

Mueller's findings came to the attention of Norman E. Rosenthal, Thomas

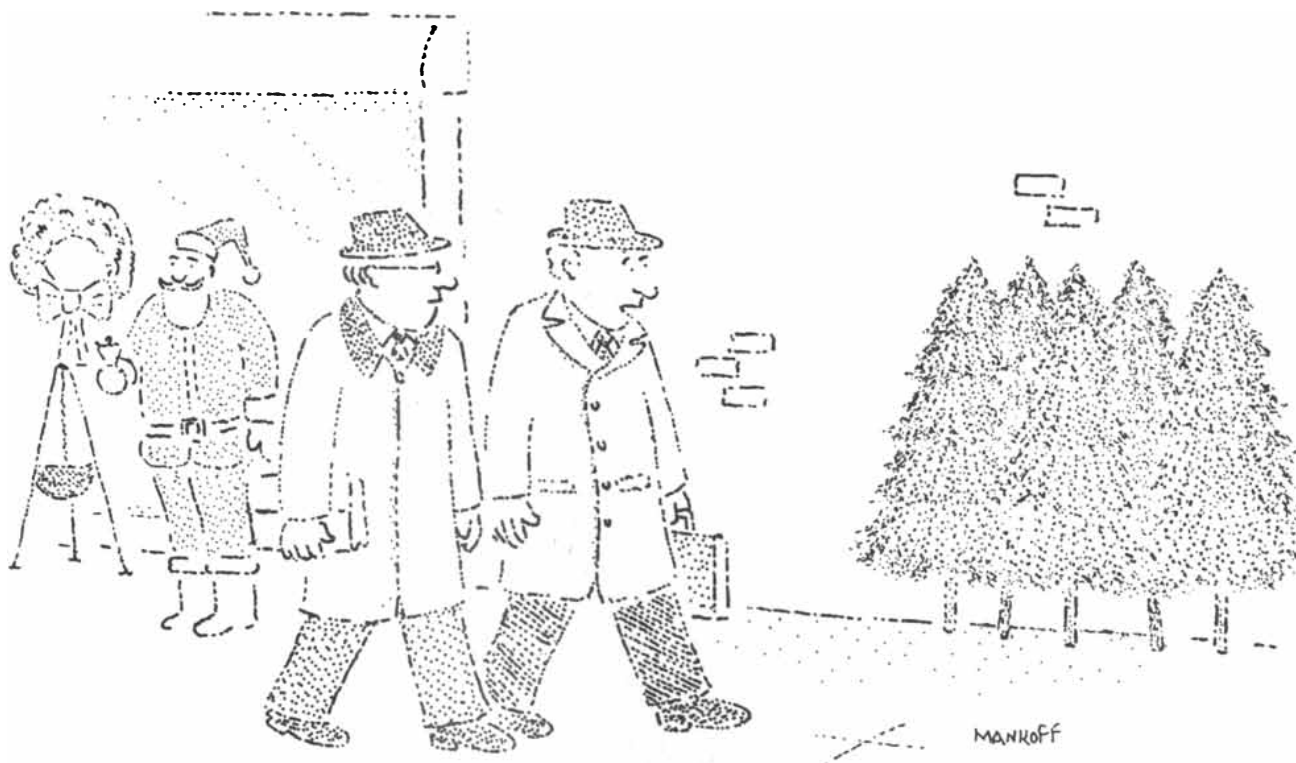
A. Wehr and Alfred J. Lewy, also at the NIMH, who were interested in the various manifestations of clinical depression. They launched a full-scale investigation into the natural history of winter depression, recruiting large numbers of volunteers for observation and treatment. The results were both revealing and intriguing. They confirmed the therapeutic effect of supplemental light in treating winter depression with phototherapy. In addition their data provided the first link between winter depression and carbohydrate craving.

A subsequent study by Steven G. Potkin, Daniel F. Kripke, William Bunney and their colleagues at the University of California at Irvine provided more complete data on the correlation in the U.S. between SAD and latitude. A questionnaire published in the newspaper *USA Today* in March of 1985 provided a description of SAD but omitted any reference to its presumed association with day length. Readers were asked to provide yes or no responses to 15 statements thought to characterize the disease. Those who responded yes to eight or more statements (and thus presumptively had SAD) were asked to send the question-

naire to the authors; 723 did so. The prevalence of SAD in each state was determined by dividing the number of respondents by average daily sales of the newspaper in that state. Results indicated that 100 people per 100,000 in the northern regions of the U.S. are affected by SAD; in the south the incidence is less than six people per 100,000. These estimates, however, are undoubtedly low because people with SAD are less likely to read newspapers and to answer questionnaires than unaffected people.

At about the same time, we began to investigate eating disorders at the Massachusetts Institute of Technology's Clinical Research Center, an inpatient clinic on the university campus. A typical study at the CRC might last for two weeks and focus on carbohydrate consumption among 20 patients in one of two weight groups: moderately obese (from 20 to 39 percent above ideal body weight) and obese (those who are from 40 to 80 percent above ideal body weight).

The eating habits of our study subjects were closely monitored—both at regularly scheduled meals and between meals. Snack intake was meas-



*“Yes, I’m somewhat depressed,
but seasonally adjusted I’m probably happy enough.”*

PUBLIC AWARENESS of seasonal affective disorder, or SAD, has increased in recent years. The spirit of the disorder is cap-

tured in this drawing by Robert Mankoff, which appeared in the December 10, 1984, issue of the *New Yorker* magazine.

ured by a computer-operated vending machine (based on a design by J. Trevor Silverstone of St. Bartholomew's Hospital Medical College in London) that was available around the clock and contained a variety of snack foods ranging from carbohydrate-rich cookies to protein-rich sardines. All the selections contained roughly equal amounts of fat (six grams, for example) and calories (about 110). The foods could be obtained only by typing a special access number into a keyboard connected to a computer that kept a continuous record of the number and type of snacks selected by each patient. Participants in the study were asked to eat as they normally would and not be embarrassed about their caloric intake; most cooperated, believing the data we obtained would eventually help them to overcome their weight problem.

Food consumption during regular meals was measured by giving participants unlimited portions of food in preweighed, labeled containers that were color-coded and set on a table in the dining room. The different foods, like the snacks in the vending machine, varied in their protein and car-

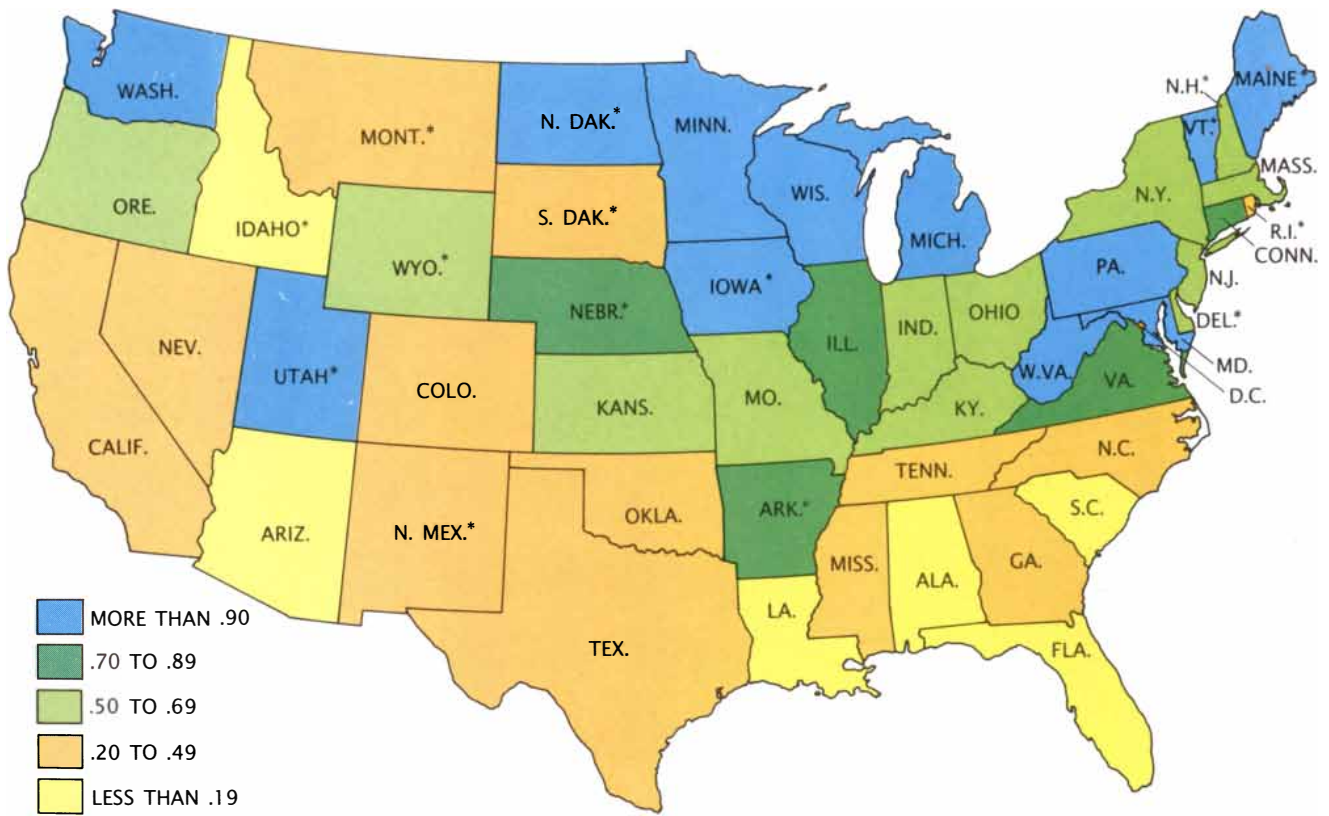
bohydrate content and were equal to one another in fat and calories. At the end of each meal a dietitian reweighed the containers (and their leftovers) to determine how much of each food type a person had eaten.

Our studies at the CRC have enabled us to test and discard a number of myths concerning obesity, specifically with regard to carbohydrate craving. Most prominent among them, perhaps, is the notion that all obese people overeat anything that is tasty, whenever it is available. Instead it appears that those who are carbohydrate cravers overeat only carbohydrates and do so only at characteristic times of the day. At mealtime they behave like normal eaters, consuming a total of some 1,940 calories per day. (An average adult female consumes a total of from 1,500 to 2,000 calories, a male from 2,200 to 2,700 calories.) Toward the late afternoon or early evening, however, the volunteers begin to snack, often consuming an additional 800 or more calories per person per day. A similar pattern has been observed among women with PMS: they increased their snack intake by about 460 more calories per day than wom-

en at the same stage of the menstrual cycle who are not affected by PMS.

We were also intrigued to discover (with the help of the computer-run vending machine) that patients almost invariably underreport their consumption of snacks. It seems that a snack, if eaten quickly, is easily forgotten, as if it somehow "doesn't count." Yet for those concerned about their weight, snacks do count. In some cases they provide 30 percent or more of an individual's caloric intake.

Moreover, we found that most of the snacks consumed by CCO and PMS patients are carbohydrates. We observed, in fact, that more than half of the carbohydrate-craving obese people at the CRC never select a protein snack, although most will readily eat proteins at mealtime. A possible (but still unproved) explanation for such selective eating behavior is that in carbohydrate cravers the ability to regulate nutrient intake is impaired in the late afternoon or early evening. In a non-craver the desire for something sweet is infrequent, noncyclic and readily appeased, say by eating a cookie or two; in a carbohydrate craver, however, the desire may continue unabated



PREVALENCE OF SAD in the U.S. varies with latitude. In a northern state such as Minnesota, SAD affects more than 100 people per 1,000, whereas in Florida it affects fewer than six

people per 100,000. Asterisks indicate a sample that is too small to be reliable. The data were collected by Steven G. Potkin and his associates at the University of California at Irvine.

until nine or 10 cookies are eaten. This suggests there is a malfunction in the feedback mechanism by which the brain knows carbohydrates have been eaten. Another possibility is that carbohydrate cravers snack not because they are hungry but because carbohydrate-rich foods improve their mood.

Why snacking takes place at certain times of day for CCO patients is not clear; its cyclic occurrence, which is monthly in PMS or seasonal in SAD, may reflect the actions of ovarian hormones or melatonin on the brain, but no such relationship has been established for CCO. It is clear, in any case, that carbohydrate snacks tend to exacerbate obesity because they are often rich in fat and thus in calories.

It appears that carbohydrate craving is a multifaceted disorder. As many as two-thirds of all obese people are carbohydrate cravers, but not all carbohydrate cravers are obese; many control their weight by exercising, eating low-calorie meals or satisfying their craving with such low-fat carbohydrates as popcorn (without butter) or candies such as jelly beans. Conversely, not all obesity is linked to carbohydrate craving. Some obese people show no preference for carbohydrates, and some overeat chiefly at mealtime, consuming snacks infrequently.

Our research also focused on mood fluctuations among carbohydrate cravers. When these people were given standardized psychiatric tests based either on an interview (the Hamilton Scale) or a written questionnaire (the Beck Depression Inventory), a high susceptibility to clinical depression was revealed. When carbohydrate cravers were asked why they succumb to foods they know will exacerbate their obesity, their explanation sounded much like the one provided by SAD sufferers. It almost never had to do with hunger or with the taste of the food; instead most said they eat to combat tension, anxiety or mental fatigue. After eating, the majority reported feeling calm and clearheaded. We wondered whether the consumption of excessive amounts of snack carbohydrates leading to severe obesity might not represent a kind of substance abuse, in which the decision to consume carbohydrates for their calming and antidepressant effects is carried to an extreme—at substantial cost to the abuser's health and appearance.

With the help of Harris R. Lieberman and Beverly R. Chew of M.I.T., one of us (Judith Wurtman) set out to test the relation between carbohydrate snack-

ing and mood. Forty-six volunteers, including both carbohydrate cravers and noncravers, were given standard psychological tests before and after eating a carbohydrate-rich, protein-free meal. The carbohydrate cravers were significantly less depressed after snacking, whereas noncravers experienced fatigue and sleepiness. These findings suggest that carbohydrate cravers may eat snacks high in carbohydrates in order to restore flagging vitality, much as some people will pour another cup of coffee when they feel that their energy level or attention span is flagging.

The discovery that one's carbohydrate craving, like SAD, has a distinct periodicity led us to believe photoperiod might in some way be linked to the cyclic manifestations of appetite and mood disorders. We knew from work carried out some 25 years ago that the secretion of melatonin follows a distinct circadian rhythm coupled to daily and seasonal changes in light, which seems to match, at least conceptually, the rhythm most associated with SAD.

Melatonin was discovered in 1958 by Aaron B. Lerner and his colleagues at the Yale University School of Medicine, who isolated it from the pineal



SNACK MACHINE at the Massachusetts Institute of Technology's Clinical Research Center Research has provided data on the food preferences of both carbohydrate cravers and noncravers. The machine contains snacks that have equal amounts of fat and calories but are either rich in carbohydrates or rich in protein. To get a snack a person enters an access code into the machine, which is connected to a computer. The types of snacks and the time they are taken are recorded for each person.

glands of cattle and found that it lightened excised pieces of tadpole skin. Five years later Julius Axelrod and one of us (Richard Wurtman), then at the NIMH, suggested that melatonin was a hormone in mammals, based on its ability to suppress gonadal function when injected into rats. Subsequently we found that melatonin synthesis decreased when rats were exposed to light and that this effect was mediated by interactions among the retina, the brain and special sympathetic nerves that innervate the pineal gland [see "The Pineal Gland," by Richard J. Wurtman and Julius Axelrod; *SCIENTIFIC AMERICAN*, July, 1965].

At about the same time, Wilbur B. Quay of the University of California at Berkeley demonstrated that melatonin levels in the pineal gland of rats exhibit a daily rhythm, rising at night and falling during the day. A few years later Russell Pelham and his colleagues at the University of Pittsburgh described similar fluctuations of melatonin in the plasma of humans. Soon thereafter one of us (Richard Wurtman) and Harry J. Lynch of M.I.T. found that melatonin levels in human urine exhibit pronounced time-dependent fluctuations in samples taken from the same subjects: they are at least five times higher at night than they are during the day.

In order to prove that the timing of melatonin rhythms in humans is affected by the day-night, light-dark cycle, David C. Jimerson of the NIMH, Lynch and one of us (Richard Wurtman) examined the effects of abruptly shifting a person's photoperiod. We recruited a number of volunteers, monitored their plasma and

urinary melatonin rhythms and then changed their photoperiod. We kept them indoors and on the test day left the lights on until 11:00 A.M., shifting the daily dark period 12 hours—to between 11:00 A.M. and 7:00 P.M.

We found it took four or five days for the subjects to reentrain and adjust physiologically to the new light cycle by secreting melatonin when it was dark and suppressing its secretion when it was light. Thus we showed that melatonin secretion follows a circadian rhythm in humans, as it does in other mammals, that the rhythm is endogenous (generated by a clock somewhere in the brain) and that it is entrained by the light-dark cycle.

Neither we nor other investigators were able to demonstrate in humans, however, what Axelrod and one of us (Richard Wurtman) had shown more than a decade earlier in rats: that melatonin secretion is acutely suppressed if subjects are exposed to light during the dark part of the cycle. Perplexed, we concluded that the pineal gland of humans was inexplicably insensitive to the effects of light.

It was not until 1980 that Lewy discovered that melatonin secretion in humans can be acutely suppressed by light—if the light is of sufficient intensity. When the participants in his study were awakened at 2:00 A.M. and exposed to 2,500 lux for one and a half hours, their plasma-melatonin levels declined abruptly. Thus light has two effects on melatonin rhythms in humans, just as it does in rats. It can either reentrain the melatonin rhythm (as when daytime was artificially reversed in our experimental study) or suppress melatonin secretion entirely (if the dark period is eliminated).

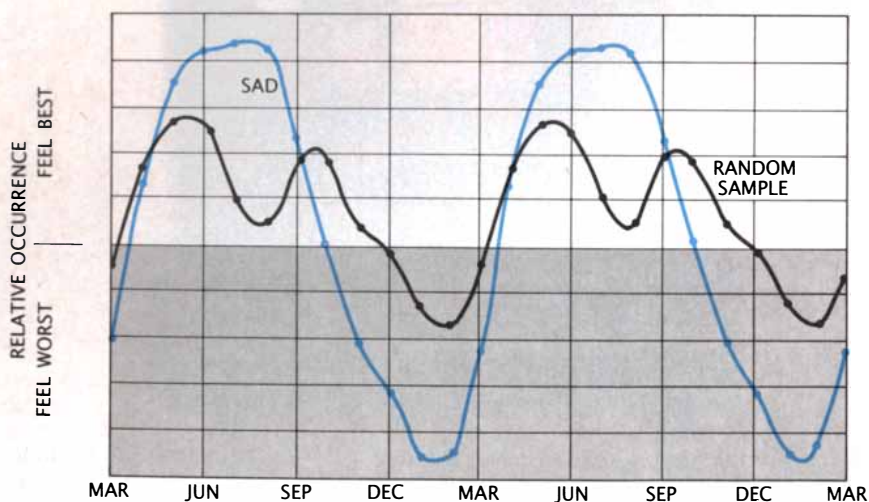
Either action or both could underlie light's therapeutic effect on SAD.

The work of Mueller, Rosenthal and others demonstrated that exposing SAD patients to intense supplemental light for a few hours each morning could eliminate their depression and carbohydrate craving after a few days. Obese carbohydrate cravers have not yet been treated with phototherapy, but a preliminary study by Barbara L. Parry of the NIMH suggests that supplemental light may be effective in treating women with PMS, whose symptoms worsen in the winter.

Michael Terman of Columbia University has found that exposing SAD patients to 2,500 lux for two hours in the morning brings complete remission from both depression and carbohydrate craving in roughly half of them, usually after only a few days of treatment. Most of the remaining patients show some improvement short of full remission. Although his study is not yet complete, Terman thinks it may be possible to enhance the efficacy of treatment by increasing the amount of time patients are exposed to light or by increasing the intensity of the light, say to 10,000 lux. Certainly such levels would more closely approximate sunlight, which ranges from 10,000 lux on a cloudy day in northern Europe to 80,000 lux on a sunny day close to the Equator. Other investigators, however, propose that it is the duration of phototherapy, rather than its timing, that is important in treating SAD. In any case it is now clear that the light must be at least 2,500 lux; customary indoor lights (which range in intensity from 250 to 500 lux) suppress neither the symptoms of SAD nor melatonin synthesis.

Researchers find that light administered in the morning is more effective than light administered later in the day. That finding has been interpreted by Lewy, Terman and others as an indication that light advances a person's circadian rhythm and shortens the dark phase of melatonin secretion. Terman and his associates have noted that the decline in plasma melatonin, which normally occurs early in the morning, is delayed in SAD patients by about two hours. Perhaps high-intensity light induces clinical remission when it is administered in the morning by shortening the daily period of melatonin secretion by several hours.

Is SAD caused by melatonin—either too much of it or when it is secreted for too long? Or is melatonin simply a convenient indicator for another process that underlies the disease? At the moment we cannot answer that ques-



SEASONAL FLUCTUATIONS in mood are common among people in New York City (and in other northern areas) but are severest in patients diagnosed with SAD. The data are from a study by Michael Terman of the New York State Psychiatric Institute.

tion, but circumstantial evidence does suggest a direct link between melatonin and SAD. Lieberman, Lynch and one of us (Richard Wurtman) found that the administration of rather large doses of melatonin to normal individuals induces sleepiness, decreases alertness and slows reaction time. Perhaps the onset of melatonin secretion in the evening is an important promoter of sleep, sensitizing the brain to other sleep-inducing factors. That may explain why SAD patients are hypersomnic in winter, when the daily dark period is almost twice as long as it is in spring. A link between melatonin and mood is also suggested by the ability of oral melatonin to worsen a patient's depression; unfortunately no drug has been developed that selectively blocks melatonin's production or its actions.

But why do patients with SAD, CCO and PMS have a tendency to crave carbohydrate snacks? Why is it that only some people are vulnerable to CCO? And how is it that the brain normally knows when carbohydrates have been or should be consumed? Inhabitants of developed countries habitually eat from 12 to 14 percent of their calories in the form of protein and about three or four times as much in the form of carbohydrates. Even a bear will eventually forsake honey for an occasional fish. How is such a phenomenon regulated? We now know that the answer to these questions involves serotonin, one of the neurotransmitters: substances that are released from a neuron when it fires and that convey the nerve impulse across the synapse to the next neuron.

Serotonin is a derivative of tryptophan, an amino acid that is normally present at low levels in the bloodstream. The rate of conversion is affected by the proportion of carbohydrates in a person's diet: carbohydrates stimulate the secretion of insulin, which facilitates the uptake of most amino acids into peripheral tissues, such as muscle. Blood tryptophan levels, however, are unaffected by insulin and so the proportion of tryptophan in the blood relative to the other amino acids increases when carbohydrates are consumed. Since tryptophan competes with other amino acids for transport across the blood-brain barrier, insulin secretion speeds its entry into the central nervous system, where it enters, among other cells, a special cluster of neurons known as the raphe nuclei. There it is converted into serotonin.

The level of serotonin in turn figures in a feedback mechanism affecting the



PHOTOTHERAPY is effective in relieving the depression and carbohydrate craving associated with SAD. Patients who are exposed in the morning to between 45 and 60 minutes of high-intensity light improve after only two or three days of treatment.

amount of carbohydrate an individual subsequently chooses to eat [see "Nutrients That Modify Brain Function," by Richard J. Wurtman; *SCIENTIFIC AMERICAN*, April, 1982]. When the feedback mechanism is disturbed, as we believe happens cyclically in patients with SAD, CCO and PMS, the brain fails to respond when carbohydrates are eaten, and so the desire for them persists longer than it should.

Serotonin regulates other behaviors too, including mood and sleepiness. Bonnie Spring, now at the University of Health Sciences/Chicago Medical School, found that noncarbohydrate-craving women become sleepy and prone to committing errors following the consumption of a high-carbohydrate lunch (which is expected to in-

crease brain serotonin levels). Similar responses among noncarbohydrate-craving obese individuals were noted by Lieberman and one of us (Judith Wurtman). In contrast, carbohydrate cravers reported feeling refreshed and invigorated after eating a meal rich in carbohydrates.

The mechanisms affecting the relative proportions of carbohydrate and protein in one's diet are most apparent when feedback loops are disrupted, as they are when a patient is given drugs that affect serotonin-mediated neurotransmission. Rats that are allowed to choose between two or more synthetic foods containing different proportions of carbohydrate and protein will normal-

ly alternate between them. If, however, the rats are either injected directly with serotonin (into the brain) or given drugs that enhance the effect of serotonin by promoting its release into nerve synapses, prolonging its activity or stimulating its receptors, then car-

bohydrate intake in experimental rats is selectively reduced.

Drug trials carried out on humans show that a serotoninlike drug, *d*-fenfluramine (which releases serotonin into brain synapses and then prolongs its action by blocking its reuptake into

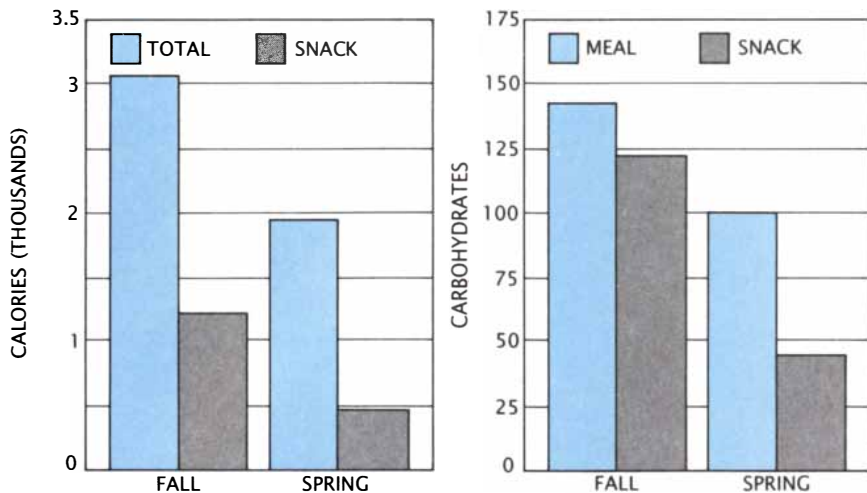
the presynaptic neuron), has a similar effect, selectively suppressing carbohydrate snacking in patients affected by CCO. We also found, while collaborating with Dermot A. O'Rourke, a psychiatrist at the Massachusetts General Hospital, that *d*-fenfluramine can also be effective in treating SAD: it reduces carbohydrate snacking (and its associated weight gain) while simultaneously ameliorating the symptoms of depression. More recently, with Amnon Brzezinski of the Hebrew University-Hadassah Medical School in Jerusalem, we found that *d*-fenfluramine may also be effective in treating similar symptoms in patients with PMS. In 12 of 17 individuals studied, administration of the drug over a six-month period led to a reduction in both carbohydrate craving and depression.

Another disorder, which we think may be linked to serotonin (and thus to SAD, CCO and PMS), is a form of bulimia that is associated with severe bingeing, often on carbohydrate-rich foods, but with little or no vomiting. Most such patients are mildly obese women; many are severely depressed and come from families with histories of depression and alcohol abuse. Preliminary studies by G. F. M. Russell of the University of London and Arthur G. A. Blouin of the University of Ottawa suggest that *d*-fenfluramine can be effective in treating such women; those that respond to the antidepressant effects of the drug are most likely to benefit from its effects on appetite suppression.

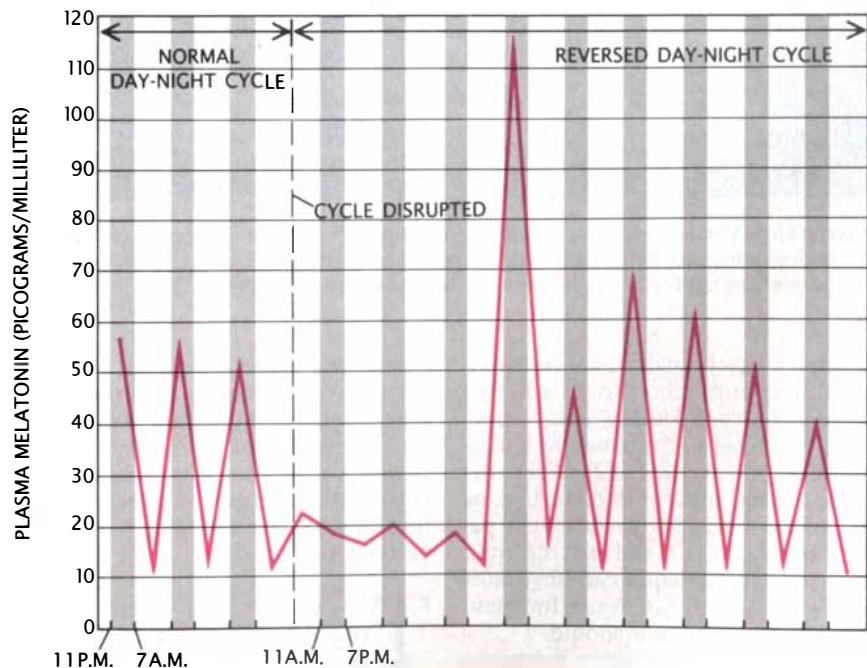
In general we have found that drugs that selectively facilitate serotonin-mediated neurotransmission (such as *d*-fenfluramine, fenoxetine, fluoxetine, zymelidine and fluvoxamine) tend to cause weight loss, whereas drugs that block serotonin-mediated transmission or antidepressants that interact with neurotransmitters other than serotonin have the opposite effect: they often induce carbohydrate craving and subsequent weight gain.

No one could reasonably claim that the symptoms of SAD, CCO or PMS are inconsequential. Prolonged periods of deep depression and irritability can sorely compromise a person's ability to sustain essential human relations. But surely it is not abnormal to feel one's spirits flagging in the fall, to sometimes crave chocolate or pasta, to put on a few pounds every winter or to feel grumpy when beset by menstrual cramps.

Indeed, seasonal changes in behavior afflict normal people as well as those with SAD. Among 200 subjects



PROPORTION OF CALORIES AND CARBOHYDRATES consumed in the form of snacks by SAD patients varies enormously depending on the season. In the fall patients consume more than 3,000 calories per day, of which about 1,200 are from snacks; in the spring their total caloric intake falls below 2,000, of which fewer than 500 come from snacks (left). A similar pattern is apparent for carbohydrate consumption. In the fall almost 50 percent of the carbohydrates eaten per day are in the form of snacks; in the spring the proportion drops to roughly 30 percent (right).



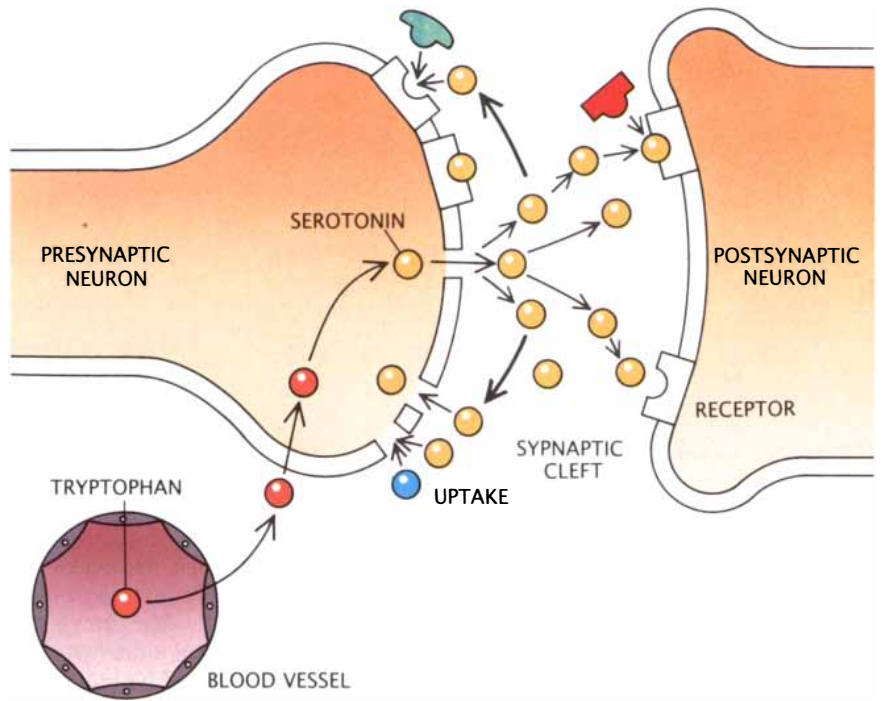
MELATONIN SECRETION follows a daily rhythm in humans, as it does in other mammals. During the day (white columns) secretion of the hormone from the pineal gland is suppressed and levels of melatonin in the plasma are low. At night (dark columns) melatonin is released from the pineal gland and its levels in the plasma are high. If the daily light cycle is abruptly shifted by 12 hours so that the dark period runs from 11:00 A.M. to 7:00 P.M. (instead of 11:00 P.M. to 7:00 A.M.), several days are needed before the new cycle of melatonin secretion reentrains (adjusts) to the new photoperiod. Once adjusted, melatonin secretion again follows a rhythmic pattern.

chosen at random from the New York City telephone directory and surveyed by Terman and his associates, half said they are less energetic in the fall and winter. Forty-seven percent said they gained weight during those months, 31 percent said they slept more and 31 percent said they were not as interested in social activities. Among respondents who reported a decline in energy at some time during the year, about 50 percent said their slump occurs in the fall and winter; only 12 percent said it occurs in summer. Terman concluded that a significant percentage of New York's population suffers from a mild form of SAD; we suspect that the inhabitants of other northern cities, such as Boston or Minneapolis, are similarly affected.

In Tromsø, Norway, which at a latitude of 69 degrees does not see the sun rise above the horizon between November 20 and January 20, midwinter insomnia is thought to affect 24 percent of the population. Charles S. Mullin, Jr., of the U.S. Naval Academy has described widespread sleeplessness, depression, irritability, impaired cognition and the gain of as much as from 20 to 30 pounds among scientists and military personnel who overwinter in Antarctica. The fact that SAD reaches its peak incidence in the Southern Hemisphere in June and July, incidentally, indicates that it is not simply a form of holiday blues or the result of melancholy reflections occasioned by the ending of another year.

Is SAD, then, merely an exaggeration of the normal human response to diminishing light levels in fall and winter? Is it perhaps analogous to hibernation? Probably not. Hibernating animals characteristically lower their body temperature, cease reproductive activity and spend the winter in deep sleep. People with SAD do none of those things; if anything, the time they spend in deep sleep (measured by electroencephalogram) is reduced.

Perhaps contemporary lifestyles increase vulnerability to seasonal depression by diminishing the amount of time we expose ourselves to light. Daniel Kripke and his fellow workers measured the amount of time per day that healthy elderly subjects in San Diego—a region of particularly favorable climate—were exposed to sunlight. Surprisingly, the men were in sunlight for only 75 minutes out of each 24-hour period, the women for only 20 minutes. We need not all live in California, but perhaps most of us need to be exposed to more light, as our ancestors were. Perhaps much as



SEROTONIN regulates carbohydrate consumption. The process begins with the amino acid tryptophan (orange), which circulates through the blood to the brain, where it enters the raphe nuclei. After entering a presynaptic neuron, tryptophan is then converted by way of a two-step process into serotonin (yellow). Serotonin is then released into the synaptic cleft separating the presynaptic neuron from the postsynaptic neuron. Serotonin that reaches the postsynaptic neuron binds to special receptors. Serotonin levels rise in response to carbohydrate consumption. As more serotonin is released, more information is thus transferred to the postsynaptic neuron, where it activates a feedback mechanism. When its concentration is high, serotonin binds to presynaptic receptors, thereby suppressing the release of additional serotonin from the presynaptic neuron. It can also be rapidly removed from the synapse by uptake into the presynaptic neuron. Drugs that enhance serotonin's release (green) or that block its reuptake (blue) increase information transfer across the synapse and diminish carbohydrate snacking; drugs that block postsynaptic serotonin receptors (red) increase appetite, particularly for carbohydrates.

office workers join health clubs to compensate for the lack of exercise, people with indoor jobs need to arrange for adequate exposure to light.

Much remains to be learned about mood and appetite disorders and about the link between serotonin and melatonin. Why does a SAD patient, for example, respond equally well to supplemental lighting, which presumably acts by affecting melatonin, and to drugs that stimulate the release of serotonin? And where might those treatments act in the sequence of pathophysiologic processes leading to SAD? Before we can answer those two questions, it would help to know whether or not light or melatonin has a direct effect on serotonin-releasing neurons. Until we have better answers, we can at least be grateful for the fact that these disorders respond to novel and effective therapies—even if the mechanisms by which the therapies work remain a mystery.

FURTHER READING

SEASONAL AFFECTIVE DISORDER: A DESCRIPTION OF THE SYNDROME AND PRELIMINARY FINDINGS WITH LIGHT THERAPY. Norman E. Rosenthal, David A. Sack, Christian Gillin, Alfred J. Lewy, Frederick K. Goodwin, Yolande Davenport, Peter S. Mueller, David A. Newsome and Thomas A. Wehr in *Archives of General Psychiatry*, Vol. 41, No. 1, pages 72-80; January, 1984.

D-FENFLURAMINE SELECTIVELY SUPPRESSES CARBOHYDRATE SNACKING BY OBESE SUBJECTS. Judith Wurtman, Richard Wurtman, Sharon Mark, Rita Tsay, William Gilbert and John Growdon in *The International Journal of Eating Disorders*, Vol. 4, No. 1, pages 89-99; February, 1985.

ON THE QUESTION OF MECHANISM IN PHOTOTHERAPY FOR SEASONAL AFFECTIVE DISORDER: CONSIDERATIONS OF CLINICAL EFFICACY AND EPIDEMIOLOGY. Michael Terman in *Journal of Biological Rhythms*, Vol. 3, No. 2, pages 155-172; 1988.

The Hunt for *Proconsul*

Sixty years after its discovery, Proconsul is now known to be the last common ancestor of great apes and human beings rather than an extinct ancestor of the chimpanzee and the gorilla

by Alan Walker and Mark Teaford

The prehistoric ape *Proconsul* is now the best-known of our ancestors, yet its route from the obscurity of an excavation pit to fame in the scientific spotlight is as full of twists and surprises as any soap opera. It is a story of implausibilities, in which various pieces of important specimens, once unearthed, became separated and sent to distant lands until fortune brought them together again decades later. It is also a tale with a happy ending; recent expeditions to the excavation sites have revealed nearly 800 new specimens of hominoid primates—the superfamily of primates that includes the great apes, the lesser apes and human beings. These have vastly increased the sample of *Proconsul* fossils, and the new finds show that *Proconsul* is a useful model of the last common ancestor of the great apes and man.

The story began in 1927 when H. L. Gordon, a settler in western Kenya, found some fossils while digging limestone from a quarry. Thinking they might be important, Gordon had them sent to the paleontologist A. Tindell Hopwood of the British Museum. One

of the fossils was apparently nothing more than a single tooth that protruded from a palm-size nodule of rock. When the stony matrix was removed, however, the specimen proved to be the left maxilla, or left upper jawbone, of a hominoid primate. The remaining fossils from the quarry indicated that the deposit was about 18 million years old, from the Lower Miocene epoch.

Fossil apes were hardly known at the time and none had been discovered of such antiquity, so that these specimens were extremely important. Nevertheless, Hopwood decided not to publish until he had more evidence that he had found a new primate. After raising money for an expedition, he went in 1931 to Kenya, where he succeeded in collecting additional hominoid fossils. Two years later Hopwood published his findings and stated his conviction that the Gordon jawbone was that of a new genus ancestral to the chimpanzee.

At the time London's vaudeville patrons were being entertained by a chimpanzee that wore a suit and cap, rode a bicycle and smoked a pipe. The chimp's name was Consul, and in a bit of scientific whimsy Hopwood named the new anthropoid ape after him: *Proconsul africanus*.

The next chapter of the story was written by Louis and Mary Leakey, who made a series of expeditions to western Kenya in the 1940's and early 1950's. On Lake Victoria's Rusinga Island, Mary Leakey discovered in 1948 what would become the most famous *Proconsul* specimen—a skull. When she found it, on a slope of soft sedimentary rock, the back of the skull was exposed to the elements and parts of it had eroded away. The face and jaws were nearly complete but the back and sides were represented only by fragments recovered from the lower slopes of the incline.

The 1948 skull was assumed to belong to the same species as the jawbone found by Gordon and described

by Hopwood. Other discoveries by the Leakeys, however, suggested that there were two species of *Proconsul* on Rusinga Island and on nearby Mfangano Island: a large, chimpanzee-size species called *Proconsul nyanzae* and the smaller *Proconsul africanus* represented by Hopwood's jawbone and the 1948 skull.

The next important specimen was found in 1951 by geologist Tom Whitworth, who was surveying the Kiakanga area of Rusinga Island. In a vertical pipe of green rock, which cut through gray silts that did not contain fossils, he found a pig skeleton and other bones. Among those that were chiseled out of the very hard matrix were pieces of a subadult (late juvenile) skull, much of a forelimb and hand and bits of a foot—all from a single *Proconsul*.

At the time, the pipe of green rock, composed of coarse, volcanic ash once laid down by water, was thought to be a large, river-cut "pothole" into which animals had been washed and later fossilized. If that was the case, investigators would be faced with two problems. First, because the bones and skeletons could be washed in from any distance upstream, the assemblage of animals in the pipe might be quite different from the actual community in which *Proconsul* lived. Second, since a pothole must be cut by water into preexisting compacted sediments, the filling of the pothole might have occurred much later than the deposition of the 18-million-year-old surrounding rock.

The 1948 and 1951 finds have recently "resurfaced" to take on new roles in the story of *Proconsul*. The revival began a few years ago when Martin Pickford of the Institute of Paleontology in Paris noticed an entry in Louis Leakey's field notebook for 1947. The entry referred to fragments of possible primate skull bones picked up at the same spot where

ALAN WALKER and MARK TEAFORD have worked together since 1984 in excavating early Miocene ape fossils in Kenya. They are respectively professor and assistant professor in the department of cell biology and anatomy at the Johns Hopkins University School of Medicine. Walker received his Ph.D. in anatomy and paleontology from the University of London. He collaborates frequently with Richard E. F. Leakey and other colleagues at the National Museums of Kenya in the study of primate evolution and is a 1988 MacArthur fellow. Teaford got his Ph.D. in anthropology from the University of Illinois at Urbana-Champaign in 1981. He has concentrated on the functioning and wear of primate teeth and the diet of extinct animals. In these studies he has made extensive use of the scanning electron microscope.



FIRST *PROCONSUL AFRICANUS* SKULL is shown at the top in full frontal view and profile. It was found (with pieces missing) by Mary Leaky in 1948 on Lake Victoria's Rusinga Island. Additional pieces were added to the skull in 1981 by Martin Pickford and one of the authors (Walker), who discovered them

mixed with turtle-bone fragments. The brown bones in this *Proconsul* hand (*bottom left*) and foot (*bottom right*) were found on Rusinga Island in 1951 by Louis Leakey's team. The white bones were found 30 years later in the National Museums of Kenya. The color difference is due to different preservatives.

Mary Leakey found the skull a year later. Pickford was quick to see that these might have been the parts of the skull that had been washed away by erosion. He managed to trace the pieces to a collection of turtle fragments stored in the National Museum, Nairobi. They were indeed the missing pieces from the back of the *Proconsul* skull.

Pickford and one of us (Walker) managed to glue them back onto the original, making the cranium complete from the snout, over the top of the skull and to the foramen magnum (the large hole in the base of the skull) underneath.

With the more complete skull we hoped to acquire a piece of information that is important to evolutionary biologists: how encephalized, or relatively "brainy," *Proconsul* was. The degree of encephalization is given by the ratio of brain volume to body weight, and the *Proconsul* skull enabled us to make a good estimate of the cranial capacity of a Miocene hominoid. To estimate cranial capacity, one normally measures the volume of water displaced by a cast of the skull's interior. Such a procedure obviously requires an undistorted skull; unfortunately the *Proconsul* specimen had been somewhat flattened and folded, that is, squashed.

Nevertheless, Pickford, Dean Falk of the State University of New York at Albany, Richard J. Smith of Washington University and one of us (Walker), devised a simple way to make the estimate. Because the skull is made from an inelastic material, the lengths of the arcs along the skull's interior were unchanged. Moreover, the gener-

al shape of the *Proconsul* brain is similar to that of catarrhine monkeys, also called Old World monkeys, and we had at our disposal a set of casts of the insides of Old World monkey craniums. Measurement of the cranial arcs on the casts showed a statistical relation between arc length and cranial capacity of Old World monkeys. We were able to measure the arc length of the *Proconsul* skull from the front of the skull to the posterior edge of the foramen magnum. Assuming that the same relation held for *Proconsul* as for catarrhine monkeys, we concluded that the cranial capacity of the fossil skull was between 154 and 180 cubic centimeters, with a best guess of 167 cubic centimeters.

As for body weight, it can be estimated from various measurements of limb bones. By making the appropriate measurements on the 1951 limb bones and on some others at our disposal, we could conclude that *Proconsul africanus* was more encephalized than modern monkeys of comparable size. We suspect that pronounced encephalization is a characteristic of modern great apes, although we cannot be certain. Modern great apes, which include the orangutan, chimpanzee and gorilla, are much larger than extinct species, and although brain size gets absolutely larger as body size increases, it gets relatively smaller. If, however, encephalization is a great ape trait, it developed very early in their evolutionary history.

The 1948 skull has recently revealed another connection between modern and ancient primates. Sir Wilfrid Le Gros Clark, who originally described the skull, noted that it had a frontal air

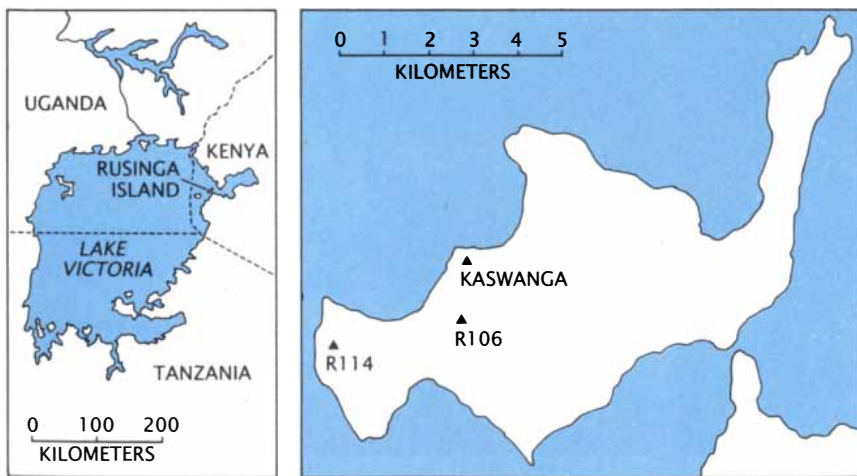
sinus, or air space within the frontal bone of the skull. This is a feature of some importance, because frontal sinuses are found in humans and African apes but not in orangutans, Old World monkeys or lesser apes (the siamang and gibbon). Roughly three years ago Steven C. Ward of Kent State University was doing a study of the size and shape of facial sinuses in higher primates. During the study he examined the *Proconsul* skull. Unfortunately the original sinus region had been filled in with plaster of Paris and could only be seen by X rays. Because of the peculiar fossilization in Rusinga Island sediments, however, the method failed to distinguish the rock matrix that filled the sinus cavity from the bone and plaster. Consequently Ward was not able to verify Le Gros Clark's observation.

Recently one of us (Walker) was able to remove the plaster of Paris. A particularly large matrix-filled frontal sinus was exposed that extended far to the posterior; this confirmed that these Miocene apes had at least some affinities with modern-day great apes, as opposed to lesser apes and Old World monkeys.

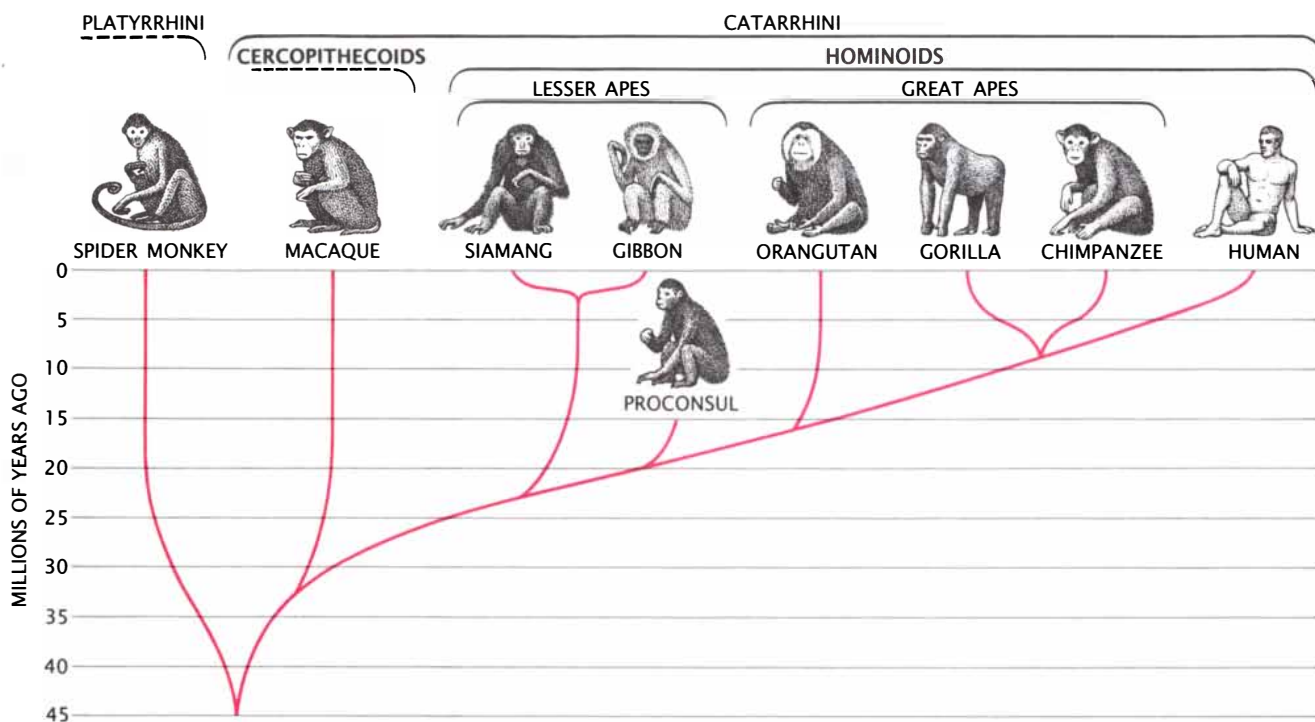
Even more convoluted is the story of the 1951 *Proconsul* specimen. When it was originally discovered at a site called R114, the bones were excavated from blocks of the hard green pothole rock and sent to Le Gros Clark, who passed them on to John R. Napier. Napier and his colleague Peter R. Davis, both of the Royal Free Hospital in London, used the bones to develop an outline of primate limb and hand evolution and wrote a classic monograph on the subject. In 1964 Napier returned the bones to Kenya.

They were reunited with *Proconsul* by a fortuitous event. In the early 1980's a collection of Miocene fossil pigs that had been on loan to a paleontologist was returned to the National Museums of Kenya. The collection included a fossil pig skeleton from the pothole as well as a block of green rock containing a number of articulated bones also thought to be those of a pig. The pig specialist, however, had recognized that the latter specimens were not pig bones and had set them aside.

During a visit to the museum in 1980 one of us (Walker) examined the bones and realized that they were the lower leg and foot bones from the same *Proconsul* individual described by Napier and Davis. The identification was facilitated by the fact that the



RUSINGA ISLAND has been the site of many *Proconsul* finds. Mary Leakey found the 1948 skull at the site labeled R106. The 1951 specimen was found at site R114 by Tom Whitworth. The Kaswanga site, discovered during the authors' first expedition, yielded hundreds of primate fossils, including at least nine *Proconsul* skeletons.



EVOLUTIONARY TREE of hominoids is shown. Hominoids are the superfamily of primates that includes the lesser apes (siamangs and gibbons), great apes (chimpanzees, gorillas and orangutans) and humans. *Proconsul* lived 18 million years ago

during the Lower Miocene. Although it shares some features with modern gorillas and chimpanzees, most of these features are relatively unspecialized, suggesting that *Proconsul* represents the last common ancestor of all great apes and humans.

individual was a subadult: many of the epiphyses—the ends of a growing bone separated by growth cartilages from their shafts—had yet to fuse with the bones proper, and this was exactly the state of Napier's specimen. The new bones also allowed investigators for the first time to see the proportions of *Proconsul's* hand and feet, which had both ape- and monkeylike features, and showed that *Proconsul* was a slow-moving quadruped.

The return of the pig block led Pickford and one of us (Walker) to suspect that other *Proconsul* bones might be misplaced in the Kenya museum drawers. Indeed, when every bone from the excavation site had been retrieved, more *Proconsul* pieces came to light: a scapula, parts of a humerus, additional hand bones, ulnas, tibias and bits of both femurs. It was now possible to estimate the limb proportions of the animal, which again revealed both apelike features (the forelimb) and monkeylike features (the leg).

If ape and pig bones could be mixed together in a museum drawer, the same might be true at the original excavation. Pickford and one of us (Walker) therefore set out to search for overlooked *Proconsul* bones at site R114, which had lain untouched and overgrown since 1951. They found the

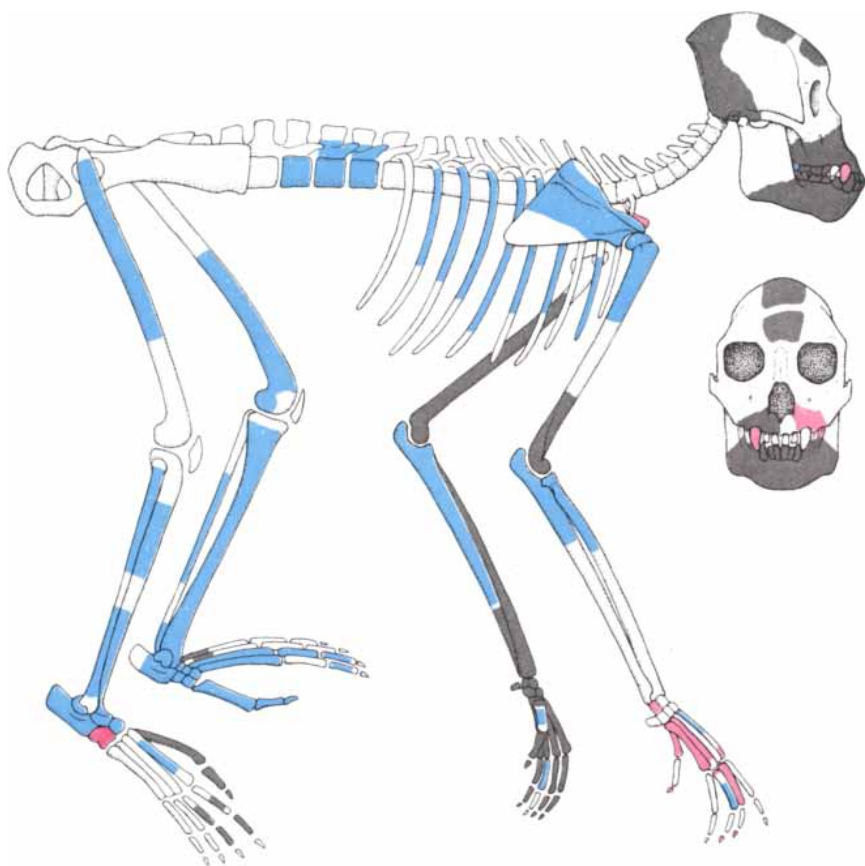
site but did not have time to excavate. Excavations at R114 in 1984 by us were rewarded with the missing maxilla, half a clavicle, a canine tooth, an important bone from the foot and a block containing bones from the right hand, not to mention a number of other mammalian specimens. Thus, after more than 30 years of separation, the various and diverse fragments of *Proconsul africanus* were reunited from their previous residencies in the pothole and mislabeled museum bins. The "1951" skeleton is now the most complete of any large-bodied Miocene hominoid.

The 1984 expedition to R114 also revealed important information about the structure of the pothole itself, which in turn had repercussions for understanding the paleoecology of *Proconsul*. Several weeks of cutting through the sedimentary rock next to the pothole showed that it was at least four meters deep. The individual layers of silt did not lie symmetrically around the pothole but had apparently been deposited unevenly around an object standing where the rock pipe is today. The small geological faults on each side of the pothole verified this suspicion: something had been in the way of the stream-carried silts as they were laid

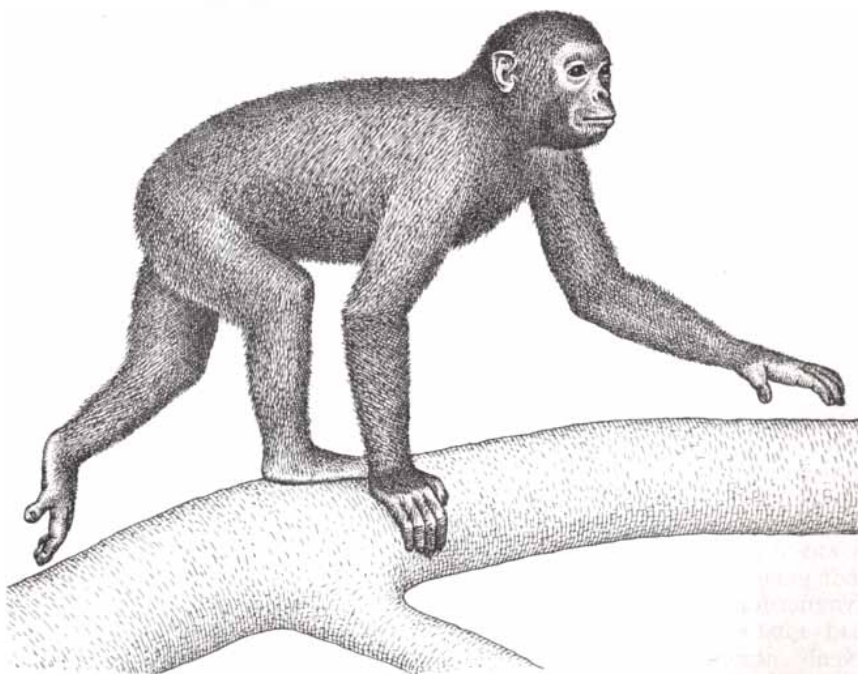
down, and it eventually collapsed, leading to faulting in the silts. This hypothesis also explained the asymmetric deposition of the silt: more pebbles and grit were deposited on the upstream side than on the downstream side.

Eventually it became clear that the pothole was not a pothole at all. Rather, a large tree had stood in this position 18 million years ago and had become partially buried by silts and sands. The tree withstood being buried, but it probably died as a result and became hollow soon thereafter. Once it was hollowed out, monitor lizards, pythons, bats and small carnivores began to occupy it and left their bones—and those of their prey—in the trunk. One of the carnivores probably caught the *Proconsul* and carried it into the tree to eat; in fact, certain joints of the *Proconsul* specimen display tooth marks. With time the tree was filled with a mixture of bones and sediment, which eventually solidified to form the pipe of rock that has come down to the present.

The solution to the pothole mystery solved two major problems. First, it is now certain that the green filling and the fossils in the deposit are, geologically speaking, of the same age as the surrounding gray silt and sands; the latter have been dated by applying



PROCONSUL SKELETON has taken 30 years to piece together. The parts found at its initial discovery by Whitworth in 1951 are shaded in gray. The parts colored in blue were found in collections at the National Museums of Kenya in 1984 by Pickford and Walker. The parts colored in red were found in 1984 by the authors at R114.



PROCONSUL is shown as it might have looked 18 million years ago. The smaller of the two species that lived on Rusinga Island was the size of a female baboon, with fore and hind limbs of about equal length. It was relatively slow-moving and probably had not developed specializations for leaping, arm swinging, knuckle walking or ground living. The larger species weighed about four times more than the smaller.

the potassium-argon method to the lava above and below them and have been found to be about 18 million years old. Second, because the tree contains fossils of animals that used it as a roost or refuge, this assemblage of fauna clearly represents the local community in which *Proconsul* lived.

Excavation of the R114 site led by accident to another major find. One of the frequent thunderstorms over Lake Victoria filled the excavation at R114 to the brim with water. Unable to work on the pothole, the National Museums of Kenya crew, led by Kamoya Kimeu, took the opportunity to prospect in nearby areas. The side trip turned out to be a fruitful one: a new site was discovered that has since revealed hundreds of whole bones and thousands of bone fragments. Indeed, the new site, now called the Kaswanga Primate Site, proved to be so spectacular that we negotiated an agreement with the chief of Rusinga Island to keep the area from being turned into a maize field for five years.

Among the important finds were at least nine whole or partial *Proconsul* skeletons that had been washed into a small gully in the ancient Miocene terrain. They represent apes ranging in age from very small babies to adults; probably both males and females are represented. By now practically every part of the *Proconsul* skeleton is known from one or more of the individuals.

Unfortunately most of the Kaswanga bones were cracked or splintered by the swelling of the clays and silts in which they were embedded or by plant roots. This has forced us to assemble the specimens by gluing, a process that is continuing at the present time. Although the work has been quite successful with the larger bones, it has been difficult to differentiate small cylindrical fragments of major infant limb bones from the finger and toe bones of juveniles and adults.

The fossils from Kaswanga have enabled us to calculate the proportions of *Proconsul's* hands, feet and limbs; when our gluing is complete we shall also be able to trace the various stages of *Proconsul* growth. Because the analysis of the new bones will take some years, we have chosen to concentrate first on those bones that have been the subject of previous research. In doing so we have already been able to resolve a controversy surrounding the functional anatomy and evolutionary meaning of the *Proconsul* wrist.

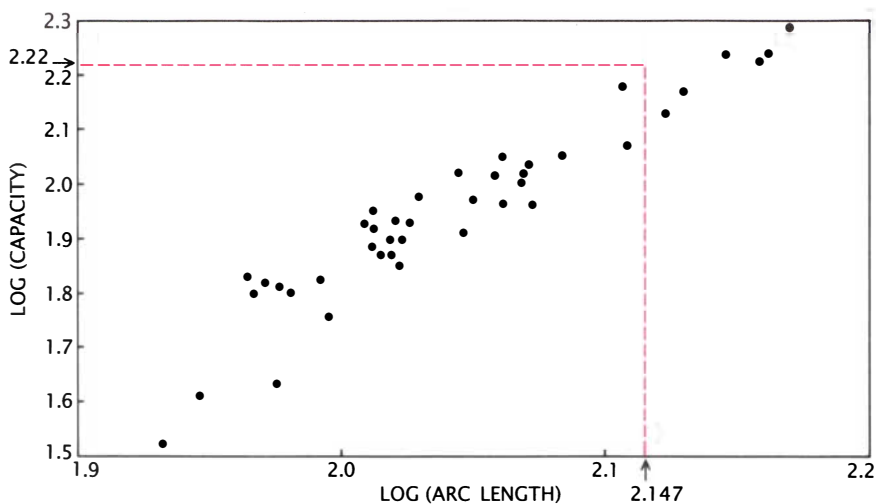
The wrist of apes and humans differs from that of most mammals, in-

cluding Old World monkeys: it does not have a proper articulation between one of the forearm bones—the ulna—and the wrist proper. In Old World monkeys, on the other hand, a small, bony extension from the ulna rests in a socket formed from two bones of the wrist, the pisiform and the triquetrum.

Among the bones of the 1951 *Proconsul* skeleton were those of the left hand, which gave rise to numerous papers, not to mention several Ph.D. theses. The triquetrum bone was badly damaged, however, leading to disagreement over the nature of *Proconsul*'s wrist articulation. As a result of the Kaswanga finds we have a number of new specimens of the pisiform and triquetrum. Working with K. Christopher Beard of Johns Hopkins University, we have been able to demonstrate that the *Proconsul* and Old World monkey wrists are similar in that there is a direct articulation between the wrist and the ulna. Although the *Proconsul* wrist is primitive in this one respect, in other respects it foreshadows the greater mobility found in certain parts of the wrist in present-day hominoids.

This "hybrid," or mosaic, pattern is often evident as one analyzes Miocene hominoids. As Michael D. Rose of the University of Medicine and Dentistry of New Jersey in Newark has pointed out, Miocene hominoids are neither like Old World monkeys nor like great apes; rather, they are like Miocene hominoids. The unique combination of characteristics makes functional interpretation of *Proconsul* fossils difficult: not only are there no modern animal models of *Proconsul*'s anatomy but also each of *Proconsul*'s anatomical complexes exhibits unique combinations of features. For example, some of the ankle bones of the foot are slender and monkeylike but the big toe is robust and apelike. The same hybrid pattern is found in the *Proconsul* pelvis: the ilium, or upper portion, is like that of Old World monkeys, whereas the acetabulum (the site of articulation with the head of the femur) is large and shallow, like that of the great apes.

Investigators are now coming to appreciate, from several lines of evidence, that at some point ancestors of Old World monkeys must have spent more time on the ground than their modern descendants do. This in turn has led to the realization that many of the features *Proconsul* shares with modern great apes are primitive features that have merely been retained in modern great apes for millions of years. Such unspecialized features are



CRANIAL CAPACITY of contemporary Old World monkeys is related to arc length of skull as is shown. Direct measurement of *Proconsul*'s brain capacity was impossible to obtain because the 1948 skull was squashed (see illustration on page 77). The arc lengths along the skull, however, were not changed by the deformation. Moreover, the *Proconsul* brain is similar in shape to that of Old World monkeys, so that one might assume the same relation holds between arc length and cranial capacity. The 1948 skull had an inner arc length from its front to the foramen magnum (the large hole in the base of the cranium) of about 140 millimeters. Plotting the logarithm of that value on the horizontal axis enables one to read off from the vertical axis this *Proconsul*'s approximate cranial capacity: 167 cubic centimeters.

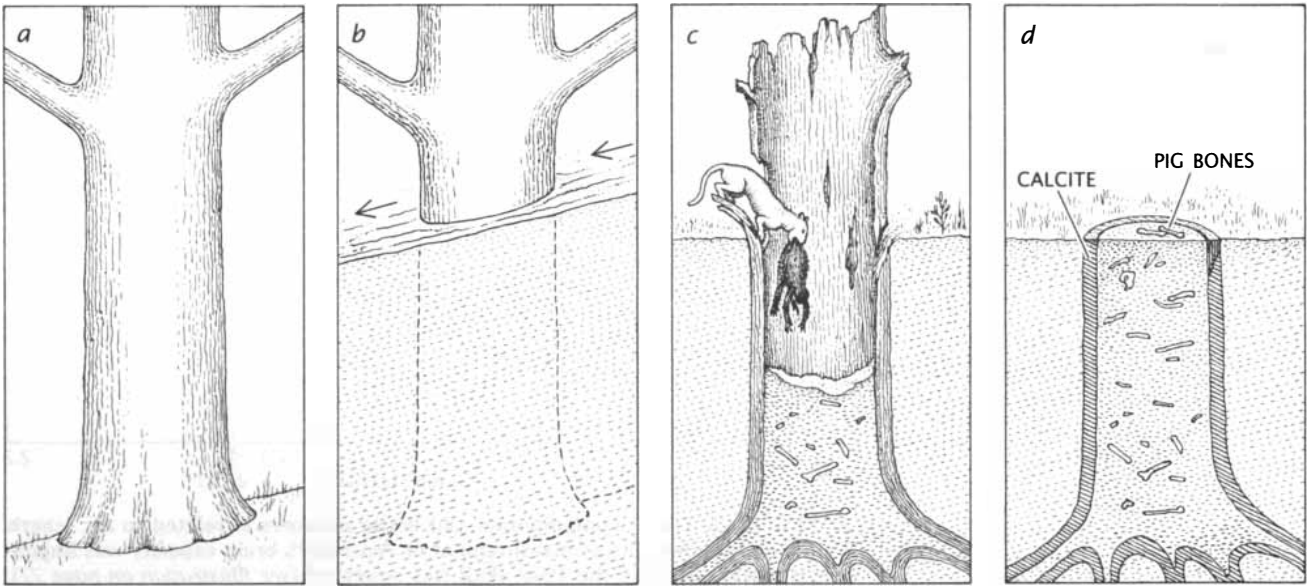
not signs of close evolutionary affinities. (After all, both humans and opossums have five fingers.) The pelvis of *Proconsul*, for instance, shows that these creatures did not have ischial callosities, the roughened buttock pads on which gibbons and Old World monkeys sit. The lack of callosities, however, is a primitive condition; the presence of these structures is a specialization that serves to unite gibbons and Old World monkeys, but their absence does not link *Proconsul* closely with chimpanzees, which also lack ischial callosities. The chimpanzee has simply retained its ancestral condition, which dates from times earlier than the time of *Proconsul*. On the other hand, the frontal sinus is present in *Proconsul*, humans and great apes but not in gibbons and monkeys. Such shared specializations indicate that *Proconsul* is more closely related to modern great apes than it is to modern monkeys.

Our recent excavations have also shed light on the number of *Proconsul* species that lived on Rusinga and Mfangano islands. Several investigators have agreed for some time that the small species of *Proconsul* found on the islands and represented by Mary Leakey's 1948 skull is in fact not the same as the original *Proconsul africanus* discovered at the mainland site and represented by Hopwood's jawbone. If this point of

view is correct, there must be at least three species—two small ones and the larger *Proconsul nyanzae*, which is also found on the islands. Other investigators such as Pickford, Jay Kelley of Brown University and David Pilbeam of Harvard University, have argued that there are only two species—the small island specimens are actually the females of the larger *Proconsul nyanzae*.

Together with Christopher B. Ruff of our department at the Johns Hopkins University School of Medicine, we have been able to test the idea by estimating the body sizes of both the large and the small specimens and comparing the differences found with those actually documented for males and females of living primates. Among the hundreds of new fossils from the Kaswanga site are several femurs of both *Proconsul* sizes. Ruff has found that he can estimate the body weight of quadrupedal primates accurately from measurements of the bone distribution in cross sections of the femur. Two large and two small complete *Proconsul* femurs (which, as far as we can tell from all the specimens on hand, are quite representative) give body-weight estimates of 37 and 9.6 kilograms for the two sizes. In other words, the proposed females were only about one-fourth the weight of the proposed males.

No living terrestrial mammals, let alone primates, show such extreme



"POTHOLE" in which 1951 *Proconsul* was found turns out to be the fossilized remains of a tree trunk. During the Lower Miocene a tree stood by a river (a) and was gradually buried by gray sediments (b). The tree rotted and became hollow

(c), whereupon animals used it as a home. Dead *Proconsul* was probably carried into the trunk by a creodont. The trunk filled up with bones and green sediment that encased the fossils until 1951 (d). Calcite eventually replaced the tree bark.

sexual dimorphism in body weight. This has led us to reject the idea that the large and small forms of *Proconsul* on Rusinga Island represent males and females of the same species. We conclude that two species of early ape lived on Rusinga and Mfangano islands some 18 million years ago: a small species that differed from Hopwood's mainland species, and the larger *Proconsul nyanzae*. They seem to have been similar to each other in shape and proportions but very different in body size.

The similarities in shape and proportions point to important similarities in posture and locomotion. In many respects the island *Proconsul* specimens display a primitive pattern in their limb proportions and joint surfaces; their fore and hind limbs, for instance, were of similar length. On the other hand, they were also both tailless, a specialized condition. In their classic monograph Napier and Davis concluded, based on an analysis of just the forelimb bones, that *Proconsul* was an active, leaping quadruped, moving rather like today's langurs of Asia and having some bony features that are indicative of arm-swinging habits.

The new fossils show, on the contrary, that *Proconsul* was a relatively slow-moving, probably cautious, arboreal species that had no obvious specializations for leaping, arm swinging, knuckle walking or ground living.

The status of *Proconsul* has changed considerably in the 60 years since Gor-

don found the first jaw fragment. Hopwood thought *Proconsul* was ancestral to the chimpanzee, and this idea was extended even further in the 1960's and 1970's by some anthropologists who saw the different species of *Proconsul* as ancestral to the different species of modern great apes.

Over the past decade, however, a wealth of new material has been discovered, and not just on Rusinga Island. Richard and Maeve Leakey of the National Museums of Kenya, for example, have found at least three new genera of 17-million-year-old apes at a site in northern Kenya. The new fossils have some similarities with *Proconsul* and also some differences. The diversity of apes in the Lower Miocene period, in only a small part of East Africa, was clearly much greater than has been thought.

These discoveries have shown that the traditional interpretations of early hominoid evolution were gross oversimplifications based on samples that were limited in space as well as in time. This understanding, together with the realization that many characteristics thought to be special to hominoids may actually be primitive characteristics of anthropoids as a whole, has left investigators with a very different picture of *Proconsul*.

Proconsul was not a specialized ancestor of the modern chimpanzee or the gorilla. In fact, it has few special features that link it to these modern

primates. Instead *Proconsul* appears to be a generalized ancestor of all the great apes and humans. It is so generalized that if one compared *Proconsul* to an earlier, theoretical ancestor of all hominoids, only a few features—such as the frontal sinus and the lack of ischial callosities—would reveal that this early Miocene primate lived after the divergence of the lesser and the great apes.

FURTHER READING

THE FORE-LIMB SKELETON AND ASSOCIATED REMAINS OF *PROCONSUL AFRICANUS*. John R. Napier and Peter R. Davis in *Fossil Mammals of Africa*, Vol. 16, pages 1-69; 1959.

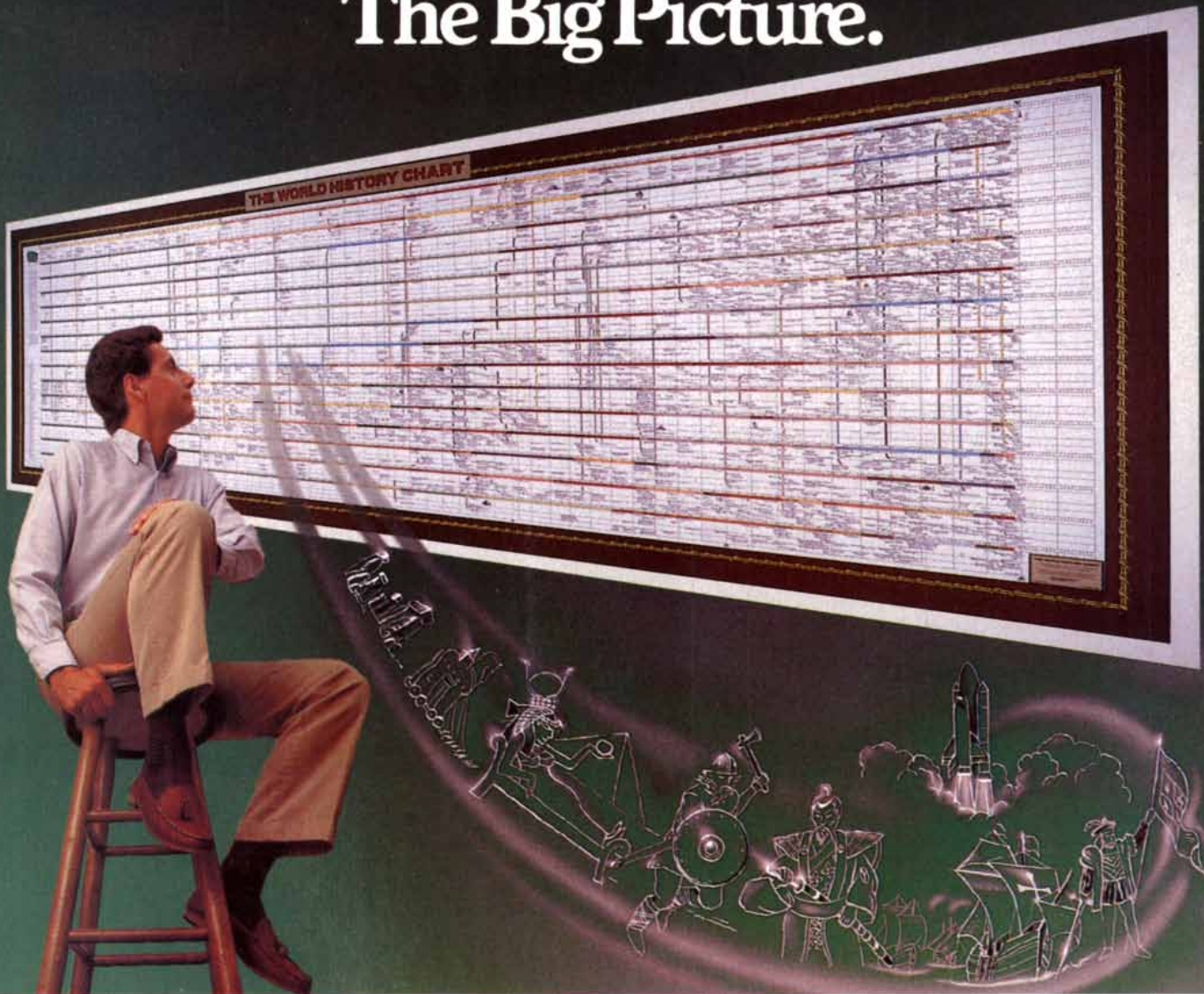
MIOCENE HOMINOID POSTCRANIAL MORPHOLOGY. Michael D. Rose in *New Interpretations of Ape and Human Ancestry*, edited by Russell L. Ciochon and Robert S. Corruccini. Plenum Press, 1983.

NEW POSTCRANIAL FOSSILS OF *PROCONSUL AFRICANUS* AND *PROCONSUL NYANZAE*. Alan Walker and Martin Pickford in *New Interpretations of Ape and Human Ancestry*, edited by Russell L. Ciochon and Robert S. Corruccini. Plenum Press, 1983.

THE SKULL OF *PROCONSUL AFRICANUS*: RECONSTRUCTION AND CRANIAL CAPACITY. Alan Walker, Dean Falk, Richard Smith and Martin Pickford in *Nature*, Vol. 305, No. 5934, pages 525-527; October 6, 1983.

NEW WRIST BONES OF *PROCONSUL AFRICANUS* AND *P. NYANZAE* FROM RUSINGA ISLAND, KENYA. K. Christopher Beard, Mark F. Teaford and Alan Walker in *Folia Primatologica*, Vol. 47, No. 2, pages 97-118; 1986.

The History Of The World Comes Alive When You See The Big Picture.



Never has there been a clearer, more concise overview of our history, archaeology, theology and mythology than with The World History Chart from International Timeline, Inc.

When you look at 6,000 years of history in chronological order—on a chart almost three feet high and eight feet long—important dates and events just seem to fall into place!

Test your talent for historical trivia. What do the Great Wall of China and the Chin Dynasty have in common with the Roman Empire? When the Normans migrated to

Northern France, where did they come from?

The world takes on new dimensions as you witness the development of 15 major civilizations from 4,000 B.C. to the present. To fully understand the scope of information on the Chart—picture U.S. history as just two inches over the space of eight feet of time!

This invaluable teaching and reference tool has taken years to compile. Its easy-to-read format appeals to everyone from the casual admirer to the scholar.

The World History Chart has been on

display in the Library of Congress. Now, you can have your own for just \$29.95 (U.S. currency) plus \$5.00 for shipping and handling (\$34.95 each). Add \$15.00 for each laminated chart (\$49.95). **Simply call toll-free 1-800-621-5559 (1-800-972-5855, in Illinois)**, and charge it to your MasterCard, VISA or Choice. Or, complete the coupon below and mail with check or money order payable to: **International Timeline, Inc., 2565 Chain Bridge Road, Vienna, VA 22180.** (Virginia residents add 4% sales tax. Outside U.S. add \$10.00 instead of \$5.00 for shipping and handling. Allow 4-6 weeks for delivery.)

Please send me the World History Chart for \$29.95 (U.S. currency) plus \$5.00 for shipping and handling for each chart ordered (\$34.95). Add \$15.00 for each laminated chart (\$49.95). Outside U.S. include an additional \$5.00 for shipping and handling.

6,000 Years Of History At A Glance For Only \$29.95!

Send to: _____ (please print) Address: _____

City: _____ State: _____ Zip: _____ # Of Charts Desired: _____ Laminated _____ Unlaminated

Enclosed is _____ check _____ money order for \$ _____ Checks payable to: **International Timeline, Inc.**

Virginia residents add 4% sales tax. Please allow 4-6 weeks for delivery. SA-2 2565 Chain Bridge Road, Vienna, VA 22180

The Shortest-Network Problem

What is the shortest network of line segments interconnecting an arbitrary set of, say, 100 points? The solution to this problem has eluded the fastest computers and the sharpest mathematical minds

by Marshall W. Bern and Ronald L. Graham

The imaginary Steiner Telephone Company figured that it would save millions of dollars if it could find the shortest possible network of telephone lines to interconnect its 100 customers. In search of a solution, Steiner hired the Cavalieri Computer Company, known for the world's fastest programmers and computers. After a week Cavalieri presented a program to solve Steiner's problem and showed that the program had indeed found a shortest network for 15 of the customers in just one hour. Steiner paid Cavalieri \$1,000 for the program and promised to pay one cent per second for the time it would take a computer to generate the complete solution. By the time the computer had finished the calculation for all 100 customers, the telephone company owed trillions of dollars in computer expenses and its customers had all moved many kilometers away—either by choice or by continental drift!

Did Cavalieri sell Steiner a faulty program? This dilemma is one example of the Steiner problem, which asks for the shortest network of line segments that will interconnect a set of given points. The Steiner problem can-

not be solved by simply drawing lines between the given points, but it can be solved by adding new ones, called Steiner points, that serve as junctions in a shortest network. To determine the location and number of Steiner points, mathematicians and computer scientists have developed algorithms, or precise procedures. Yet even the best of these algorithms running on the fastest computers cannot provide a solution for a large set of given points because the time it would take to solve such a problem is impractically long. Furthermore, the Steiner problem belongs to a class of problems for which many computer scientists now believe an efficient algorithm may never be found. For this reason the Cavalieri Computer Company should be exonerated.

On the other hand, Cavalieri could have developed a practical program that would have yielded solutions somewhat longer than the shortest network. Approximate solutions to the shortest-network problem are computed routinely for numerous applications, among them designing integrated circuits, determining the evolutionary tree of a group of organisms and minimizing materials used for networks of telephone lines, pipelines and roadways.

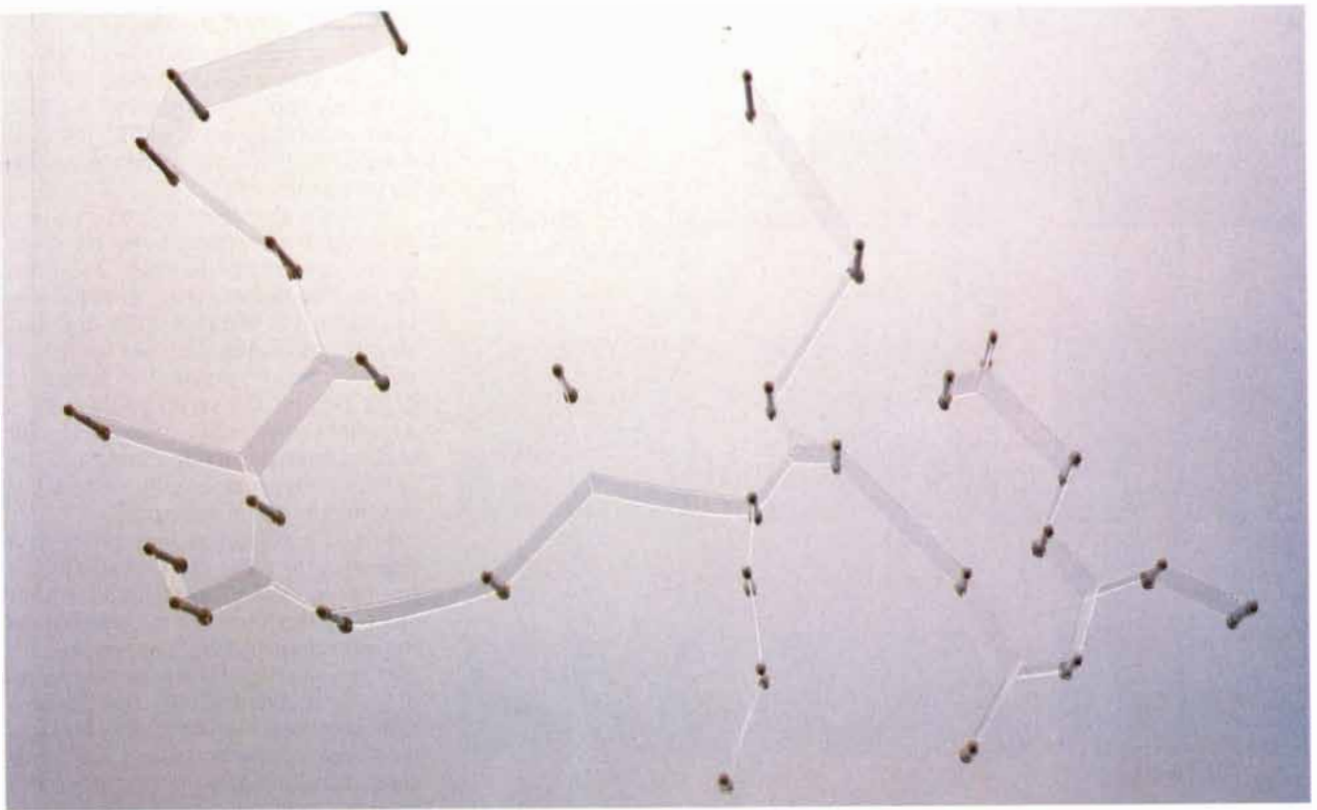
The Steiner problem, in its general form, first appeared in a paper by Miloš Kossler and Vojtěch Jarník in 1934, but the problem did not become popular until 1941, when Richard Courant and Herbert E. Robbins included it in their book *What Is Mathematics?* Courant and Robbins linked the problem to the work of Jakob Steiner, a 19th-century mathematician at the University of Berlin. Steiner's work sought the single point whose connections to a set of given points had the shortest possible total length. In about 1640, however, a special case of both problems—the one Steiner worked on and the one that

bears his name—was first posed: Find the point P that minimizes the sum of the distances from P to each of three given points. Evangelista Torricelli and Francesco Cavalieri solved the problem independently. Torricelli and Cavalieri deduced that if the angles at point P are all 120 degrees or more, then the total distance is minimized.

Knowing that the angles at P measure at least 120 degrees, Torricelli and Cavalieri developed a geometric construction for finding P [see top illustration on page 87]. Line segments are drawn connecting the given points (call them A , B and C , with B at the vertex of the largest angle). If angle B is greater than or equal to 120 degrees, then point P coincides with point B . In other words, the shortest network is simply the line segments between A and B and between B and C . If angle B is less than 120 degrees, then point P must be somewhere inside the triangle. To find it, an equilateral triangle is drawn along the longest side of the triangle, namely the side between points A and C . The third vertex of the equilateral triangle, labeled X , is opposite point B . The equilateral triangle is circumscribed, and a line segment is drawn from point B to X . Point P lies where the line intersects the circle. Joining points A , B and C to P creates three angles of exactly 120 degrees and yields the shortest network. Furthermore, the length of the line from B to X turns out to be equal to the length of the shortest network. For the purpose of our article we shall call X the replacement point, since replacing points A and C with X leaves the length of the network unchanged.

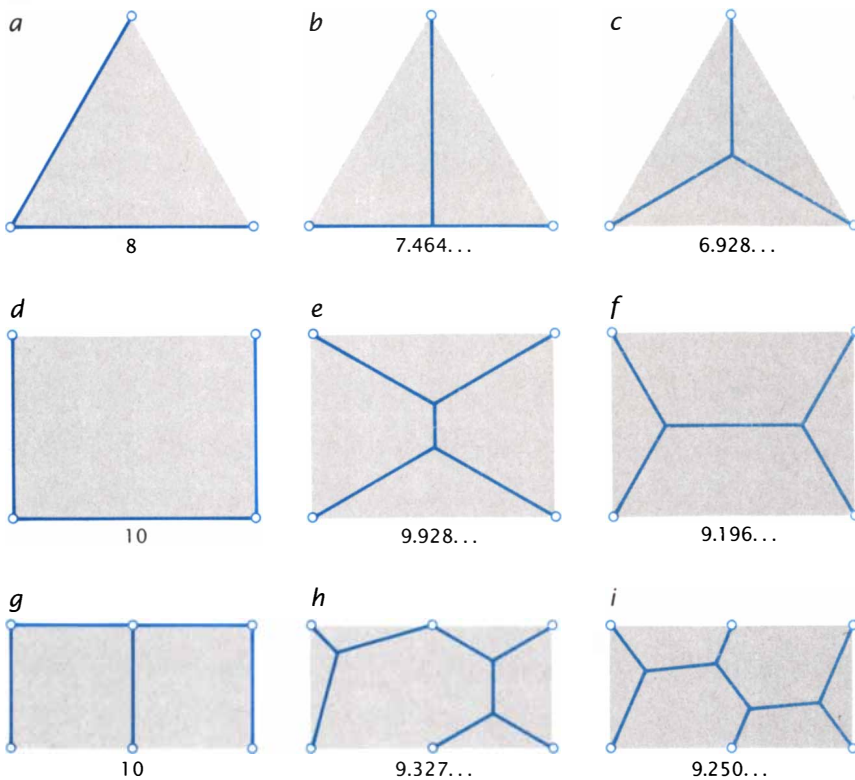
The three-point and the multipoint Steiner problem share many properties. The form of the solutions, known as trees, is such that removing any line segment from the shortest network detaches one of the given points. In other words, one cannot follow the network from a given point back to the same point without retracing lines.

MARSHALL W. BERN and RONALD L. GRAHAM have studied shortest-network problems for many years. Bern is a researcher at the Xerox Palo Alto Research Center. After earning his M.A. at the University of Texas at Austin in 1980, he worked in the signal-processing division of Enso, Inc. He received his Ph.D. in computer science from the University of California, Berkeley, in 1987. In his spare time Bern grows cacti and dabbles in painting and printmaking. Graham is Adjunct Director for Research, Information Sciences Division, at the AT&T Bell Laboratories and university professor at Rutgers University. He got his Ph.D. in mathematics from Berkeley in 1962. He enjoys pointing out that he is a former president of the International Jugglers Association. This is Graham's second article for SCIENTIFIC AMERICAN.



SOAP-BUBBLE COMPUTER (*top*) challenges an electronic computer (*bottom*) to find the shortest network of line segments interconnecting 29 cities. The soap-bubble computer, in which the placement of pegs mimics the geography, minimizes the length of soap films in a local area. It provides a short net-

work, but not necessarily the shortest one. The electronic computer implements an algorithm, authored by Ernest J. Cockayne and Denton E. Hewgill of the University of Victoria, that guarantees the true shortest network. The 29-point problem is close to the current limit of computing capabilities.



NETWORK PROBLEM for points arranged at the vertices of an equilateral triangle, a rectangle and a "ladder" has various solutions. In *a*, *d* and *g* the points are connected without adding new points, in what is known as the minimum spanning solution, or tree. Steiner trees, which are made by adding additional junction points, are shown in *c*, *e*, *f*, *h* and *i*. Only *c*, *f* and *i* are shortest Steiner trees, or shortest networks. The number under each solution gives the approximate total length of the line segments.

Solutions to the three-point and multipoint problem are therefore called Steiner trees, the segments are known as edges, and the points—analogueous to *P*—that must be added to construct the tree are called Steiner points.

The three-point Steiner problem also provides information about the geometry of shortest Steiner trees. First, every angle measures at least 120 degrees, which implies that every given point is connected to the tree by no more than three edges. Second, at every Steiner point exactly three edges meet, at 120-degree angles. Third, the number of edges in a tree is always one less than the number of given points added to the Steiner points. And last, since exactly three edges meet at every Steiner point and at least one edge must touch every given point, the maximum number of Steiner points in any problem is two fewer than the number of given points.

For the same number and arrangement of given points, many different Steiner trees can be constructed that have those properties. Some of the trees, known as locally minimal solutions, cannot be short-

ened by a small perturbation, such as moving an edge slightly or splitting a Steiner point. But not every locally minimal Steiner tree is a shortest solution possible. Large-scale rearrangements of the Steiner points may be necessary to transform a network into a shortest possible tree, called a globally minimal Steiner tree.

Let us demonstrate with a set of given points that define the four corners of a rectangle measuring three meters by four. The solutions have two Steiner points, which can be arranged in two different ways. Each arrangement forms a Steiner tree that has three edges connected to each Steiner point at 120 degrees. If the Steiner points are arranged parallel to the width, the locally minimal Steiner tree that results is about 9.9 meters long. If the Steiner points are arranged parallel to the length, a globally minimal Steiner tree results, measuring about 9.2 meters.

A brute-force approach to discovering a shortest network is to search through all possible locally minimal Steiner trees, calculate their lengths and choose the shortest one. Because Steiner points can be placed anywhere,

however, it is not clear that all possible locally minimal Steiner trees can be computed in a finite amount of time. Z. A. Melzak of the University of British Columbia overcame the difficulty and developed the first algorithm for the Steiner problem.

Melzak's algorithm considers many possible connections between given points and many possible locations for Steiner points. The algorithm can be outlined in two parts. The first part simply separates the set of given points into every possible subset of given points. The second part creates a number of possible Steiner trees for each subset by using a construction similar to the one employed to solve the three-point problem.

Just as in the three-point problem, a replacement point can be substituted for two of the given points without changing the length of the solution. In the general problem, however, the algorithm must guess which pair to replace, and eventually it tries all possible guesses. Moreover, the replacement point may be placed on either side of a line segment that joins the pair, because the equilateral triangle used in the construction can be oriented in two directions. After one pair of points in the subset is replaced by one of the two possible replacement points, each subsequent step of the algorithm replaces either two given points, a given point and a replacement point or two replacement points with another replacement point until the subset is reduced to three points.

Once the Steiner point for those three points is found, the algorithm works backward, attempting to determine the Steiner point corresponding to each replacement point [see bottom illustration on opposite page]. An attempt can fail because of contradictory constraints on the placement of Steiner points. A successful attempt, however, creates a Steiner tree connecting each given point in the subset with one edge. After considering all replacement sequences, the algorithm chooses the shortest of these Steiner trees for the subset. Combining shortest Steiner trees for subsets in all possible ways to span the original set of given points yields all possible locally minimal Steiner trees, and the geometry of the overall shortest network can be determined.

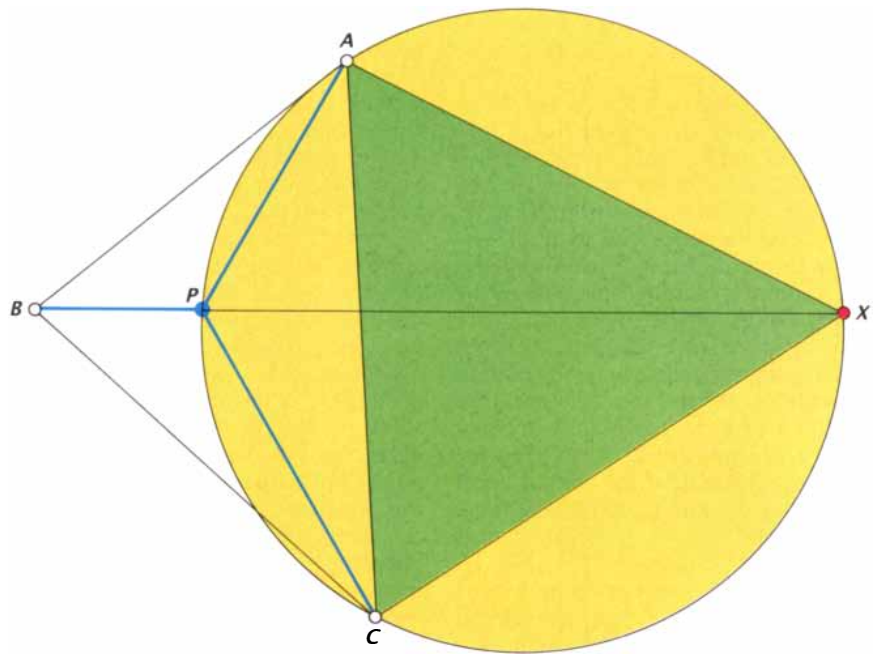
Melzak's algorithm can take an enormous amount of time even for small problems because it considers so many possibilities. A 10-point problem, for instance, can be separated into 512 subsets of given points. Although two-point subsets do not re-

quire much work, each of the 45 subsets of eight points has two million replacement sequences. Furthermore, there are more than 18,000 ways to recombine the subsets into trees.

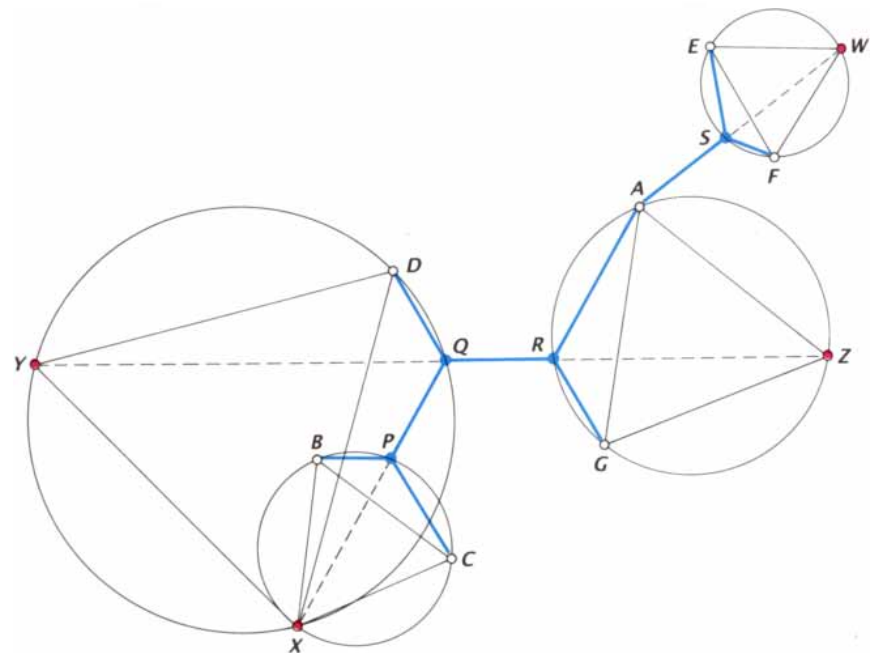
To be sure, investigators have found better ways to organize the computation and increase the speed of the algorithm. Instead of considering the problem's geometry, they focus on possible patterns of connections in the network—what is known as the network's topology. A topology specifies which points are connected to one another, but not the actual locations of Steiner points. Assuming a certain topology, one can find an appropriate replacement sequence relatively quickly. This organization greatly increases the speed of computing shortest Steiner trees for the subsets. For an eight-point subset, for example, the algorithm needs to consider only about 10,000 different topologies rather than two million different replacement sequences.

Because the number of topologies grows rapidly with the size of the subset, Steiner problems might become more manageable if only very small subsets of the set of given points needed to be considered. Experiments with Melzak's algorithm suggest that the shortest network for more than six random points can usually be separated into shortest networks for smaller sets of points. By considering special arrangements of points called ladders, however, Fan R. K. Chung of Bell Communications Research and one of us (Graham) demonstrated that there are arbitrarily large sets of points for which the shortest Steiner tree cannot be separated. A ladder is an arrangement in which all the given points are equally spaced along two parallel lines. A general solution was discovered for this quite special Steiner problem. It showed that the number of Steiner points in a shortest Steiner tree for a ladder with an odd number of "rungs" is the maximum: the number of given points minus two. Such a Steiner tree cannot be separated because the placement of every Steiner point requires that every given point be considered simultaneously. Thus one cannot simply declare a cutoff on the size of subsets considered in Melzak's algorithm.

A number of investigators improved on Melzak's algorithm by finding subtler ways to reduce the amount of work [see illustration on next page]. These methods prune, or eliminate, parts of the computation that would only yield long networks. New pruning



SHORTEST NETWORK for three points A , B and C can be constructed. An equilateral triangle ACX (green) is erected along the longest side of triangle ABC and then circumscribed with a circle (yellow). The intersection of the circle and a line segment from B to X , the equilateral triangle's third vertex, marks point P , known as the Steiner point. Joining points A , B and C to P forms three angles of 120 degrees and yields the shortest network. The length of line segment BX equals the network's length.



MELZAK'S ALGORITHM reduces a shortest-network problem into smaller problems. Point A is the correct place to separate the problem into a three-point problem and a five-point problem. To construct possible Steiner trees for the five-point problem, a pair of points (B and C , for instance) can be replaced with a single point (X in this case) by constructing an equilateral triangle on one side of B and C . The problem is thus reduced to four points: X , D , G and A . A pair of these points can then be replaced—in this case, first D and X with Y and then G and A with Z . Each of the equilateral triangles that results (XDY , AGZ and BCX) is circumscribed with a circle. The points at which a line from Y to Z intersects two of the circles give the Steiner points Q and R , and the intersection of a line from X to Q with the remaining circle determines the Steiner point P . Since the best partitioning and pairing cannot be determined in advance, all possibilities must be considered to find the shortest tree.

techniques have reduced computation times substantially. Programs based on Melzak's algorithm, such as one written in 1969 by Ernest J. Cockayne of the University of Victoria, could solve all nine-point problems and some 12-point problems in about half an hour. A program written recently by Cockayne and a colleague at Victoria, Denton E. Hewgill, uses a powerful pruning technique introduced by Pawel Winter of the University of Copenhagen to solve all 17-point problems and most randomly generated 30-point problems in a few minutes. Winter's pruning method is so successful at eliminating possible topologies that the bulk of the computation is now the recombination of solutions for subsets.

For any of these programs, however, running times can depend quite sensitively on the geometry as well as on the number of points. Moreover, the computation time of even the most sophisticated algorithm grows exponentially with the number of points, and Steiner problems of 100 points are still well out of reach. Will an efficient algorithm ever be found to

compute solutions for large Steiner problems?

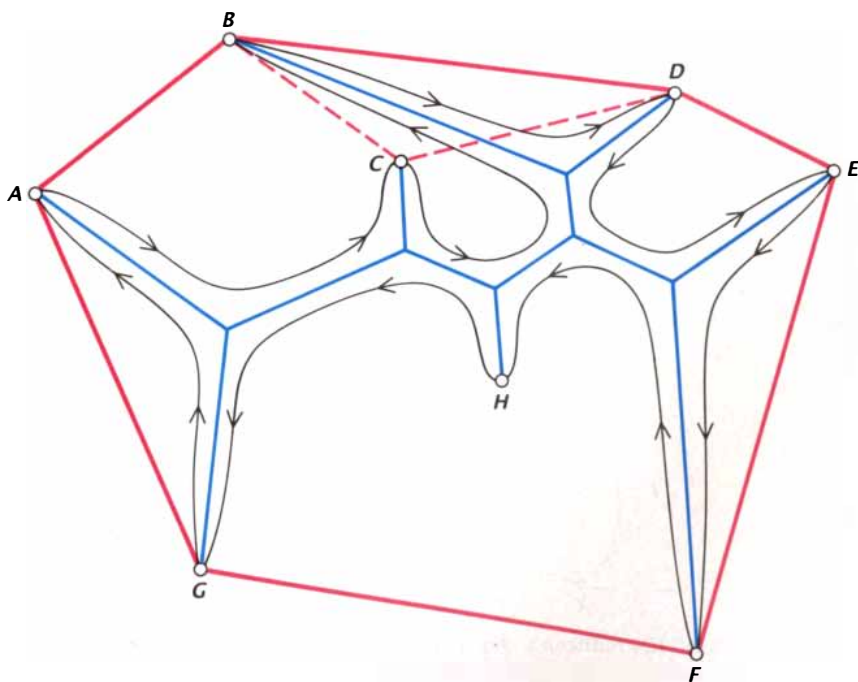
Advances in theoretical computer science have convinced most investigators that the existing algorithms for Steiner problems cannot be substantially improved. This theory assigns a size to each instance, or example, of a problem. For Steiner problems there is a natural measure of size: the number of given points. One then considers the number of basic computer operations—such as additions, subtractions or comparisons—an algorithm may need in order to solve an instance of a certain size. Since different instances of the same size may require different numbers of operations, one looks at the maximum number of operations as a function of size. If the number of operations increases by the size of the instance (n) to some power, as in the expressions n^2 , $5n$ or $6n + n^{10}$, the procedure is called a polynomial-time algorithm. These algorithms are considered efficient, at least in a theoretical sense. If the number of operations increases exponentially with size, as in cases such as 2^n , 5^n or $3n^2 \times 4^n$, the procedure is known as an exponential-time algorithm.

Although for very small problems both polynomial- and exponential-time algorithms are practical, for large problems the solution times of exponential-time algorithms are so slow that these algorithms are hopelessly impractical [see "The Efficiency of Algorithms," by Harry R. Lewis and Christos H. Papadimitriou; SCIENTIFIC AMERICAN, January, 1978]. For sufficiently large problems a polynomial-time algorithm executed on even the slowest machine will yield an answer sooner than an exponential-time algorithm running on a supercomputer.

Even though exponential-time algorithms have been found for the Steiner problem (Melzak's algorithm, for example), no polynomial-time algorithms have yet been found. The prospects for an efficient algorithm are not good. In 1971 Stephen A. Cook of the University of Toronto proved that if a polynomial-time algorithm could be found for any single problem in a group now known as NP-hard problems, that algorithm could be used to solve all other problems efficiently in a large class of hard problems including NP-hard problems. Later one of the authors (Graham), working with Michael R. Garey and David S. Johnson of the AT&T Bell Laboratories, proved that the Steiner problem is an NP-hard problem. Since all NP-hard problems have to date foiled the efforts of thousands of workers, it is considered unlikely that any NP-hard problem, including the Steiner problem, can be solved by a polynomial-time algorithm. Proving that NP-hard problems cannot be solved efficiently, however, is the preeminent problem in theoretical computer science.

Although it does not appear likely that an efficient, polynomial-time algorithm will be found for computing shortest networks, there are practical algorithms producing slightly longer networks. One example is the algorithm for solving the minimum-spanning-tree problem, which searches for the shortest network of line segments that will interconnect a set of given points without adding any new ones. To solve it one connects the two given points that are closest together, and in each subsequent step one connects the closest pair of points that can be joined without forming a closed path. After all, an edge can be removed from a closed path and leave given points still connected by the remaining edges.

Edgar N. Gilbert and Henry O. Pollak of Bell Laboratories have conjectured that the ratio of a shortest Steiner tree to a minimum spanning tree is at least



PRUNING METHODS increase the efficiency of algorithms for finding short networks. One way to prune, or rule out, possible networks (devised by Cockayne) is to consider the order in which a rubber band (red) stretched around the set of given points touches them. The rubber band touches all the points except C and H, but C can be included in the sequence because the angle formed by point C and two consecutive points in contact with the rubber band measures at least 120 degrees. The order of points is then ABCDEFG. An unbroken path (purple) traced around a possible network (blue) touches the points in the order ACBDEFHG. Since B and C are reversed with respect to the order established by the rubber band, this network can be pruned.

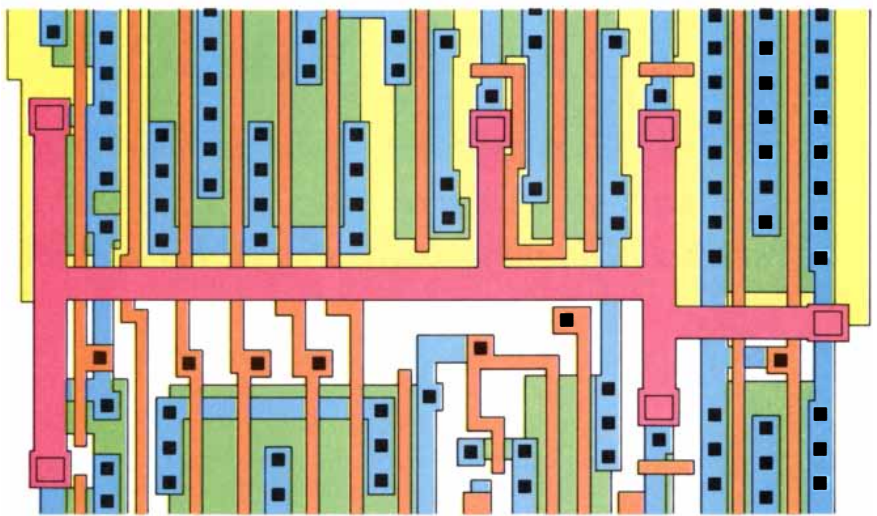
$\sqrt{3}/2$, that is, the Steiner tree's length is at most about 13.4 percent shorter than the minimum spanning tree's length. The ratio of $\sqrt{3}/2$ occurs in a simple example: three given points forming an equilateral triangle. Although the conjecture remains unproved, Chung and one of the authors (Graham) have proved that the Steiner tree is at most 17.6 percent shorter than the spanning tree.

Minimum spanning trees can often be shortened 3 to 4 percent by carefully adding Steiner points and adjusting the tree. One of the authors (Bern) has shown that this kind of inexact algorithm has some theoretical justification, since the average length of an adjusted tree will be a little less than the average length of a minimum spanning tree.

The minimum-spanning-tree and shortest-network problems have been applied to constructing telephone, pipeline and roadway networks. The solutions, whether approximate or exact, can provide guidelines for the layout of the networks and the necessary amounts of materials. More complicated versions of the Steiner problem can accommodate the need to avoid certain geographic features or to find the shortest connections along preexisting networks.

Perhaps the most practical application of the Steiner problem is in the design of electronic circuits. A short network of wires on an integrated circuit requires less time to charge and discharge than a long network and so increases the circuit's speed of operation. The shortest-network problem on circuits, however, involves a different kind of geometry, since wires on a circuit generally run in only two directions, vertical and horizontal.

The problem, known as the rectilinear Steiner problem, was first investigated in 1965 by Maurice Hanan of the IBM Corporation's Thomas J. Watson Research Center in Yorktown Heights, N.Y. As in the original Steiner problem, the solution to the rectilinear version is also a tree containing Steiner points and given points, but edges meet at 90 or 180 degrees. Although Steiner points could conceivably lie anywhere in the rectilinear problem, Hanan showed that it is possible to restrict the locations of Steiner points in a shortest rectilinear network. A vertical and a horizontal line are drawn through each given point, and each intersection of two lines defines a possible Steiner point. An algorithm can try all subsets of possible Steiner points in order to compute a short-



VARIATIONS of the shortest-network problem have been applied to the design of electronic circuits in order to increase operating speeds. The shortest network of horizontal and vertical wires that interconnect a set of terminals is shown in magenta. The background shows other wires and terminals arranged in deeper layers.

est network. As the number of given points increases, however, the solution time of such a brute-force algorithm grows exponentially. More sophisticated but still exponential-time algorithms can solve rectilinear Steiner problems that have about 40 points.

A rectilinear version of the minimum-spanning-tree problem, which can be solved efficiently by the algorithm that chooses the shortest connection at each step, unless that connection forms a closed path. Frank K. Hwang of Bell Laboratories has proved that a rectilinear Steiner tree is never shorter than a rectilinear minimum spanning tree by more than one-third.

The most surprising application of the Steiner problem is in the area of phylogeny. David Sankoff of the University of Montreal and other investigators defined a version of the Steiner problem in order to compute plausible phylogenetic trees. The workers first isolate a particular protein that is common to the organisms they want to classify. For each organism they then determine the sequence of the amino acids that make up the protein and define a point at a position determined by the number of differences between the corresponding organism's protein and the protein of other organisms. Organisms with similar sequences are thus defined as being close together and organisms with dissimilar sequences are defined as being far apart. In a shortest network for this abstract arrangement of given points, the Steiner points correspond to the most plausible ancestors, and edges correspond to a relation be-

tween organism and ancestor that assumes the fewest mutations. Since the phylogenetic Steiner problem is no easier than other Steiner problems, however, the problem—except as it is applied to small numbers of organisms—has served more as a thought experiment than as a practical research tool.

Although knowledge about algorithms has progressed greatly in recent years, the shortest-network problem remains tantalizingly difficult. The problem can be stated in simple terms, and yet solutions defy analysis. A tiny variation in the geometry of a problem may appear to be insignificant, and yet it can radically alter the shortest network for the problem. This sensitivity renders even peripheral questions about shortest networks quite challenging. The shortest-network problem will continue to frustrate and fascinate us for years to come.

FURTHER READING

- STEINER MINIMAL TREES. E. N. Gilbert and H. O. Pollak in *SIAM Journal on Applied Mathematics*, Vol. 16, No. 1, pages 1-29; January, 1968.
- COMPANION TO CONCRETE MATHEMATICS. Z. A. Melzak. John Wiley & Sons, Inc., 1973.
- AN ALGORITHM FOR THE STEINER PROBLEM IN THE EUCLIDEAN PLANE. Pawel Winter in *Networks*, Vol. 15, No. 3, pages 323-345; Fall, 1985.
- STEINER PROBLEM IN NETWORKS: A SURVEY. Pawel Winter in *Networks*, Vol. 17, No. 2, pages 129-167; Summer, 1987.

André-Marie Ampère

The first investigator to quantify the magnetic effects of electric current, Ampère was also a pioneer in the philosophy of science. His philosophy shaped his method of scientific discovery

by L. Pearce Williams

When my freshmen students ask me who André-Marie Ampère was, I often reply, "You probably know him better as Amps; he comes in two sizes, 15 and 35." In fact, Ampère is best known as the founder of the science of electrodynamics. In the early 19th century he carried out the first systematic investigations of the magnetic fields produced by electric currents, discovered and quantified the forces that act between current-carrying wires, and was the first investigator to propose that the magnetism observed in permanent magnets is caused by tiny electric currents circulating within the molecules of magnetic material. His achievements are commemorated in the name of the international unit of electric current: the ampere, or amp—the unit in which electric fuses, such as the common 15- and 35-amp fuses, are sized.

Ampère is less well known for his achievements as a philosopher of science, although they were in some ways just as significant. He was the first major scientist to adapt the views of the German philosopher Immanuel Kant to provide a philosophical foundation for the study of physics and

chemistry. Ampère took Kant's theory of knowledge—his theory of what human beings can know and how they can come to know it—as a starting point in developing a practical method of scientific discovery. Kantian philosophy eventually came to pervade the physical sciences, particularly in the late 19th and early 20th centuries, and Ampère's method of discovery, which guided his investigations of electrodynamics, survives today in modified form as one of the most commonly accepted scientific methodologies.

Born in 1775, Ampère grew up in the village of Poleymieux, outside Lyons. As a youth he educated himself, avidly reading books from his father's library and volumes brought from Lyons. He devoured Denis Diderot's newly completed *Encyclopédie*, committing entire articles to his photographic memory. He was particularly interested in mathematics; he worked his way through the treatises of the Swiss mathematician Leonhard Euler on advanced algebra, probability theory and the calculus, and in his late teens he mastered Joseph-Louis Lagrange's groundbreaking book on analytical mechanics. His interests outside mathematics were literally encyclopedic. Ampère studied Georges de Buffon's works on natural history, learned Greek, Latin and Italian, tried to develop a universal language that would be based on the most up-to-date principles of linguistics, immersed himself in French literature, wrote poetry, studied botany and developed ways to systematize his own observations of the natural world. This broad range of interests never left him.

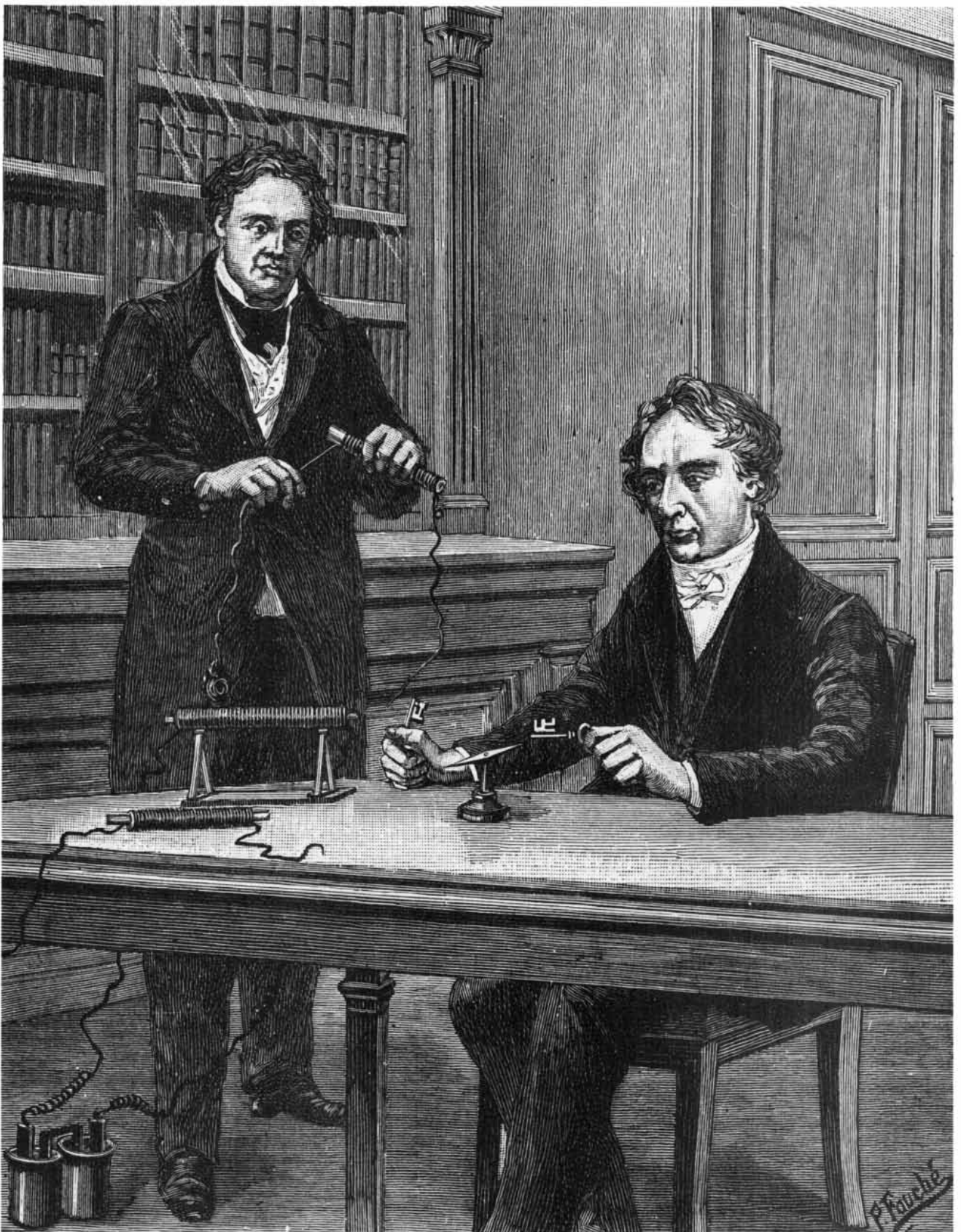
For his first 18 years, Ampère, as the only son in a family of comfortable means, led a nearly idyllic existence, wandering wherever he liked—physically and intellectually—in a secure, stable world. In 1793, however, began a series of personal tragedies that

punctuated the rest of his life. In 1793 his father was guillotined as a counterrevolutionary. In 1803 his beloved wife of four years (and mother of his son, Jean-Jacques) died, and four years later he entered into a disastrous second marriage that ended in divorce, leaving him with a young daughter. A few years of relative calm were shattered in 1819, when his son left all prospects behind to join the retinue of Mme Récamier, the great beauty of the Napoleonic era, who collected male admirers and resolutely kept them at arm's length. Jean-Jacques attended her for the next 20 years, ignoring his father's pleas to return home and make something of himself. In 1827 Ampère married his daughter to a former officer of Napoleon's army, only to find later that the man was both an alcoholic and insane.

Throughout his life Ampère suffered from a steady deterioration of his health, which increasingly impeded his scientific efforts. He also suffered from financial insecurity for most of his life and had to take a variety of low-paying jobs, most of which involved teaching mathematics at one level or another.

Ampère's development as a scientist was hindered not only by these personal circumstances but also by the breadth of his own intellectual activity. In a recently discovered collection of letters, written when he was 20 years old, he jumps from theoretical mechanics to the construction of practical machinery, to the theory of kite flying, to artificial languages, to music, to astronomy, to botany, to systems of classification. He was never able to focus narrowly on a given area and bring his full genius to bear on it. While he was feverishly developing his electrodynamics he continued to speculate on problems of metaphysics and philosophy. He never saw any contradiction in all of this, for he was convinced that a basic unity underlies all knowledge. As we shall see, his last

L. PEARCE WILLIAMS is professor of the history of science and director of the Program in History and Philosophy of Science and Technology at Cornell University. He entered Cornell as an undergraduate in 1944 and returned there after service in the Navy to get his Ph.D., which he received in 1952. He taught at Yale University and the University of Delaware before joining the faculty at Cornell. Williams writes: "I became fascinated with Ampère when I was working on a biography of Michael Faraday. The two were so dissimilar in lives, methods of research and ideas that I knew a biography of Ampère (which I am now working on) would be fun to write. My interests outside work include breeding and training Weimaraners as bird dogs and the study and teaching of Shito-ryu karate, in which I hold a black belt."



ANDRÉ-MARIE AMPÈRE (*left*) and his friend François Arago examine the magnetic effects of electric currents in an imaginative (and somewhat inaccurate) late 19th-century reconstruction of their collaboration. Ampère, who was a founder of electrodynamics (the study of electric currents), adapted phil-

osophical ideas to provide a methodological framework for scientific investigation. Arago was the investigator who told Ampère and his colleagues in the French Academy of Sciences of the Danish physicist Hans Christian Oersted's discovery that an electric current could deflect a nearby compass needle.

major work was an attempt to demonstrate that point.

When Ampère first began his scientific work, French philosophy was dominated by a school whose members had been contemptuously dubbed *Idéologues* by Napoleon. The *Idéologues* professed to have determined the complete rules for proper scientific procedure. According to their beliefs, the human mind is a passive receptor of sensory impressions. On the basis of these impressions the mind creates a series of pictures (which include sensations of smell, sound, taste and touch in addition to sight) that represent the outside world. The mind can recall pictures and compare them, in order to determine their differences and to ascertain how the succession of images changes over time. Any observed regularities can then serve as the basis for scientific laws. Yet there is no way to tell whether there really is an outside world. The only known reality lies in the pictures.

In this world view there is no causation, merely a succession of images. It is thus impossible, in one sense, to explain phenomena—to speak of the physical forces by which a certain cause brings about a particular effect. Hence there can be no scientific theories, in one modern sense of the word. On the one hand, science becomes taxonomic: the scientist arranges similar sensations in properly organized groups. On the other, it becomes positivistic: the scientist expresses observed regularities in mathematically rigorous laws. A good example of the first kind of science is Antoine Lavoisier's system of chemistry, which was based on exact description, precise nomenclature and proper classification of chemical elements and compounds. The second kind of science was typified by Jean-Baptiste Fourier's studies of heat, in which Fourier ignored the causes of heat, concentrating instead on determining the mathematical laws of thermal propagation. Such a philosophical framework is adequate for many kinds of scientific

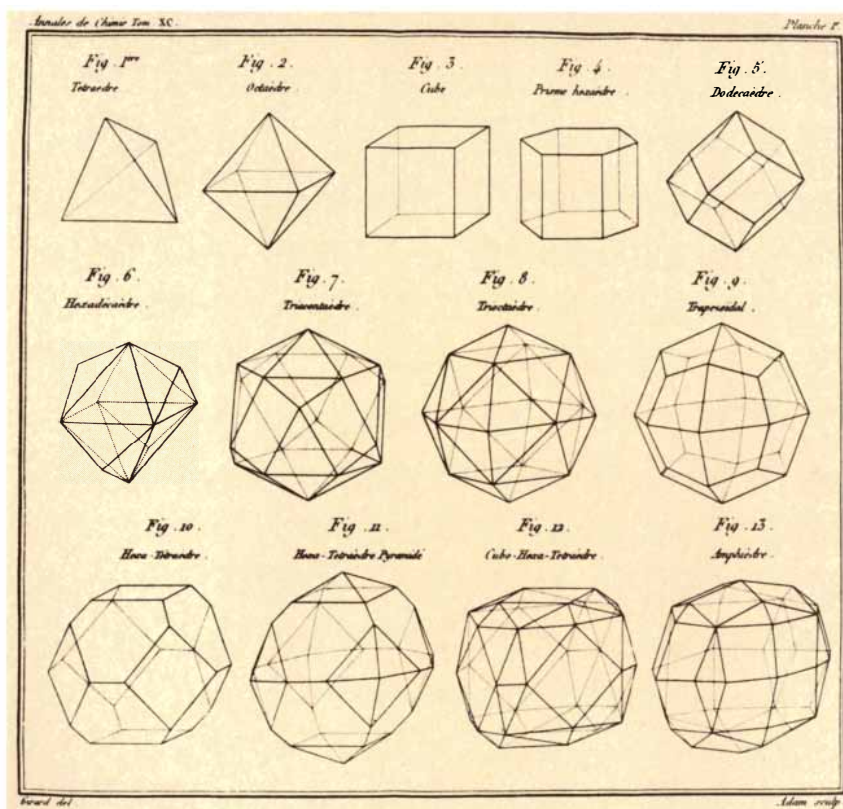
study, but it essentially rules out such fields as microphysics, which relies on theoretically posited entities that cannot be observed directly, such as atoms and molecules.

Ampère's early scientific work was not limited by this framework; he began his career as a mathematician, and mathematics need not refer to an external physical reality. Ampère's first published paper, written in 1802 to gain him the reputation necessary to be named professor at the Napoleonic lycée in Lyons, was on the mathematical theory of gambling. In this small treatise Ampère showed that a gambler who has a finite amount of money but faces either a single opponent with infinite financial resources or a large number of opponents with finite resources will necessarily lose all within a finite amount of time. Ampère also wrote papers on theoretical mechanics, and his most extensive mathematical work—written to gain election to the French Academy of Sciences—was a treatise on partial differential equations. A number of other mathematical papers complete this phase of his career. Had his efforts been confined to mathematics, he would now be known, if at all, as a competent and sometimes innovative mathematician who was overshadowed by his contemporaries Laplace, Poisson, Cauchy and Fourier.

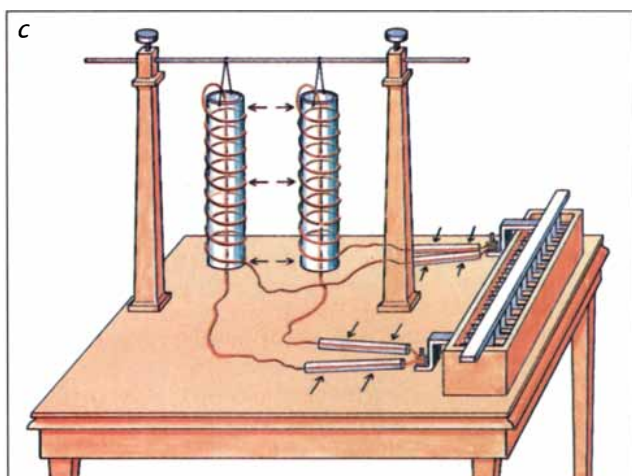
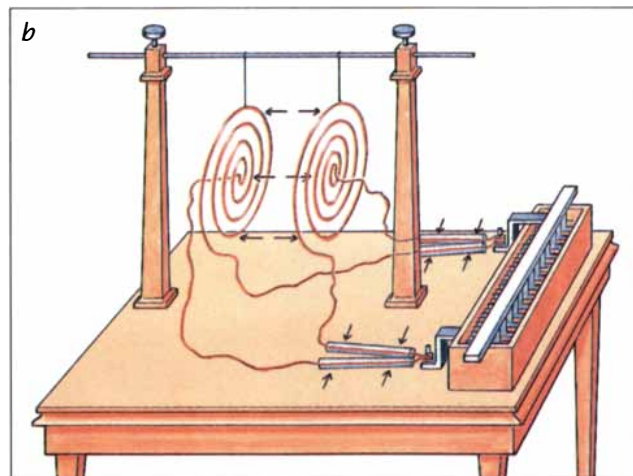
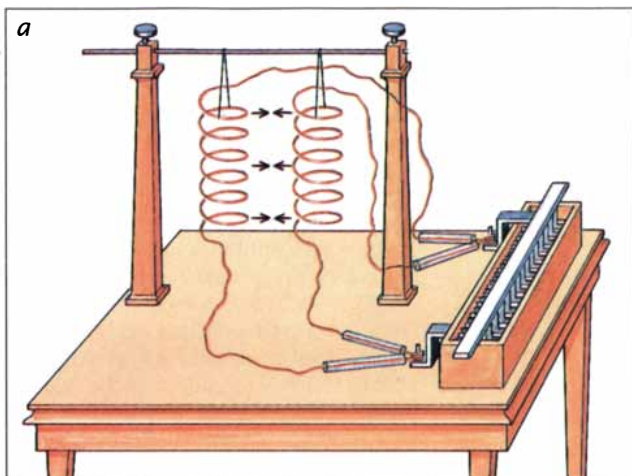
One reason for Ampère's failure to break new ground in mathematics was that by 1805 he was bored with the subject. He had found two new passions: metaphysics and chemistry.

Ampère's fascination with metaphysics sprang from his introduction to the *Idéologues* in 1804, when he moved from Lyons to Paris after the death of his first wife. Ampère was one of a small group of philosophers who gathered in the suburban village of Auteuil to discuss the work of the Abbé de Condillac, the founder of the *Idéologue* school. He was soon repelled by the views of the *Idéologues* because they ruled out the existence of God and the immortal soul, and he left the group to search for an alternative. He found inspiration for his own philosophy in the writings of Kant.

Kant divided the world into two domains: the domain of phenomena and the domain of noumena. Phenomena are events as perceived by the human mind—they are sensations. Noumena are the causes of phenomena—they are the so-called things-in-themselves, the objects that really exist. Human beings can never know the noumena



POLYHEDRAL MOLECULES formed the conceptual basis of Ampère's early work in theoretical chemistry. Motivated by Immanuel Kant's assertion that true sciences should be based on mathematical principles, Ampère attempted to explain chemistry in terms of geometry. In Ampère's scheme the most fundamental molecules were made of points arranged as the vertexes of the regular or nearly regular solids in the top row of this chart (which is from a paper by Ampère in *Annales de Chimie*). Reactions between these molecules could occur only if they resulted in solids with a degree of regularity and symmetry, such as those in the other two rows.



SERIES OF EXPERIMENTS tested Ampère's fundamental hypothesis that all magnetism (even that of permanent magnets) is caused by electric currents flowing in circles. If the hypothesis is true, Ampère reasoned, then helices carrying current in the same direction (a) should repel each other, like permanent magnets whose north poles point in the same direction. Instead the helices attracted each other. When Ampère tried the experiment with flattened spirals, the spirals did act like magnets (b). In that experiment Ampère noticed that straight wires carrying current in the same direction near his battery (right) attracted each other. He decided that because his original helices had been wound loosely, the circular flow of the current had been insignificant compared with its longitudinal flow, and so the helices had attracted each other as straight wires would have. He tested the hypothesis by winding helices around glass tubes and passing the wire back down each tube, so that the returning current would cancel the effects of the longitudinal current. Helices so wound, as shown in a simplified version of the experiment (c), did repel each other.

directly: noumena are the sources of the signals that act on our senses, and we can perceive only the signals, not the sources. According to Kant, then, we cannot ever really know anything definite about the noumena.

It was on this point that Ampère broke from Kant. Ampère noted that there are often relations—*rappports*, to use his word—between phenomena. Those *rappports*, he believed, must be analogous to *rappports* between the noumena underlying those phenomena. Hence it should be possible to learn about the interactions between unobservable noumena by studying phenomenal *rappports*.

This philosophical assertion was the foundation on which Ampère built his method of scientific discovery. He explained his method in a letter written in 1810 to his old friend Marie-François Pierre Maine de Biran. In Ampère's method the scientist accounts for phenomena by hypothesizing the existence of certain noumenal entities. Then he engages in a process of deduction: Accepting the existence of

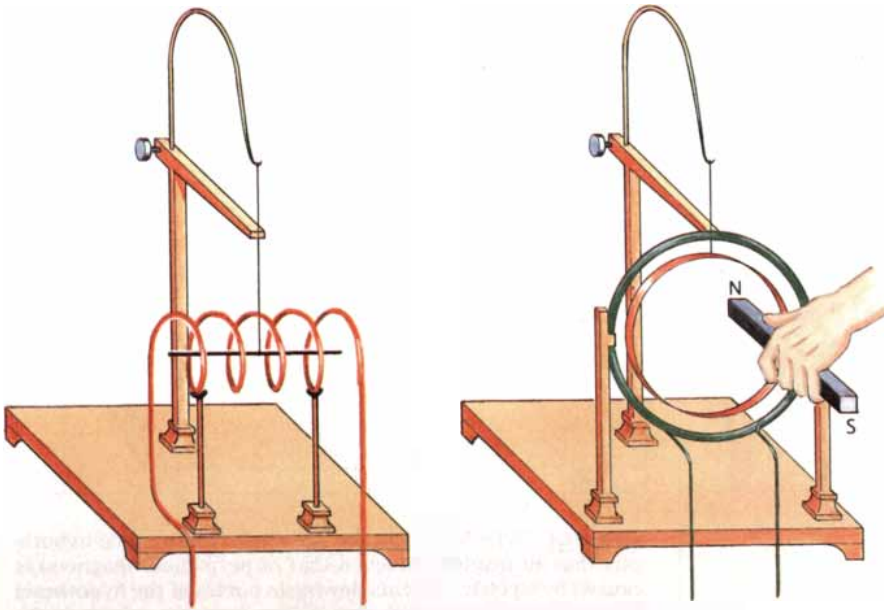
the theoretical entities, what new experimental results—phenomena—can be expected? The deduction is then tested by experiment. The theoretical entities can be assumed to exist as long as they give a testable account of phenomena. The probability that the theory is true increases as its ability to withstand experimental attack is demonstrated. Ampère's method is known today as the hypothetico-deductive approach, and it is viewed by many as the proper way to conduct scientific research.

A good example of the method in action is Ampère's analysis of Joseph-Louis Gay-Lussac's law of combining volumes. According to this phenomenal law, when integral volumes of gases are combined in a reaction, they must produce an integral volume of product; for example, when two liters of hydrogen are combined with one liter of oxygen, the product is precisely two liters of water vapor. For the *Idéologues*, that was as far as one could go; one could not say *why* the law is true. Ampère was determined to

go further. In 1814 he argued, in one of the first modern papers on theoretical chemistry, that one could account for the phenomena only by assuming that equal volumes of different gases at the same temperature and pressure must contain the same number of molecules. The existence of molecules—unobservable noumena—is the basis of Ampère's explanation of the phenomena described by Gay-Lussac.

Soon after devising his method, Ampère applied it to develop a brilliant theoretical framework for chemistry. Kant had insisted that sciences must be founded on a priori mathematical principles—principles that he believed are inherent in the structure of the human mind and are not based on observation. Following Kant's teaching, Ampère tried to deduce the laws of chemical affinity (the laws that determine which chemical reactions are possible) by assuming the existence of hypothetical molecules that were geometric in nature.

Ampère assumed that each molecule was composed of pointlike atoms



NATURE OF CURRENTS underlying magnetism was explored in a pair of experiments carried out separately by Arago and Ampère. In the first experiment (*left*) Arago wound a helix of copper wire, suspended an iron needle in the center and ran a current through the helix. The needle became magnetized. Ampère thought to test whether the circular currents responsible for permanent magnetism flow around every particle of magnetic material or around the axis of the magnet as a whole. In his experiment (*right*) he wound a coil of insulated copper wire (*green*) and suspended within it a ring made of copper ribbon. If the current in the coil had magnetized the needle by making currents flow around its central axis, Ampère thought, it should also magnetize the copper ring, because currents could circulate around its axis. He ran a current through the coil and held a bar magnet near the ring to see if it had become a “permanent” magnet. It had not, convincing Ampère that the currents in permanent magnets circulate around individual particles and not the magnetic axis. (The experiment—but not the conclusion—is flawed from a modern perspective.)

arranged in space as the vertexes of a simple geometric solid, such as a tetrahedron, an octahedron or a cube. The only chemical combinations that could occur were those that produced geometric solids having a certain degree of three-dimensional symmetry and regularity. In Ampère’s theoretical framework the puzzling arbitrariness of chemical activity could be reduced to mathematical certainty: chemistry could be based on geometry, which Kant believed was the purest form of mathematics. I should note that neither of Ampère’s chemical papers attracted much support from chemists, who were hostile to both his speculations and his mathematics.

Ampère’s accomplishments by the year 1819, when he was 44, would probably rate only a footnote in the history of physics. Unlike his contemporaries Augustin Fresnel (Ampère’s close friend and a creator of the undulatory theory of light) and Sadi Carnot (a founder of thermodynamics), who died at the ages of 39 and 36 respectively, Am-

père did his greatest scientific work in middle age, after he had despaired of ever making a serious mark.

Ampère’s first experiments in electrodynamics involved the voltaic pile, which had been invented by Alessandro Volta in 1800. A voltaic pile is an electrochemical cell, much like a modern automobile battery. If the poles of a voltaic pile are connected by a wire, a current will flow along the wire while chemical reactions inside the cell act to preserve the voltage difference between the poles.

Continuous electric current was a new phenomenon in the early 19th century, and early theories of current relied heavily on theories of static electricity. Few people expected current to have any magnetic effects, because Charles de Coulomb had shown in the 1780’s that the forces associated with static electricity are distinct from magnetism. Only a few German and German-influenced “nature philosophers,” who believed in the unity of all forces, sought a relationship between electricity and magnetism.

One of these philosophers was Hans

Christian Oersted of Denmark. In 1807 and 1812 he published works insisting, on philosophical grounds, that electricity and magnetism must be related. In the winter and spring of 1820 he finally detected a relationship by holding a compass needle near a long wire; when current flowed in the wire, the needle was deflected. Oersted’s finding was published in all the leading scientific journals of the day.

The news was brought to Paris by François Arago, a friend of Ampère’s, who had witnessed the effect during a visit to Geneva. Members of the Academy of Sciences were skeptical of Arago’s report, and they were convinced only by his actual demonstration of the effect on September 11 of that year. Ampère was present at the demonstration, and he went home to investigate the effect for himself. He immediately realized that Oersted had not fully understood the experiment—he had not taken account of terrestrial magnetism. The amount by which his compass needle had been deflected depended on the angle between the current-carrying wire and the earth’s magnetic field.

Ampère set out immediately to find the true effect of electric current on a compass needle by devising an arrangement of freely rotating magnets that neutralized the earth’s magnetic field in a small region. To his great satisfaction he found that the compass needle now always aligned itself at right angles to the current-carrying wire. He then realized that a compass needle could be used as part of an instrument to detect an electric current. With his new instrument, which he called a galvanometer, he mapped the current throughout a circuit made up of a wire and a voltaic pile.

Until then it had been assumed that the mechanisms operating within a voltaic pile were distinct from the current flowing along a wire connecting the pile’s two poles. Ampère found to his surprise that the current flowing through the voltaic pile was the same as the current in the rest of the circuit. What would happen, he wondered, if he constructed a circular pile—a pile bent into a ring so that the positive pole touched the negative pole? It appears from indirect evidence that he built such a pile during the same fertile September of 1820 and found that it created a symmetrical magnetic field. At this point Ampère’s mind leaped to the hypothesis that he would defend for the rest of his life: Magnetism is no more than electric currents moving in circles. This was as far as he had got by September 18,

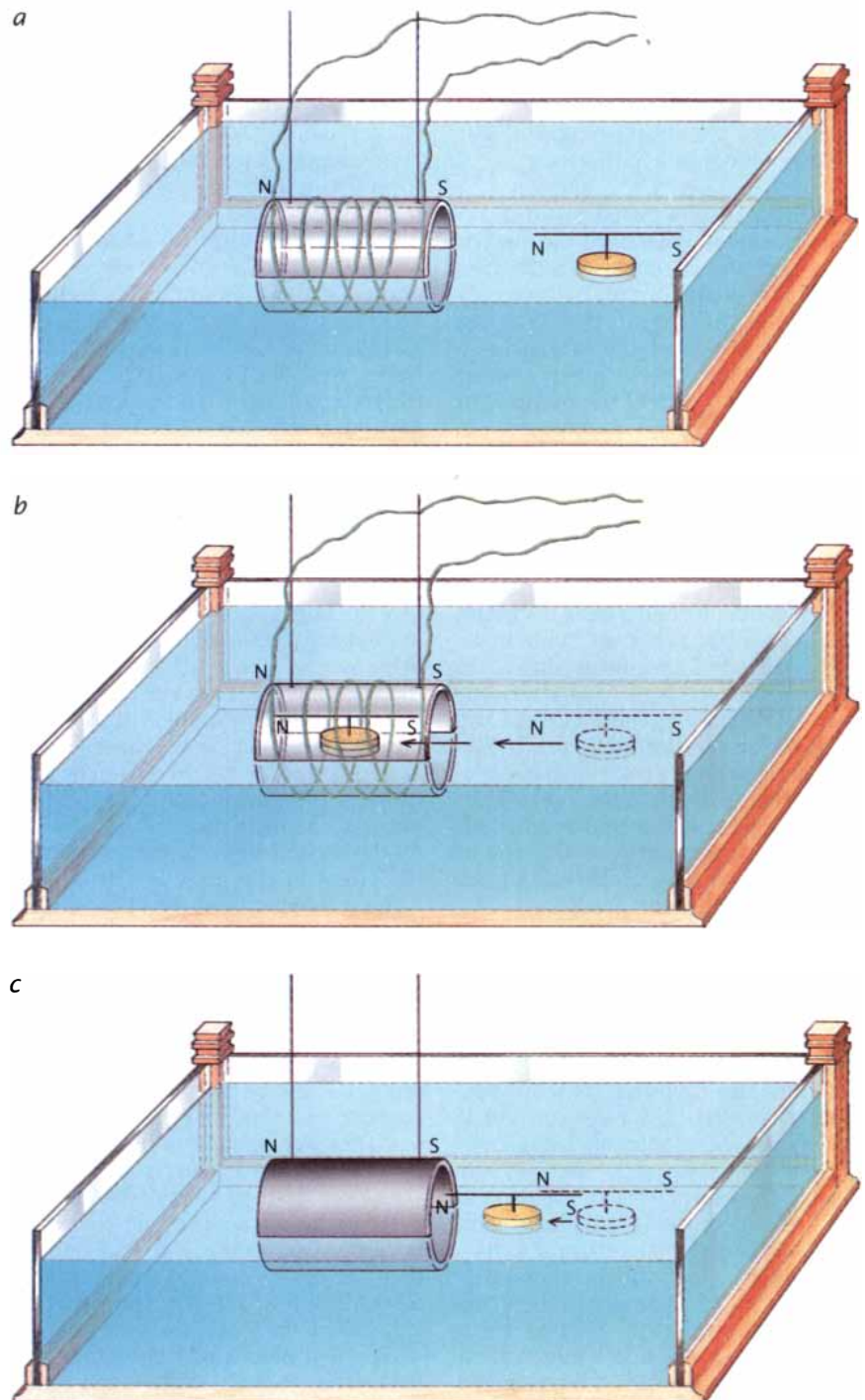
when he read his first report to the academy. No one paid much attention.

Between September 18 and September 25, the date of the academy's next meeting, Ampère applied his method. Having hypothesized a noumenal explanation—circular electric currents—for the phenomenon of magnetism, the next step was to test deductions based on the hypothesis: to show that circular electric currents produce the same effects as permanent magnets. He first tried to demonstrate the effect in copper wire that had been wound into helices (coils shaped like springs). When he placed two helices side by side and ran a current in the same direction through both of them at once [see illustration on page 93], he expected them to repel each other, like two magnets placed side by side with their north poles pointing in the same direction. Instead they attracted each other.

Ampère then tried a second approach. He wound his copper wire into flat spirals, with one end of the wire coming out of the center of the spiral and the other coming out of the spiral's edge. This time, when he turned on the current and held the spirals near each other, they did indeed act like magnets. On September 25 he demonstrated their various attractions and repulsions to his colleagues.

Why had helices acted differently? The answer came from a chance observation. The wires leading to the two spirals were connected to the same battery, and so the wires connected to the same pole of the battery passed near each other. When Ampère turned on the current, he noticed that wires carrying current in the same direction attracted each other, even though they were not wound in circles. This interaction between straight wires, never before observed, gave Ampère the answer to the helix problem. Because his wires were uninsulated, he had wound his helices very loosely, so that the individual coils in each helix were relatively far apart. The circular motion of the current had therefore been insignificant compared with its longitudinal travel from one end of the helix to the other. Hence the helices had been much like two straight wires carrying current in the same direction, and so they had attracted each other.

To test this conclusion, Ampère wound a helix around a glass tube and passed the end of the wire back down the center of the tube. He expected the current passing down the center of the tube to cancel the longitudinal effect of the current flowing from one



APPARENT CONTRADICTION of Ampère's fundamental hypothesis of magnetism came from experiments done by Michael Faraday. Faraday wound a helix of insulated copper wire around a hollow glass tube and suspended the tube in a tank of water (a). He then attached a long, magnetized needle to a cork floating near the tube and ran a current through the helix. If the current-carrying helix were exactly like a magnet, he argued, the cork should float up to the tube until the needle was just inside the helix; then one magnetic pole of the needle would be as close as possible to the opposite pole of the helix-magnet. Instead (b) the cork floated all the way through the tube until both of the needle's poles were near the corresponding poles of the helix. In another experiment (c) Faraday rolled a sheet of steel into a tube, magnetized it and suspended it in the water. This time, as expected, the needle floated up to one end of the tube and stopped. Faraday argued that the current-carrying helix and the magnet were therefore not identical. (Ampère was able to reconcile this result with his hypothesis by pointing out that in the helix the needle had been in the center of the current and in the steel tube it had been outside all the small loops of current; one could therefore expect the needle to behave differently in the two situations.)

end of the helix to the other, allowing the effects of the current's circular component to become apparent. Helixes constructed in this way behaved exactly like permanent magnets, confirming Ampère's hypothesis.

These results raised one difficult question: Where are the electric currents in a permanent magnet? There are really only two possibilities. Either the currents flow in circles around the axis of the magnet as a whole or they flow in much smaller circles around each of the particles of which the magnet is composed. On Fresnel's suggestion, Ampère hypothesized that the currents flow around individual molecules of the magnet.

Such a hypothesis required experimental test; Ampère based his test on an experiment done by his friend Arago. Arago had wound a helix of copper wire, and at the center of the helix he had suspended an iron needle. When he connected the helix to a battery, the resulting circular electric current had magnetized the needle. Hence if Ampère's theory was correct, the circular current in the helix had created circular currents within the needle. But were the currents around the axis of the needle or around the individual molecules within the needle?

Ampère decided to answer this question by bending a thin ribbon of copper into a ring and suspending it within a cylindrical coil of insulated copper wire. The diameter of the ring was slightly smaller than that of the coil, and the ring and the coil were aligned so that they were concentric and their axes were parallel. If current in a helix had created circular currents around the axis of Arago's needle, Ampère reasoned, then a similar current should cause a circular current to flow around the copper ribbon, temporarily causing it to act as a magnet. He tested for such a current by holding a bar magnet up to the ring while current was flowing in the coil. If the ring had been magnetized, it would have been deflected by the bar magnet, but, as Ampère had expected, there was no deflection. Ampère cited this experiment publicly as weighty evidence for his molecular-current hypothesis.

Just at that time, however, his entire theory concerning the cause of permanent magnetism faced contradiction. In late 1821 an anonymous history of electromagnetism—written, it turns out, by the English physicist Michael Faraday—was published in England and translated immediately into French. The book outlined a pair of experiments designed to refute Am-

père's central hypothesis that permanent magnetism is simply the result of circular electric currents. According to Faraday, the experimental results had shown that the magnetism of permanent magnets was distinctly different from the magnetism of current-carrying helixes.

In the first experiment a helix made of insulated copper wire was wound around a wide, hollow glass tube. A tank was filled with water and the glass tube was half-submerged in the water, with its long axis parallel to the water's surface. Then a long, magnetized needle was placed on a cork floating near the tube and an electric current was passed through the helix. If the helix were exactly like a magnet, Faraday argued, the cork should float up to the glass tube and then stop, because then one pole of the needle (say the north pole) would be as close as possible to the opposite pole of the helix-magnet (the south pole). Instead the cork floated up to one end of the tube and then proceeded through it (down the center of the helix), until both poles of the needle rested directly under the similar (not the opposite) poles of the helix-magnet.

The second experiment examined the effect in the case of a true permanent magnet. A sheet of steel was rolled up to form a hollow tube, magnetized, and suspended halfway in the water. This time the cork floated up to the hollow tube and stopped directly under the magnet's south pole, confirming, said Faraday, that current-carrying helixes are not the same as permanent magnets, and thus that permanent magnetism is not the result of circulating electric currents.

Ampère found a way out. If the currents in magnets circulate around the individual molecules, he reasoned, then the center of a steel tube *should* be qualitatively different from that of a helix. In the helix the magnetized needle had been enclosed within the circulating current; in the steel tube, on the other hand, the compass needle had been outside all the many molecular currents. Hence one could expect the needle to act differently in the two cases. Having given this explanation (which is essentially correct), Ampère was now publicly committed to the concept of molecular currents.

The strength of Ampère's commitment became apparent in the summer of 1822, when he repeated the experiment that involved a circle of copper ribbon and a copper coil. This time, however, he tested the ribbon with a powerful horseshoe magnet instead of a weaker bar magnet, and the rib-

bon actually was deflected. Ampère's reaction to this result is somewhat puzzling. The result seems to contradict his theory of molecular currents, but he did not respond by testing the theory further. He made only a passing mention of the experiment in his report to the academy that September, with the astonishing remark that the effect had no theoretical significance. In fact, Ampère had unwittingly observed the electromagnetic induction of one current by another, but it was not until 1832, after Faraday had discovered and investigated electromagnetic induction, that Ampère realized how narrowly he had missed an important discovery.

Ampère's basic concepts of electrodynamics never changed after 1822. What did change was his ability to quantify his theory. After he had formulated his theory of permanent magnetism, the next task was to determine experimentally how strong the various electromagnetic forces are. Ampère decided that the fundamental interaction in electrodynamics is the force acting between two current-carrying wires, and so he set about the difficult job of measuring that force. Once again, his ability to hypothesize aided him. Unlike gravitational forces, which can be treated mathematically as forces acting between simple geometric points, the forces due to a current along a wire cannot always be treated as simple local phenomena.

Ampère's idea was to consider infinitesimally small segments of the current-carrying wires and to assume that the forces acting between such segments varied as the inverse of the square of the distance between the segments. He then found the total force between the two wires by integrating—by mathematically adding up all the infinitesimal units of force, taking into account the directions in which the forces might act depending on how the wires were bent. Ampère originally considered the special case in which the two wires lie in the same plane. He later generalized his result to take account of wires lying in separate planes, no matter how the planes are angled in relation to each other. That generalization made it possible to consider wires twisted in any way whatever in three-dimensional space. His end result was a famous and compact formula by which one can calculate the electrodynamic force acting between two wires as long as one knows the strength of the currents and the geometric arrangement of the

wires. In 1826 Ampère reworked his earlier research papers to produce his definitive book, *Mémoire sur la théorie mathématique de phénomènes électrodynamiques uniquement déduite de l'expérience* (Theory of Electrodynam-ic Phenomena Deduced Uniquely from Experiment).

After 1827 Ampère's health began to decline rapidly. He abandoned creative scientific research and turned to his final work on the philosophy of science. Here he rediscovered some of the inspiration of his early youth. He was enraptured by Gottfried Wilhelm von Leibniz' doctrine of pre-established harmony, which held that Man's mind is a copy, albeit an imperfect one, of the mind of God. Since Man's reasoning process is an image of God's reasoning process, Leibniz said, and since God's reason created the universe, the human mind should be able to understand the universe through a process of pure reason—in other words, there should be a preexisting harmony between the laws of the universe and the reasoning powers of mankind.

Ampère decided that the correspondence between God's reason, human reason and the inherent rationality of the universe should make it possible to use what is essentially a process of taxonomy to find the ultimate Truth. If one could outline all the sciences the human mind could possibly construct, Ampère argued, one would have the basic key to all possible truth, since the mind is structured in a way that corresponds directly with the structure of the universe. It would remain only to fill in the contents of the cosmic taxonomic chart.

Ampère produced many of these charts, intended to be tools of fundamental research, before he died, probably of pneumonia, in 1836. And so Ampère ended almost as he had begun: as an encyclopedist committed to the unity of all knowledge, for all knowledge is but the reflection of the unity of the divine mind.

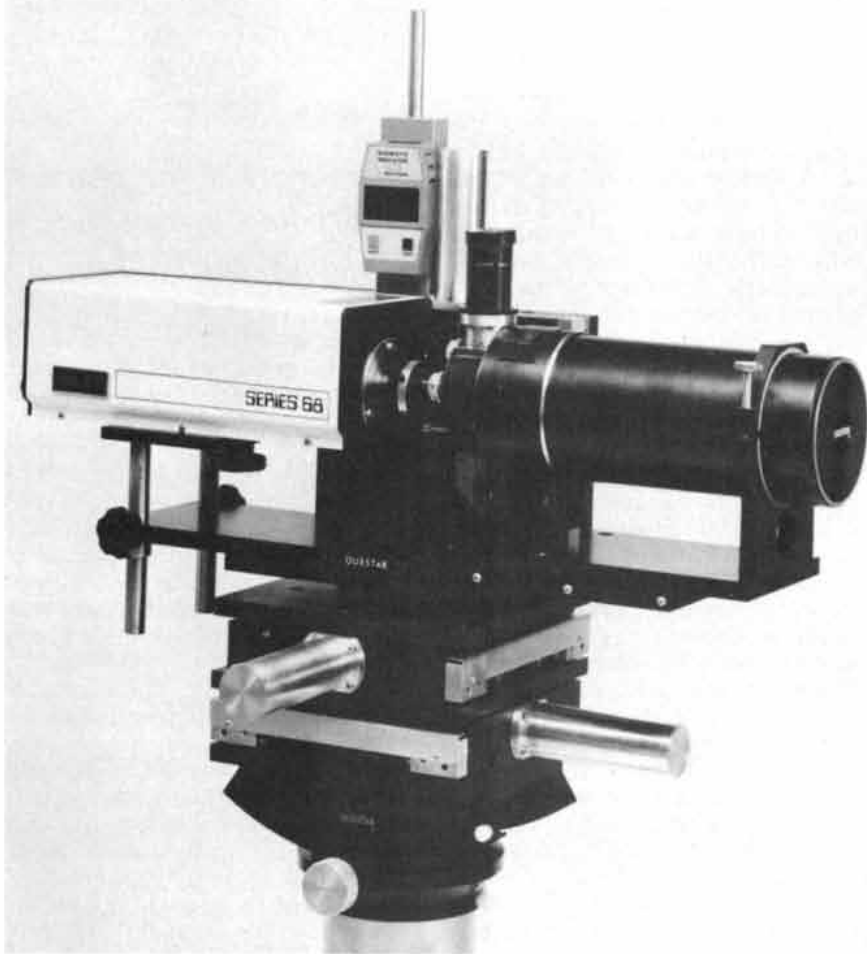
FURTHER READING

AMPÈRE. L. PEARCE WILLIAMS IN *Dictionary of Scientific Biography*, edited by Charles Coulston Gillispie. Charles Scribner's Sons, 1970.

A.-M. AMPÈRE ET LA CRÉATION DE L'ÉLECTRODYNAMIQUE (1820-1827). Christine Blondel. Comité des travaux historiques et scientifiques, 1982.

WHAT WERE AMPÈRE'S EARLIEST DISCOVERIES IN ELECTRODYNAMICS? L. Pearce Williams in *Isis*, Vol. 74, No. 274, pages 492-508; December, 1983.

A Questar® System for the Laboratory



In the laboratory this Questar visible-data imaging system is an indispensable tool capable of remote non-contact gauging and alignment to .0001 precision. It is a matched photo-visual system, with translation stages that position in 3 axes, with a digital indicator calibrated in ten thousandths. This is basically the resolution of the QM 1 which is the heart of the system. One of the unique features of the QM 1 is a variable focal length over a working range of 22 to 66 inches, covering fields of view from 1 to 32 mm. The equipment, shown above, is mounted on a specially designed, highly stable floor stand with rising centerpost and is easily portable.

This is no single-purpose system, but a remarkable Jack-of-all-trades that is used in projects as varied as crack propagation analysis, documentation of crystal growth, measurement of thermal expansion of a part in real time, or whatever difficult problem of measurement, imaging and recording has a need of solution. So why not start thinking about all the ways you can put this self-contained little laboratory to work, not only on your present project but in special applications on down the road.

© 1987 Questar Corporation

QUESTAR

P.O. Box 59, Dept. 215, New Hope, PA 18938 (215) 862-5277

Merchants of Peace

Defense contractors eye the impact of arms control



The market for peace, Stirling prospects, smart-card slump

A "very tough card game" is the way President-elect Bush has described one of his major legacies from the Reagan Administration, the strategic arms reduction talks (START). The talks are aimed at achieving bilateral reductions of up to 50 percent in strategic nuclear weapons. Bush has also said he will pursue a treaty to ban chemical weapons and will seek limits on conventional weapons. If peace continues to break out, what will be the consequences for the defense industry?

Certainly there is no prospect that a START treaty will cut defense spending drastically and release thousands of engineers and managers into civilian industry. Industry figures and arms-control advocates agree that no foreseeable treaty is likely to reduce defense procurement: weapons systems such as the MX missile and the Stealth bomber, which would be deployed even if a START treaty is concluded, are much more sophisticated and expensive than the systems they replace. Moreover, requirements for costly items such as reconnaissance satellites and the computer systems needed to analyze the torrent of information from orbiting eyes and ears would, if anything, increase.

In a 10-year forecast that assumes a START agreement will be reached within five years, the Electronics Industries Association concluded that although pressure on the Federal budget makes a 5 percent decrease in total defense procurement probable over the next five years, the value of electronic systems will be virtually unchanged during that period. The forecast predicts new spending requirements for the modernization of strategic forces as well as for "elaborate and expensive verification systems."

New kinds of weapons may also help to create additional markets. Simon Ramo, cofounder of TRW, foresees opportunities in a future shift in resources from strategic nuclear weaponry to high-technology conventional weaponry and to defensive and surveillance systems that would offset the Warsaw Pact's advantage in manpower and matériel.

What new spending would a START

treaty require? The form of any treaty cannot be predicted yet. The virtual impossibility of counting nuclear-submarine-launched cruise missiles without unacceptably intrusive inspections remains a major problem, and so does an impasse over whether mobile missiles should be banned. It does seem likely that instead of banning outright any class of missiles, a START agreement would confine strategic forces for most of the time to specified areas that could be inspected.

In the near term, monitoring compliance with a START agreement would rely largely on reconnaissance systems that are already planned, ones the intelligence services want regardless of any START treaty. Nevertheless, treaties strengthen the case for such systems. The Senate Select Committee on Intelligence is said to have successfully justified its demand for six new Lacrosse spy satellites costing a total of between \$6 and \$12 billion by citing the need to verify the intermediate-range nuclear forces (INF) treaty.

Even though the first systems are already planned, an arms agreement would place greater demands on them. A START treaty would put "an unprecedented level of pressure on National Technical Means [satellites and other sensors]," according to Roger L. Hagengruber, who is developing treaty verification systems at the Sandia National Laboratories. Although it is generally accepted that no verification system can be proof against all cheating, the challenge to technology is to ensure that any cheating is militarily insignificant.

The likely form of a START treaty will make satellite reconnaissance even more vital: a treaty that does not eliminate entire classes of weapons is intrinsically harder to verify than the INF treaty, which eliminates intermediate-range missiles altogether. Hagengruber says intelligence analysts would

require much more frequent and detailed images of the Soviet Union than are now available in order to be sure that illegal mobile missiles were not being concealed.

Kosta Tsipis of the Massachusetts Institute of Technology estimates that total spending on satellite and related reconnaissance is now in the region of \$20 billion per year (the figures are "black"—highly classified—as are all details of the systems). He "would expect the total to increase" as a result of a START treaty. According to Jeffrey Richelson of the National Security Archives, a private research group, major improvements in the state of the art are imminent. Future systems will be able to provide military commanders with images within hours of acquisition. Richelson says the Lacrosse satellites (newly developed by Martin Marietta) will carry synthetic-aperture radar, which is able to produce detailed images even at night or under cloud cover. (Some observers think the first Lacrosse satellite was scheduled for launch on the space shuttle Atlantis in late November of 1988.) Another satellite system, the KH-12, built by Lockheed Missiles and Space, is said to achieve a resolution of a few centimeters by dropping to a low altitude for "close-ups" and using a computer-controlled "rubber mirror," which changes shape to compensate for atmospheric distortion. Each KH-12 costs well over \$500 million.

Signals-intelligence satellites would play an essential role in verifying a START treaty by monitoring communications between command centers and mobile-missile crews, Richelson thinks. Tsipis believes infrared spectrometry—for detecting heat from industrial and military activity—will also play a growing role in monitoring a treaty. One concept being studied is an armada of lightweight, relatively inexpensive reconnaissance satellites that would provide reliability and coverage through redundancy.

Gleaning useful intelligence from the torrent of data such systems will produce is like "drinking from a firehose," according to Hagengruber. Artificial intelligence may come to play a role in such analysis, he says. Computers based on neural-network concepts are also being studied.

Compared with what is needed for electronic intelligence, the hardware employed on-site for treaty verifica-

Join the most exciting book club in science— where the only commitment you make is to reward your own curiosity

This is the Scientific American Library, where the mysteries of the human and natural worlds unfold before your eyes in lavishly illustrated volumes you'll want to read and refer to for years to come.

Here's why 75,000 readers like you have already become members of the Scientific American Library

■ Free Trial

You will receive a new volume to examine approximately every two months. (Your first volume is shipped immediately via UPS at no extra charge.) You have 15 days to look over each volume. If you decide to keep it, send in your payment. If you don't, simply return the book in the same shipping carton and we'll immediately credit your account.

■ No Prepayment and No Minimum Purchase

Send no money in advance. You pay only for the volumes you keep. You may choose as many or as few volumes as you like: there is no yearly minimum.

■ Low Members' Price

You may purchase any of these handsome volumes for only \$24.95—25% below the bookstore price.

■ No Strings, No Risk, No Commitment

You may cancel your obligation-free membership at any time for any reason, without notice, without penalty.



A one-of-a-kind book club for those who take science seriously.

Each volume in the Scientific American Library was created exclusively for this series, written by preeminent scientists—including many Nobel laureates—who describe their work in an exciting, accessible way.

Choose your premier volume

■ **Extinction**, Steven M. Stanley. In this geologic detective story, a renowned paleobiologist explores the causes of the mass extinctions that have periodically decimated thousands of thriving species. "Smack up to date scientifically...the best overview of the changing relationship between the physical environment and life... Enthralling..."

—John Gribbon,
New Scientist

■ **Molecules**, P. W. Atkins. A distinguished chemist takes us below the surface of visible reality to show how molecules make soap "soapy," determine the taste of barbecue, and give rise to the brilliant colors of fall foliage. "Undoubtedly the most beautiful chemistry book ever written.... It is more than a book of facts; *Molecules* is a work of art."

—John Emsley,
New Scientist

Some of your future selections

■ **Islands**, H. W. Menard. "Breathtaking photographs, lavish illustrations...and lively narrative.... It is a book that only Menard, [with his] encyclopedic knowledge of the oceans...could have written." —*Science*

■ **The Timing of Biological Clocks**, Arthur T. Winfree. "Winfree stands unique in the science of biological time-keeping by visiting all fields, from population dynamics...to biochemistry.... Brilliant insights." —*Nature*

■ **Einstein's Legacy**, Julian Schwinger. "Delightful.... An ideal gift for the curious non-expert." —*Nature*

Join us. You have nothing to lose. And a world of scientific discovery awaits.

To become a member of the Scientific American Library, simply detach and mail the facing reply card.

Or write to: Scientific American Library, Dept. A, P.O. Box 646, Holmes, PA 19043



TAGGING SYSTEM for missiles was developed at the Sandia National Laboratories. An identification number on a missile's casing is covered with a hard epoxy resin containing random, reflective particles of mica. The reflections under light from different angles (compare the left and right images) form a unique signature.

tion is small beer, built largely from off-the-shelf parts. One reason is the remoteness of the places where the systems will be used; another is the desire to keep the level of technology low in order to protect U.S. secrets from Soviet inspections and keep sensitive U.S. technology out of Soviet hands.

Nevertheless, the field offers challenges for companies specializing in systems analysis and implementation. For example, the U.S. has proposed that approved missiles be "tagged" in a way that identifies them for counting. Any missile lacking a tag would presumably have been produced illegally. At the Argonne National Laboratory investigators are examining the possibility of using portable electron microscopes to examine missiles for minute surface scratches that cannot be copied. Electronic tags have also been investigated, although some observers fear they might be susceptible to copying by "reverse engineering." For ensuring that conventionally armed cruise missiles are not illegally converted to nuclear use, several elaborate schemes involving tamper-proof seals have been proposed. Critics argue, however, that the seals would simply be broken in a crisis.

Several national laboratories are working on sensors for counting the number of warheads on missiles leaving controlled sites without removing the warheads' shrouds. Sensors relying on both gamma-ray and neutron beams are under study. Alexander Devolpi of Argonne says the challenge is to create a system that can count warheads without revealing significant details about their design. Laboratories are also working on unmanned sensors that would ensure that illegal items do not leave controlled areas. John R. Harvey of the Lawrence Liver-

more National Laboratory has what he describes as a "far-out idea": to use robots for patrolling compounds.

Contractors are reluctant to reveal their plans for verification hardware, but some idea of the on-site requirements for START can be had by comparing it with INF. Hughes Aircraft won a \$24-million contract to provide personnel for the one permanently monitored inspection site in the Soviet Union specified by the INF treaty, the missile assembly plant at Votkinsk. A new X-ray scanning system, made by Bechtel National and costing about \$10 million, will be used at Votkinsk to detect any prohibited missiles in containers leaving the site. A congressional source estimates the total cost of on-site inspections under the INF treaty—which also provides for short-notice inspections at other designated sites—at about \$200 million per year for 13 years. On-site inspection for a START treaty might cost a total of \$4 billion over 20 years, although the number of sites to be inspected would probably be from 10 to 20 times larger. The amount of hardware would still be small.

Actually dismantling weapons for a START treaty is itself likely to become a specialized industry worth about \$1 billion. Indeed, it might amount to a kind of high-technology mining. Theodore B. Taylor, a former nuclear weapons designer, estimates that if warheads were dismantled, the value of the enriched uranium potentially freed for use in civilian power generation would be several billion dollars.

Other treaties besides START will create small markets. The Natural Resources Defense Council is installing a network of unmanned seismic monitoring stations in the Soviet Union worth \$500,000 each; it hopes eventually to have more than 20 stations

and attract Government support. Remote sensors would also probably be needed to monitor a chemical-weapons treaty as well as a conventional-arms treaty. In each case "the idea is to accomplish the job with the lowest technology you can," Harvey says.

The other common element in all efforts to capitalize on arms control is, of course, uncertainty. Hagengruber has been approached by several companies eager to exploit the verification technologies being developed at Sandia but says they have been frustrated by the lack of clear technical requirements during protracted secret negotiations. If all goes well, however, verifying the peace may turn out to be a growth industry. —Tim Beardsley

Power to Burn

The Stirling engine begins to look practical

Cars have a delicate digestion—their internal-combustion engines can survive only on a diet of clean, high-grade gasoline or diesel fuel. External-combustion engines are much less discriminating, and after a decade of research by the National Aeronautics and Space Administration and its contractor Mechanical Technology Incorporated (MTI) one may finally be ready for the marketplace. It is the Stirling engine, a long-standing favorite of people seeking alternatives to internal combustion but one that until now has not been available at a reasonable price or performance level. The Stirling external-combustion engine runs equally well on gasoline, kerosene, ethanol or virtually any other combustible liquid. With some minor modifications it can burn natural gas and perhaps even powdered coal.

Conventional internal-combustion engines earn their name by detonating mixtures of fuel and air inside piston chambers. Hot, expanding gases from the explosions push against the pistons and provide the driving force. Unfortunately the explosions also produce gummy, partially burned hydrocarbons that can stick to the engine and eventually impede the piston's motion. To minimize this buildup cars burn only refined petroleum.

Stirling engines avoid buildup problems by isolating combustion in an external chamber. The heat generated is transferred to a sealed "working fluid"—usually hydrogen or helium—that expands and moves the pistons. Since no combustion by-products can clog the moving parts, operators of

Stirling engines have a much wider choice of fuels. In addition, because the fuel burns continuously and completely, Stirling engines incinerate many of the pollutants released by internal-combustion engines. Stirling engines therefore do not need expensive catalytic converters to curb their emissions.

Although models of Stirling engines have existed since the early 19th century, when the engine was invented by the Scottish clergyman Robert Stirling, engineering problems have until recently hampered their development. Working Stirling engines were built in the Netherlands during World War II and were later improved on in Sweden. They included high-temperature parts made from alloys containing cobalt, however, a scarce strategic metal. Moreover, these Stirling engines produced too little power to compete with less troublesome internal-combustion designs.

Interest in Stirling engines reignited following the oil crisis of the 1970's. In 1978 the Department of Energy and NASA began funding a 10-year, \$100-million research effort into making the engines more practical. Materials research has led to lighter, less expensive metals that can replace cobalt alloys, as well as to better seals that stretch the useful lifetime of the engine. Fast-acting microprocessor controls also quicken the engine's response to the accelerator.

According to MTI, the new Mod II Stirling engine is competitive in size, weight and performance with conventional spark-ignition engines. Experimental engines installed in U.S. Air Force vans have logged more than 18,000 miles burning a variety of fuels, and they have shown consistently good performance. The Air Force was gratified to learn that the vans could use a large and costly petroleum supply that had previously gone to waste: dirty fuel drained from jets undergoing repairs, which cannot be recycled to other aircraft. Last May a Mod II was installed in a U.S. postal van for more extensive driving tests under more varied weather conditions. Although official Environmental Protection Agency measurements are not yet available, MTI project manager William D. Ernst expects that the Mod II will boost the postal van's fuel efficiency by more than 20 percent.

Regardless of these preliminary successes, Stirling engines will probably not be installed in commercial automobiles until manufacturers have had years of experience mass-producing them. Stirling engines are more likely

to prove themselves first as power sources for electric generators, heat pumps, irrigation pumps and military vehicles.

—John P. Rennie

Not in the Cards?

The smart-card revolution is still waiting to happen

A few years ago "smart cards," plastic credit cards with microprocessors embedded in them, were touted as the coming wave in the credit-card business. Consumers would carry portable data banks containing everything from account balances to fingerprints. Credit-card fraud would go down and convenience would go up. But the flight from plain plastic to silicon never quite took off. Smart cards are still "a technology waiting for a good application," says Einar Asbo, manager of technology applications at Visa International.

Smart credit cards put the data and intelligence needed to authorize a sale into the card itself rather than into a computer at a clearinghouse or a card-issuing bank. The smart card keeps its owner's current account balance in memory, updating the balance with each purchase (and with each verified payment), and a personal access code (PIN) ensures that the person making a purchase is in fact the legitimate owner of the card. A single smart card can carry the data for multiple accounts, thereby putting an end to the prestige of a walletful of plastic.

Whereas conventional credit cards can be copied with readily available magnetic read-write equipment, the smart card can encrypt its data and release it only in response to specified passwords. "You'd need a factory" capable of manufacturing smart cards from scratch in order to copy or alter a card illicitly, says Robert B. J. Warnar of the National Institute of Standards and Technology.

Issuers of credit cards, governments and card manufacturers, among them Bull S.A., the French computer maker, have tried to find a good use for the smart card. Perhaps the most concerted effort has been in France, where Carte Bleue in 1985 ordered 12 million smart cards from Bull. About a million have been delivered so far. In the U.S., MasterCard championed the smart card, conducting 18-month test programs in Maryland and Florida and issuing more than 30,000 cards. And in Japan, Visa International ran experiments beginning in 1986 and continues to push smart cards, including a

"supersmart" card with a keyboard and display developed in partnership with Toshiba.

Why, then, have smart cards not become commonplace? In the U.S. lack of an agreement on standardization between MasterCard and Visa has put widespread use on hold since the end of MasterCard's trials in mid-1987, according to John C. Elliott, who guided MasterCard's development efforts.

More to the point, do smart cards offer a more cost-effective way of doing most of the things dumb cards do today? The ability of the smart card to validate transactions without approval from a central clearinghouse is valuable only where telecommunication is expensive. (In one such place, Norway, an association of banks has ordered half a million smart cards for electronic transfer of funds.) And even though smart cards could reduce credit-card losses from fraud and nonpayment, losses today are low enough—about \$1 per card every year—for the added security not to justify the extra cost of a smart card.

Meanwhile the 2,000 to 8,000 characters of storage on a smart card are looking for useful work. Current magnetic-stripe cards can hold about 240 characters of data, enough for account number, name, address, issuing institution and airline-seat preference with 100 characters to spare. The added financial information on a smart card fills only a little extra space. Test marketers in the U.S. and elsewhere are putting everything from medical records to brand-preference data in the unused space, but no application has found widespread acceptance.

Smart cards are finding niches primarily as debit cards rather than credit cards. Debit cards—ones a customer pays a specific price for and can then redeem for goods and services—based on magnetic stripes have typically been limited to low-value applications such as parking garages, mass-transit fares and photocopying machines. If a magnetic debit card were worth more than a few dollars, the incentive for copying it would be too high.

In France smart debit cards eliminate the need for small change to feed telephones and parking meters. The U.S. Marine Corps issues smart cards instead of scrip to recruits on Parris Island. Debit cards are particularly attractive to financial institutions, of course, because the cardholder pays up front, and the institution gets the use of the money until the transaction clears. In the existing credit-card system the cardholder takes advantage of the float.

—Paul Wallich

THE AMATEUR SCIENTIST

The colors seen in the sky offer lessons in optical scattering



by Jearl Walker

The colors of the sky during daylight and twilight have been notoriously hard to explain even though they have been scrutinized scientifically for well over a century. Why is a clear daytime sky mostly blue but white near the horizon? Why is the setting sun normally red and the sky just above it a tapestry of colors? At twilight why does a curved shadow with a rosy border rise in the eastern sky? Why does a purple patch sometimes appear and then fade in the western sky soon after sunset, and why does another purple patch sometimes appear as much as two hours later? These questions call for studies of how light interacts with molecules and airborne particles. For some of the questions definitive answers are still being sought.

There has been no shortage of explanations for why a clear sky is largely blue. Many popular schemes involve the scattering of sunlight from such airborne materials as dust, aerosols, ice crystals and water droplets; others depend on absorption of the red end of the visible spectrum by the water and ozone in the atmosphere. The inadequacies of these accounts were reviewed in 1985 by Craig F. Bohren

and Alistair B. Fraser of Pennsylvania State University, who also pinpointed the correct explanation, one that was introduced in 1899 by Lord Rayleigh.

Rayleigh had been slow to accept his own explanation, in part because of findings published in 1869 by John Tyndall, the British physicist who is remembered for his skill at making science accessible to nonscientists. Tyndall demonstrated how artificially produced smog took on "a colour rivalling that of the purest Italian sky" when it was illuminated with a beam of white light and viewed at an angle to the beam. For years thereafter many investigators, including Rayleigh, believed the scatter of light from particles produced the blue of the sky. Moreover, they concluded that a pure gas such as air cleansed of all particles would not be able to scatter light and break it up into different colors.

In his publication of 1899 Rayleigh finally stated that the scattering and color separation were due to the air molecules themselves—that "even in the absence of foreign particles we should still have a blue sky." By then he had constructed an elegant model of how a molecule scatters light. To understand his model, consider an

air molecule (it makes no difference which kind) illuminated by white sunlight. The light is a composite of all the colors in the visible spectrum; a wavelength is associated with each tint. The wavelength increases from blue to green, yellow and red; the wavelength associated with red light is about 1.68 times the wavelength associated with blue light.

Each color component in the sunlight is scattered from the molecule in all directions but not with uniform intensity. The brightest scatter is in the forward direction (as if the light passed directly through the molecule) and in the backward direction (back toward the sun). Light scattered at a right angle to the sunlight's initial path is only half as bright. All the colors scatter in this pattern, but the intensity scattered in any particular direction is different for each color. Rayleigh found that the intensity depends on the inverse fourth power of the wavelength. Therefore short-wavelength light (say blue) is more strongly scattered than, say, red light, which has a long wavelength. Since the ratio of their wavelengths is about 1.68, the blue scattered light is 1.68^4 (or about eight) times as bright as the red scattered light.

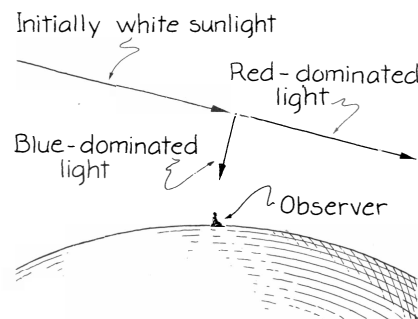
Suppose you intercept the light that is scattered to the side at an angle of about 90 degrees to the initial direction of the sunlight. If you could perceive the light from a single molecule, it would be bluish, because the blue end of the spectrum would be brightest. The picture is more realistic when abundant molecules scatter light to you, so that the light and its color are perceptible. That is the case when you look up at an area of the sky away from the sun. All the molecules along your line of sight scatter light to you that is dominated in intensity by blue; that part of the sky appears to be bluish—not pure blue, because you also intercept the other, fainter colors.

The fact that the blue end of the spectrum is strongly scattered out of the initial beam of light means the beam continuing through the atmosphere gradually becomes dominated by the red end of the spectrum. If you look toward the sun when it is high, the light reaching you has traversed too little of the atmosphere to be appreciably reddened. When the sun is low and the sunlight takes a longer path through the atmosphere to reach you, the light is noticeably reddened, and so sunsets are dominated by the red end of the spectrum.

You may detect an apparent contra-



Two kinds of scattering pattern



Color separation by Rayleigh scattering

diction in this argument. Blue light is always more strongly scattered than red light—in any direction. The statement even applies to the light that is scattered forward and continues in the initial direction of the beam. If blue light is more strongly scattered forward than red light, why does the continuing beam redden?

James A. Lock, a colleague of mine at Cleveland State University, turns to a particulate description of light to show there is no contradiction. Suppose the initial beam of light has 1,000 red and 1,000 blue photons. When the beam reaches a group of molecules, the number of blue photons that scatter in all directions (or in any particular direction) is eight times the number of red photons. Suppose a total of 80 blue and 10 red photons are scattered in all directions, of which eight blue photons and one red photon are scattered forward. In the continuing beam, then, there are 991 red photons but only 928 blue ones, and so the beam has reddened.

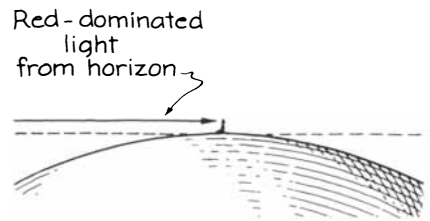
You might also question how it is that light can be scattered through a molecule in the forward direction. A molecule is not a solid barrier like a wall; rather it is largely empty, with electrons in various orbits around a remarkably small nucleus. In a classical description of scattering, the electric field of the passing light forces the electrons to oscillate, and energy is transferred from the light to the oscillations. When a charged particle such as an electron is forced to oscillate, it radiates light in all directions except directly along the line on which it oscillates. This newly emitted light is the “scattered” light, and of course some of it is emitted in the forward direction.

Bohren and Fraser raised and then removed a possible objection to the Rayleigh explanation of the blue sky. The very shortest wavelengths in the visible spectrum correspond to violet rather than blue. Why then is the sky not violet? Bohren and Fraser gave two reasons. One minor reason is that because the initial sunlight is somewhat weaker in violet than in blue, less violet than blue is scattered to you. A more important reason is that the human eye is much less sensitive to violet than to blue.

People occasionally attribute the blueness of the sky to the water vapor in the atmosphere, perhaps because bodies of water are often bluish. One reason a lake can be blue is that when white light passes through several meters of water or more, the water mole-

cules partially absorb the red end of the spectrum, and the light eventually reflected to a viewer is left primarily blue. Bohren and Fraser pointed out that the atmosphere has too little water to make this kind of absorption significant in the bluing of the sky.

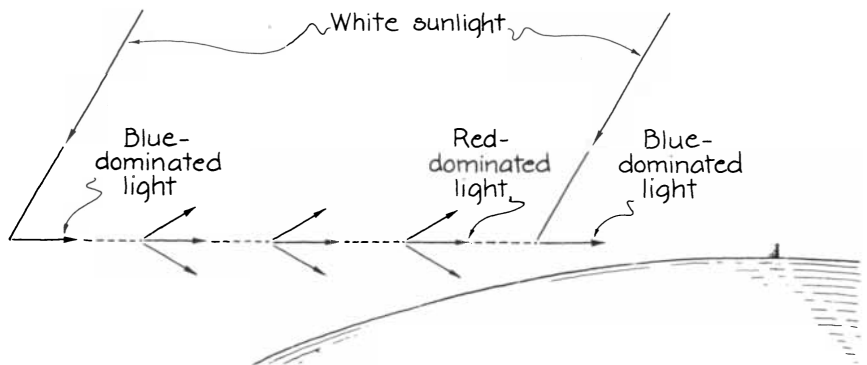
The sky's color has also been attributed to the layer of ozone molecules that extends from about 10 to about 40 kilometers in altitude, with a peak density at about 25 kilometers. The molecules have absorption bands at the red end of the spectrum. Perhaps the red components of sunlight are weakened as the light passes through the ozone layer, so that the light finally reaching the ground is dominated by blue. Bohren and Fraser argued that although there is certainly depletion of the red by ozone, it plays a minor role in determining the blue of the sky. When you look up at the daytime sky, you intercept light that has passed through too little ozone for the absorption to matter. At twilight, when the rays travel a slanted (and hence longer) path through the ozone layer to reach you, the ozone absorption is more important, but even then the main reason for the sky's blueness is the Rayleigh mechanism.



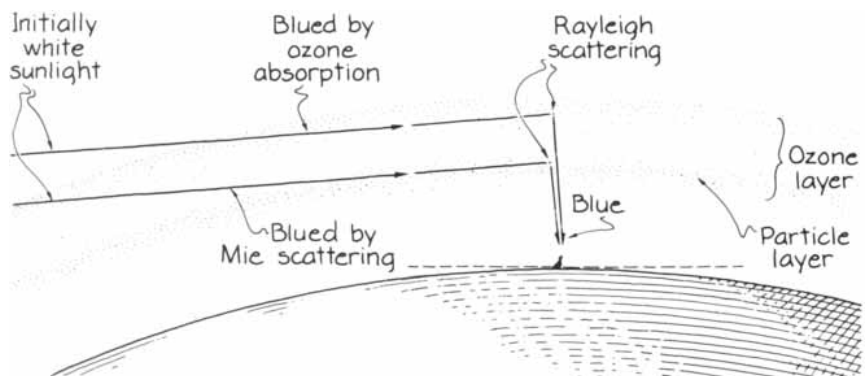
The color of the setting sun

In daytime the blue of the sky dulls toward the horizon, and in a band of about five degrees from the horizon the sky is often white. Since the air molecules along a line of sight to the horizon must, like any others, scatter light by the Rayleigh mechanism, what can have happened to the blue? Bohren and Fraser explained that the lack of color is due to the long path through the atmosphere taken by light reaching you when you look toward the horizon rather than looking up. The added distance means the light scatters many more times before it reaches you.

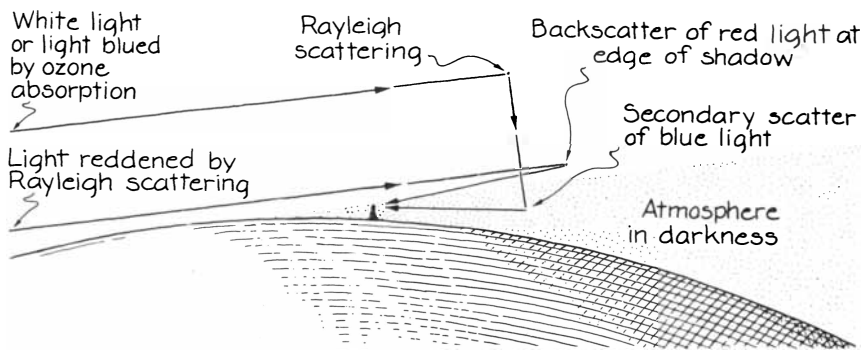
Some light scatters from molecules that are not very distant [see illustration directly below]. From them you receive light that is dominated by the



Why the daylight sky is white near the horizon



Why the blue is enhanced at the zenith just after sunset



The colors seen at the top edge of the earth's shadow

blue. Much more distant molecules also scatter blue-dominated light in your direction, but your distance from those molecules means that the light undergoes repeated scattering before it reaches you. In each scattering event the light scattered toward you is light scattered in the forward direction, so that its blue component is weakened; after many scattering events that light ends up being dominated by the red half of the spectrum. The result is that you receive mostly the blue half of the spectrum from nearer molecules and mostly the red half from more distant molecules. The combination is white, and that is the color of the sky in the direction of the horizon.

The same effects explain the colors of dark mountains seen on a clear day. If the mountains are not too distant, their image is bluish, because blue-dominated light is scattered by the air molecules between you and the mountains. Somewhat more distant mountains may be even bluer, but those that are very far off are white—just as the horizon is.

According to Bohren and Fraser, the light from a setting sun would actually be orange (between red and yellow) instead of red if it scattered only from air molecules along its way to you.

They pointed out that the reason the color is instead usually a rich red is that the light normally scatters not only from molecules but also from very small particles and aerosols in the atmosphere.

When you look in a direction close to the sun at any time of day, you intercept some of its bright light scattered forward by those same small particles and aerosols, and so that region of the sky is brighter than it would be if the particles were absent. When the sun is high, its surround is bright white. When it is low, on the other hand, the light reaching the particles has already been reddened by Rayleigh scattering and the surround is bright red. The greater the density of the particles, the brighter the surround and the sharper the circumference of the setting sun.

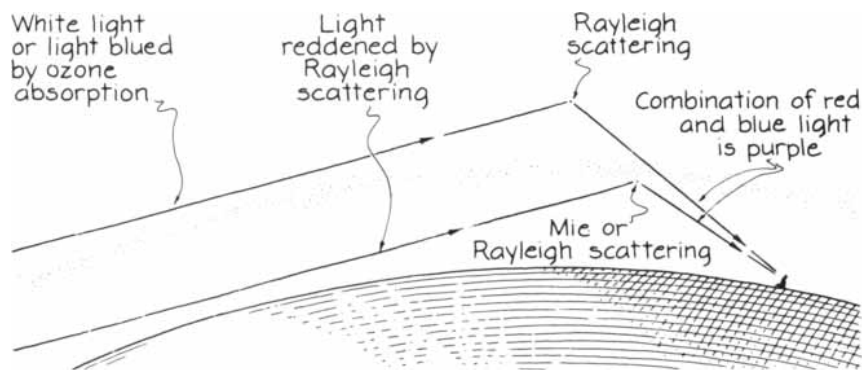
I have been assuming that the particles and aerosols alluded to so far are smaller than about .1 micron, and that as a result they scatter light by the Rayleigh mechanism just as molecules do. Particles that are somewhat larger than .1 micron scatter light by a much more complex mechanism, which is called Mie scattering after Gustav Mie, the German physicist who developed a theoretical model for such scattering

at the beginning of this century. Mie scattering by a sizable particle is actually a form of diffraction, with most of the light being sent forward in a narrow cone. Red light is spread over a wider cone than blue light, and so the continuing beam becomes bluer. (In simple diffraction the light waves that strike a particle spread out and also travel around the particle and into its "shadow region." In Mie scattering the behavior of the light is harder to interpret, and I shall not consider its details.)

During twilight on a clear evening the zenith (the sky directly overhead) becomes bluer than it is during the day. The extra blueness seems strange, considering that the horizon near the sun may be quite red. Several explanations for the bluing have been given, the likeliest of which involves the ozone layer. When light from the sun takes a slanted path through the layer at sunset, the ozone's absorption of the red end of the spectrum leaves the beam dominated by the blue end in spite of the Rayleigh scattering the beam encounters along the way. The beam becomes even more dominated by blue light if it also slants through a layer of particles large enough to introduce Mie scattering. After the light has been blued by either mechanism or both of them, some of it then scatters to you from the zenith by the Rayleigh mechanism, and you see a purer blue coming from there than you do in the daytime.

Just after sunset the shadow of the earth begins to rise from the eastern horizon. The border of the shadow is usually red or rosy purple. The color is due to light that has been reddened by Rayleigh scattering during its long passage through the lower reaches of the atmosphere. Near where you see the top "edge" of the shadow some of the light undergoes Rayleigh scattering and returns toward you. When you intercept the light, you perceive red at the top edge.

Below the edge the upper part of the shadow may be a faint blue. The blue tint is probably due to sunlight that travels through higher, less dense parts of the atmosphere, where the blue component of the beam is not weakened as much as it is in a beam that passes lower in the atmosphere, which encounters more air molecules. The light may actually be appreciably blued if it passes along a slanted path through the ozone layer or through a particle layer that forces Mie scattering. Near the earth's shadow some of the light undergoes Rayleigh scatter-



How the purple patch is formed at twilight

SCIENTIFIC AMERICAN

CORRESPONDENCE

ing and travels into the shadow, where it again undergoes Rayleigh scattering before it heads toward you. Since this scheme involves multiple scatters, the blue light you finally intercept from the shadow region is dim. It is perceptible because you see it against the dark shadow of the earth.

About 10 minutes after the sun has set a purple patch occasionally develops above it somewhere between 30 and 75 degrees from the zenith. The patch, which is often called the "purple light," seems to depend on the presence of a layer of particles at an altitude of from 16 to 20 kilometers, in the lower part of the ozone layer. The particles might be dust from a desert or fine ash from volcanic eruptions or large forest fires.

In 1967 Aden B. Meinel and Marjorie Pettit Meinel of the University of Arizona pointed out that such a layer might be formed not by particles but by another product of volcanic eruptions. If an eruption emits a large amount of sulfur dioxide, when the gas mixes up to the base of the ozone layer, it interacts with the ozone to produce sulfates. When the sulfates precipitate onto condensation nuclei, a Mie-scattering aerosol is formed.

In their beautiful book *Sunsets, Twilights, and Evening Skies* the Meinels have suggested that the purple patch is the product of very red light and very blue light that is scattered from different regions of the sky. The red component comes from sunlight that has skirted the earth, passing through so much atmosphere that Rayleigh scattering has made the light red. Some of this light scatters to you from the particle layer—presumably by Mie scattering if the particles are large enough and by Rayleigh scattering if they are smaller; in any case you receive extra red light because of the presence of the particles.

The blue component comes from sunlight that passes through higher parts of the atmosphere and so does not redden as much. (I might add that since the light passes through the ozone layer along a slanted path, it may be dominated by the blue end of the spectrum because of absorption.) Some of the light undergoes Rayleigh scattering, and blue light is sent toward you. Both the red and the blue components travel along your line of sight when you view the patch, and the combination creates the impression of purple light.

The reason other parts of the sky are not purple is that you receive different mixtures of colors, rather than simply

red and blue, when you look toward them; they may have a variety of hues that depend on your angle of view. The colors are particularly brilliant when the particle layer is dense and extensive, as it often is after a major volcanic eruption. The 1883 explosion of Krakatau near Java produced brilliantly colored sunsets for about five years, and sunsets were colored by the 1963 explosion of Agung on Bali for about three years.

A second (but much rarer) purple light, which appears in about the same part of the sky as the first one but between an hour and a half and two hours after sunset, is still not well understood. Some twilight watchers think it is caused by the same particle layer as the first purple light. If the layer is extensive, some of the light that scatters from the part of the layer well below the horizon may scatter again from the part of the layer that is in view. Provided the light is bright enough, you would then see a faint purple patch.

An alternate explanation involves a second layer of particles at an altitude of from 80 to 90 kilometers, in a region of low temperature at the junction of the mesosphere and the ionosphere above it. These particles may be of terrestrial origin but are more likely from space; the earth intercepts a vast amount of comet and asteroid debris that may occasionally produce an extensive particle layer in the low-temperature region. Sunlight scattered by the layer is certainly too faint to be perceived during the day or even at early twilight, but it might be seen when most of the atmosphere in view is in the shadow of the earth and the layer is still illuminated by the sun. The increasing amount of light pollution from urban environments may well make observations of this second purple patch even rarer in the future.

FURTHER READING

POLARIZATION AND SCATTERING CHARACTERISTICS IN THE ATMOSPHERES OF EARTH, VENUS, AND JUPITER. David L. Coffeen in *Journal of the Optical Society of America*, Vol. 69, No. 8, pages 1051-1064; August, 1979.

RAYLEIGH SCATTERING. Andrew T. Young in *Physics Today*, Vol. 35, No. 1, pages 42-48; January, 1982.

SUNSETS, TWILIGHTS, AND EVENING SKIES. Aden Meinel and Marjorie Meinel. Cambridge University Press, 1983.

COLORS OF THE SKY. Craig F. Bohren and Alistair B. Fraser in *The Physics Teacher*, Vol. 23, No. 5, pages 267-272; May, 1985.

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

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of \$1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/88. \$1.25 + 0.00.

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

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51593. The date of the last issue on subscriptions appears at the right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name _____

New Address _____

Street _____

City _____

State and ZIP _____

Old Address _____

Street _____

City _____

State and ZIP _____

COMPUTER RECREATIONS

People puzzles: theme and variations



by A. K. Dewdney

"It seems that the analysis of character is the highest human entertainment."

—Isaac Bashevis Singer,
New York Times Magazine,
November 26, 1978

Three philosophers of ancient Greece once took a noontime walk in the country near Athens. Finding shade under an olive tree, they unplugged a wine flask and began a quiet discussion of the fundamental ontological question: Why does anything exist? The discussion grew heated, then confused and rambling. Soon afterward all three philosophers fell asleep in the shade of the tree.

Later that afternoon an Athenian youth who was bent on mischief spied the three slumbering philosophers. He gently splashed drops of white paint on their foreheads. Just before sunset an owl that made its home in the tree sidled out onto a branch above the three men. It hooted once loudly and then flapped noisily away. The owl's cry awakened the philosophers, each of whom immediately assumed that the owl was responsible for his two colleagues' decorated foreheads. They all began to laugh.

The sight was no doubt amusing, and perhaps five seconds passed before one of the men abruptly stopped laughing. Why? The puzzle can be solved by accounting for the thought process of the one philosopher who stopped laughing. In doing so, one has to consider what he must have supposed his cohorts were thinking.

Many puzzles involve people merely to provide a human context in which the would-be solver feels more comfortable. But the solutions to problems I call people puzzles, such as that of the three philosophers, depend on thinking about what each person in the puzzle is thinking the other

people in the puzzle are thinking. Such reasoning about reasoning is not only an amusing theme for puzzles but also a necessary one for computer scientists who seek to create programs that mimic the way a human being might think about similar perplexing situations.

The basic theme of this column is not complete without the solution to the puzzle of the three philosophers. Let us call the philosophers Pythagoras, Plato and Aristotle. Pythagoras, the oldest and wisest, is the one who stopped laughing. Here is the reason: Pythagoras looked at Aristotle chortling away and realized that Aristotle had no idea his own forehead was anointed with a white substance. If he (Pythagoras) had a clean forehead, then Aristotle was evidently laughing at Plato. But who then did Aristotle think Plato was laughing at? "Great Athena!" Pythagoras must have exclaimed to himself, "I shouldn't be laughing." The situation is depicted in cartoon form on the opposite page.

That is the answer to the more or less traditional form of the puzzle. The puzzle can be extended somewhat by asking why a few seconds later Plato also stops laughing. As soon as Pythagoras quits laughing, his line of thought is no longer open to Plato. Indeed, it would seem that the belief that one's own forehead is clean is now reinforced in Plato's mind. "Pythagoras evidently saw my clean forehead and realized that Aristotle was laughing at him," Plato might think. Yet if Plato cogitates a bit more, he will deduce his own befouled state. I shall let the reader step into Plato's sandals to determine how.

Consider a variation of the three-philosophers theme. A sultan wanted to choose the wisest of three candidate viziers for the position of grand vizier. The sultan took the viziers into

a darkened room and put a white hat on the head of each. He then led them back to the throne room and told them: "Each of you has been given a hat that is either white or black, and at least one of the hats is white. The first of you to tell me the color of your hat will get the job of grand vizier."

The puzzle of the three viziers is essentially the same as that of the three philosophers, even though there is no laughter to serve as an indicator of what each vizier sees. In a curious way the shared knowledge that at least one hat is white works hand in hand with the silence of indecision to produce an equivalent effect.

Let us call the candidate viziers al-Khwarizmi, ibn Khaldun and ibn Sina. Here are al-Khwarizmi's thoughts, which win him the position:

"H'm, I wonder if my hat is black. If it were, what would the other viziers think? Suppose ibn Khaldun too thought his hat were black. In that case he would realize that ibn Sina would see two black hats and deduce immediately that his own is white. Yet ibn Sina has not cried out, 'My hat is white!' hence ibn Khaldun knows that ibn Sina sees at least one white hat. But if my hat were indeed black, then ibn Khaldun would know that his is the white hat ibn Sina sees, and he would say so. Yet ibn Khaldun has not done that. By the beard of the Prophet, my hat cannot be black!"

The three-vizier puzzle has a meta-puzzle level pointed out by one of my colleagues: as soon as the viziers are told the nature of the contest, each can deduce that the only fair test of their cognitive abilities would in fact require all three hats to be white.

One can easily rewrite the puzzles of the three philosophers and the three viziers in an endless number of ways. A well-known variation involves three aristocratic women traveling by train through the English countryside at the turn of the century. It is a hot day and they have opened their compartment window to let in some fresh air. The train, belching thick black smoke, enters a long tunnel. When it emerges into daylight, all three women simultaneously break into laughter at the sight of their traveling companions' soot-covered faces—until one of them ceases to laugh.

Here is a second variation: Three businessmen lunch on spinach quiche at a Manhattan expense-account restaurant. All three get spinach on their teeth... The realization that a people puzzle can take many different equivalent forms leads to the idea of puzzle transformations. How does one trans-

form philosophers into viziers, so to speak? It helps to identify the elements of one puzzle that have corresponding elements in the other:

1. mischievous youth → sultan
2. philosopher → vizier
3. forehead → head
4. white mess → white hat
5. laughed → kept silent

In addition to these elements, the transformation must also define the peculiar characteristic of a philosopher or a vizier as operationally defined in the puzzle. A philosopher is someone who, until he analyzes the situation, will laugh if he sees at least one befouled forehead but will otherwise not laugh. A vizier is someone who, until he analyzes the situation, will say nothing if he sees at least one white hat. Otherwise he will say, "My head has a white hat on it."

Readers are invited to fill in the blanks in the generic people puzzle that follows. Filling in the numbered blanks with the corresponding words in the left-hand column of the table above will result in the three-philosophers puzzle. If the blanks are replaced with the corresponding words in the right-hand column, the three-viziers puzzle emerges. (One can, of course, come up with one's own set of words that would make sense if they replaced the blanks.) Colorful details of time and place have been eliminated to simplify the demonstration.

Once upon a time a (1) _____ put a (4) _____ on the (3) _____ of each of three (2) _____s without their knowing it. As long as he did not know that his own (3) _____ had a (4) _____ on it, each (2) _____ (5) _____. Suddenly the wisest of the three (2) _____s no longer (5) _____. He then exclaimed, "My (3) _____ has a (4) _____ on it!" How was he able to deduce this?

The next variation I shall present on the theme of the three philosophers originally concerned unfaithful wives. Because it involves sexual stereotyping, I shall take advantage of a simple transformation to alter it. In any event the puzzle is interesting because it generalizes the problem of the three philosophers.

The tyrannical queen of the Amazons announced one day to her subjects that at least one husband in her realm had been unfaithful. She then issued a proclamation: "If any of you discovers your husband to be unfaithful, I order you to execute him at the

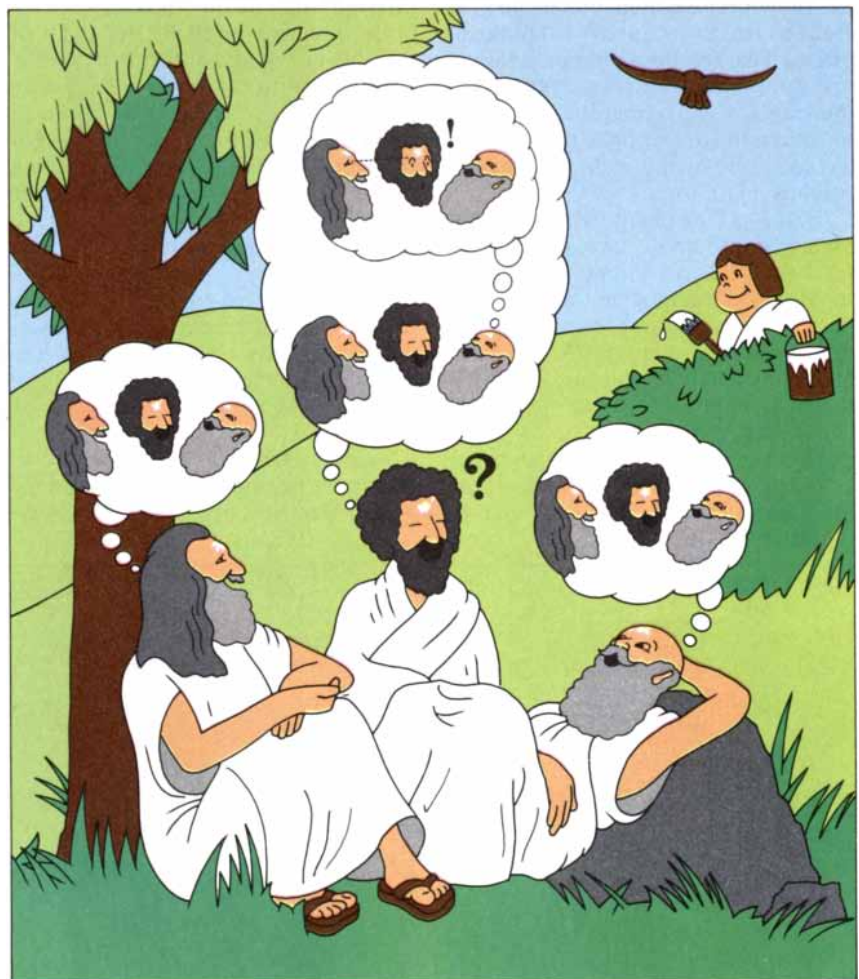
stroke of midnight of the day on which you ascertain his infidelity." In the realm of the Amazons information was shared freely but not too freely; each wife knew about the infidelities of all husbands except her own. Also, word of an execution spread throughout the realm within a day. As it happened, there were exactly 40 unfaithful husbands. Were any executed and, if so, when?

Readers will have noted that the queen announced that at least one husband had been unfaithful. If exactly one husband had been cheating, his wife would have known immediately: if any husband but her own was a philanderer, she would have heard about it. Hence on midnight of the day of the queen's announcement he would have been killed by his wife.

If there had been exactly two unfaithful husbands, their respective wives would have dispatched them at midnight of the second day. The news that there was no execution carried out at midnight of the first day would

confirm the fact that there were two husbands who were unfaithful. Because the wives of the two philanderers would have been aware of only one philanderer in the realm (although all the other Amazons would have realized that there were at least two), they would immediately have realized that their own spouse must be the second philanderer.

By now readers will have caught the drift of the argument. No execution at midnight of the n th day means that at least $n+1$ husbands were unfaithful. At the dawn of the 40th day it would be common knowledge that at least 40 husbands were unfaithful. Indeed, this would come as no surprise to Amazons married to faithful husbands, since they would know of 40 philandering husbands. Only a wife with an unfaithful husband would know of 39 unfaithful ones, meaning that her own spouse was the 40th. These wives, in carrying out their monarch's command, would then summon their husbands for a midnight tête-à-tête on the 40th



One philosopher thinks on a deeper level than his colleagues

day after the queen's proclamation.

Is the Amazon puzzle really a variation of the three-philosophers puzzle? The question can be answered by asking what would happen if there were only three Amazons, each married to an unfaithful husband. In that case at the end of the third day each Amazon would have drawn the correct conclusion. In this form the puzzle compares most directly with that of the three viziers (which I have already shown to be a transformation of the three-philosophers puzzle).

To see this, suppose the sultan had told the three candidates for grand vizier, "I shall ask several times in succession whether you know the color of your own hat. Answer only if you know, otherwise remain silent." In that case the first time the sultan asked the question, all three viziers would have remained silent. The second time he asked the question, all three would again have remained silent. The third time he asked, all three candidates would have replied yes.

The Amazon puzzle deals with 40 unfaithful husbands, not three. Can the three-philosophers puzzle be generalized to a group of 40 philosophers? Yes. For the moment, imagine just four philosophers asleep under the tree. On awakening they all begin to laugh and the fourth philosopher (actually one of the gods in disguise) reasons as follows:

"H'm. It is consonant with my divine dignity to assume I have an unsoiled forehead and can therefore laugh at the three besmirched mortals. But why does one of them not realize his condition and stop laughing?" (The Olympian now mentally recapitulates Pythagoras' argument.) "Dear me, I think I know the reason."

If such a thought process can bring a fourth philosopher to realize that his forehead has not been spared,

then it can just as easily explain how a fifth, a sixth and even a 40th philosopher might arrive at the same conclusion. In one of his "Mathematical Games" columns Martin Gardner developed a generalization of the three-viziers puzzle along these lines [see SCIENTIFIC AMERICAN, May, 1977]. As he astutely observed, however, difficulties arise. "This generalization usually provokes arguments, because the problem demands so many fuzzy assumptions about degrees of smartness and lengthening lapses of time that the problem becomes unreal."

To be sure, all puzzles have a degree of unreality about them. How likely is it, even in a trendy restaurant, that three people lurching on spinach quiche and sipping cool Chardonnay will all end up with a bit of spinach on their teeth at the same time? And even if that does happen, how likely is it that one of them would not see from another's glance that he himself was being laughed at?

Still, to investigators trying to develop new forms of artificial intelligence based on what is known as logic programming, people puzzles are serious business. John L. McCarthy of the University of California at Berkeley, among others, tested the deductive powers of logic systems by applying them to people puzzles such as that of the three viziers. Logic-programming systems exploit a kind of symbolic thought process known as predicate calculus to derive automatically various deductions from certain assumptions. In order to mimic the human capacity by which people make deductions in various social situations, logic-programming systems must have the power to model reasoning about reasoning.

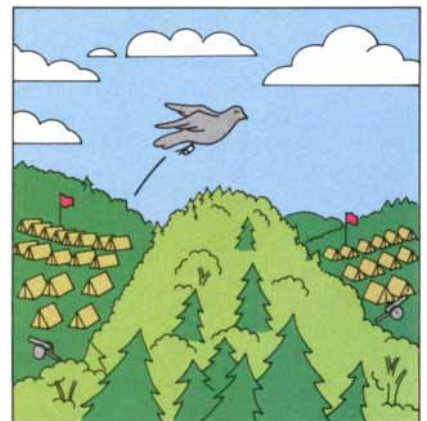
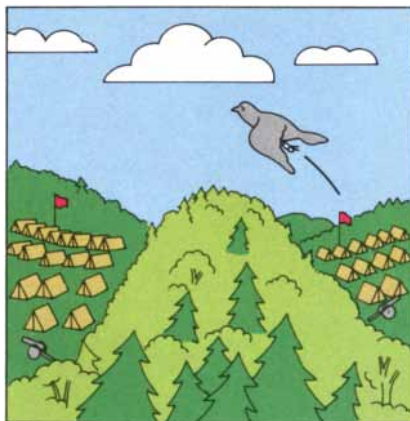
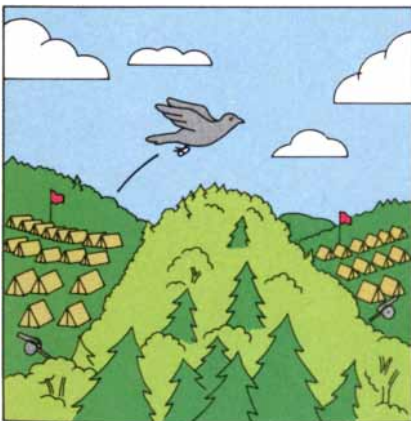
Such investigators might do well to examine the more realistic kind of people puzzles that occur in every-

day life. Although these puzzles are not logically deep and require a great many assumptions for their solution, they nonetheless reveal that a significant part of our own mental capacity is applied in thinking about what other people may be thinking we are thinking.

Erving Goffman, also at Berkeley, is a pioneer in a field of psychology called transactional analysis, which maintains that every person is constantly trying to know what others are thinking about him or her and to manipulate that thinking, rather in the manner of actors. Goffman has documented this aspect of our inner lives through a compilation of all-too-human behaviors that we recognize as valid experiences.

Here are two examples of the kind described by Goffman that actually come from my own life; it so happens that I have been collecting similar examples for years. They all seem to point to a human facility for thinking about what other people are thinking, which in most of us, I believe, is semiconscious.

I once saw a policeman descending to street level from a second-floor establishment of a questionable nature. Perhaps it was a massage parlor. (The section of town in which the incident took place makes that likely.) As he emerged onto the street, the policeman unexpectedly met a fellow officer. For his colleague he put on a very strange face: he looked not embarrassed but as though he were pretending to be embarrassed. This was at first strange to me but later understandable. The first officer's primary feeling was probably outright embarrassment. After all, the colleague would rightly suspect that the visit had been unofficial, so to speak. Yet to have expressed such embarrassment would only have confirmed those sus-



Will the confirmations of confirmations between the generals ever end?

pictions. On the other hand, if the second officer could be made to perceive that the first officer was only pretending to be embarrassed, he might pass the incident off as a joke.

Games sometimes bring out the worst in people. Two rather competitive colleagues had just finished a game of poker. The loser was visibly upset but trying hard to control himself. Finally he blurted out: "You were planning to _____ me and I was planning to _____ you. You just beat me to it!" (Here the blanks can be filled in with a suitable coarse word meaning, in this context, to cheat.)

The operative word here is *just*. At first it seems to imply that luck played the major role in his opponent's victory. On closer inspection, however, it appears to mean the loser wants the winner to think he (the loser) thinks the most significant aspect of their game was the equality of their motivation rather than the inequality implied by his losing.

I wonder if any readers have examples of their own to send in. Particularly interesting would be examples that are a little closer in actual content to traditional people puzzles.

Speaking of puzzles, a wonderful puzzle book, *The Puzzling Adventures of Dr. Ecco*, has appeared. It is written by Dennis Shasha, a computer scientist at the Courant Institute of Mathematical Sciences at New York University. The book describes the adventures of the mysterious Dr. Ecco, a fabulously intelligent eccentric who solves puzzles for a living. Besides a couple of people puzzles, it includes puzzles about elections, multiple routes, spies, circuits that check circuits and much more. There is even a contest in which entrants are challenged to solve 10 decoding puzzles. Successful solvers win not only an "Omniheurist Tee-Shirt" (symbolic of supreme intellectual ability) but also the right to have their name entered in a draw for a hand-carved onyx chess set. The contest closes on April 1, 1989. The winner of the grand prize will be announced in this column in July.

I shall close the discussion of people puzzles with one from Shasha's book. Two generals whose respective armies occupy two sides of a ridge want to coordinate their attack on a common enemy, because if either army attacks alone, it could be destroyed. Unfortunately the generals can communicate only by carrier pigeons that fly over the ridge from one camp to the other [see illustration on opposite page].

The first general sends the second general the following message: "At-

tack at 0800 hours. Confirm message received, otherwise I shall not attack." The second general has no objection to attacking at that time and intends to send a pigeon bearing the confirmation. But the second general suddenly realizes that the first general will not attack unless he receives the confirmation. Since the second general has no guarantee that a carrier pigeon will actually deliver his confirmation to the first general, he decides that he himself will not attack unless he knows that the first general knows that the attack has been confirmed. The second general then sends a pigeon off with a message to the first general. Will the seemingly infinite regress of confirmation messages ever stop? Perhaps the answer depends on the particular message that one of the generals sends at some point.

I acknowledge with thanks the help of two of my colleagues at the University of Western Ontario, Andy L. Szilard and Areski Nait Abdallah, in the preparation of this column.

I shall now announce the winners of the prime-grid challenge issued by Gordon Lee in the July column. A prime grid is a six-by-six grid of squares each of which contains a single digit. The object is to choose the digits and their positions in the grid in such a way that one gets as many prime numbers as possible by reading consecutive digits along any straight line—horizontal, vertical or diagonal. Lee, who organized a similar contest in England last year, reported a winning entry of 170 primes. Could the readers of this column do better? I never doubted it!

I received many more entries than I thought I would. They came both from high-tech hackers who employed supercomputers and from humbler folk who worked them out with pencil and paper. I mailed the entries to Lee for adjudication. Here are the results in reverse order:

Grittiest griddier: Larry J. Padden of Oklahoma City, Okla., who produced 147 different prime grids containing between 170 and 173 primes.

Sixiest gridders: David Mckenzie and Frank Endres of Austin, Tex., who discovered a grid that contained only 106 primes, but all 28 possible six-digit numbers in it were prime.

Third-place griddier: James I. Waldby of Robinson, Ill., whose score was 185.

Second-place griddier: Mckenzie and Endres share this honor with Stephen C. Root of Westboro, Mass. They were able to come up with grids that contained 186 primes.

Top griddier: Root again. His winning entry, which is given below, has a total of 188 primes:

3	1	7	3	3	3
9	9	5	6	3	9
1	1	8	1	4	2
1	3	6	3	7	3
3	4	9	1	9	9
3	7	9	3	7	9

The shortest path through the cruel three-dimensional labyrinth described in the September column was identified by many readers. It was 39 cells long (counting both the entrance and the exit cells). There are many other, longer paths through the labyrinth, but all paths to the exit lead past a specific point in its murky interior. It is at that spot that the ferocious Minotaur awaits the hapless Athenian youths and maidens who have been forced into the maze. Although some readers discovered the Minotaur's secret location by laboriously drawing a map and then spotting the bottleneck, others suspected that the high degree of symmetry in the design of the maze might be the clue to determining the Minotaur's spot. It was.

At just one point the three-dimensional maze is asymmetric. Readers might enjoy returning to the maze to discover a "missing" wall between the cells (5,3) and (5,4) on the second level. (The coordinates respectively designate the number of cells to the right of the left edge of the maze and the number of cells down from the top edge.) Why does the bottleneck have to be at the point of asymmetry?

The six readers whose correct Minotaur solutions reached me first are Michael Amling, Glen Ellyn, Ill.; Lawrence Leinweber, Cleveland Heights, Ohio; Thomas R. Lunsford, Jr., Hinesville, Ga.; Donald E. G. Maln, Rochester, Mich.; Jim Newton, Middleton, Wis., and Ken Silber, New York, N.Y.

FURTHER READING

THE PRESENTATION OF SELF IN EVERYDAY LIFE. Erving Goffman. Doubleday & Company, Inc., 1959.

THE PUZZLING ADVENTURES OF DR. ECCO. Dennis Shasha. W. H. Freeman and Company, 1988.

BOOKS

A cold-eyed look at a treatise on warm-blooded dinosaurs



by A. W. Crompton and Stephen M. Gatesy

PREDATORY DINOSAURS OF THE WORLD, by Gregory Paul. Simon and Schuster (\$19.95).

The past two decades have seen the publication of numerous books on the biology, diversity and extinction of dinosaurs. Prominent among them have been volumes advancing a new view of dinosaurs: lumbering, cold-blooded and virtually brainless giant lizards are replaced by fleet, endothermic creatures rather akin to modern birds. Gregory Paul's book stresses this view as it deals exhaustively with the structure and biology of the predatory dinosaurs.

The new view—and to a large extent the general revival of interest in dinosaurs—can be traced to a short 1968 article, "The Superiority of Dinosaurs," by Robert T. Bakker, who was then a Yale undergraduate. Bakker's conclusion was that dinosaurs were "fast, agile, energetic creatures that lived at a high physiological level reached elsewhere among land vertebrates only by the later, advanced mammals." Bakker proposed that 10-ton horned dinosaurs could have galloped at 30 miles per hour and that giant herbivores such as *Brontosaurus* could hold their 30-foot tails off the ground and move as effectively on dry land as modern elephants. Bakker suggested too that some dinosaurs were social and moved in organized herds.

Much of Bakker's later work was devoted to convincing his colleagues that dinosaurs, like birds and mammals, were "warm-blooded" [see "Dinosaur Renaissance," by Robert T. Bakker; *SCIENTIFIC AMERICAN*, April,

1975]. So-called warm-blooded animals consume energy at rest at a rate from five to 10 times greater than a "cold-blooded" reptile, amphibian or fish of comparable size. Birds and mammals control the loss or conservation of this heat energy and maintain a constant body temperature, in sharp contrast to modern reptiles, which depend on environmental sources of heat to elevate their body temperature to a preferred level. The concept that dinosaurs may have been warm-blooded was not new, but the importance of Bakker's work is that he based his views of dinosaurian energies on several novel and quite different lines of research. They ranged from simple counts of carnivore and herbivore bones in museum collections to sophisticated measurements of aspects of mammalian, avian and reptilian physiology. Paul does not build on that solid foundation. He accepts, as a given, that dinosaurs were warm-blooded and fails to present a balanced account for and against this view or for his own very controversial theories.

Perhaps the most compelling argument in favor of a high basal metabolic rate for dinosaurs derives from the empirical relations between animal size, resting metabolic rate and maximum sustained (not short-burst) speed—relations determined in the laboratories of Knut Schmidt-Nielsen at Duke and C. Richard Taylor at Harvard and, more recently, by Albert F. Bennett at the University of California at Irvine. These investigators (including Bakker when he worked in Taylor's laboratory) demonstrated that if the size and resting metabolic rate of a living animal is known, its top running speed can be calculated to within a factor of two or three.

Bakker applied these empirical formulas to dinosaurs. An assumption of a reptilian metabolic rate would mean that the maximum speed of a

10-ton *Tyrannosaurus rex* would be about six kilometers per hour and that of a 100-kilogram ostrichlike dinosaur about three kilometers an hour. On the other hand, if their resting metabolic rate were the same as that of mammals, these animals could be calculated to have sustained running speeds in the range of 50 kilometers per hour. Intuitively, these high sustained speeds are in line with dinosaur limb structure and proportions, which are more comparable to those of birds and mammals than to those of living reptiles. Other evidence for warm-bloodedness, such as bone histology, predator-to-prey ratios and posture, is still being debated.

Not all investigators accept Bakker's expanded views on dinosaur biology or his unflinching attempt to view these animals in terms of mammalian or avian physiology. Indeed, some of his ideas are met with considerable skepticism. For example, he speculates that the large, long-necked, herbivorous sauropods may have given birth to live young. He also asserts that tyrannosaurs, such as his newly coined *Nanotyrannus*, are more bird-like than *Archaeopteryx* (generally considered to be the earliest known bird).

Still, most people in the field would agree that Bakker's ideas have generated a valuable debate and have led to novel and imaginative research on a group of organisms that dominated the terrestrial scene for well over 130 million years. More important, Bakker's work has highlighted the fact that a multidisciplinary approach can throw light on the biology of extinct organisms. Unfortunately the random application of the biology of living animals to extinct vertebrates opens the door to idle speculation and fanciful scenarios comparable to Rudyard Kipling's *Just So Stories*.

Paul's *Predatory Dinosaurs of the World* is a comprehensive treatise on the bipedal, carnivorous dinosaurs called theropods and their relatives. The underlying weakness of this book is that Paul not only has accepted Bakker's view of dinosaurian biology uncritically but also has added many ideas that lack any sound biological foundation. Paul is hard on earlier books about dinosaurs; as a self-proclaimed "progressive dinosaurologist," he states that many of these works are "marred by sloppy thinking" and show a "misuse or misunderstanding of modern biology." Sad to say, this is particularly true of much of his own book.

A. W. CROMPTON and STEPHEN M. GATESY are at the Museum of Comparative Zoology at Harvard University. Crompton is Fisher Professor of Natural History and Gatesy is one of his graduate students.

In order to bolster the view that our understanding of dinosaurian biology has recently undergone a major revolution, Paul erects and demolishes a series of straw men. For example, he repeatedly stresses that until two decades ago dinosaurs were viewed as "slow, sluggish, dull-witted" creatures. Some scientists may have held this view of the large herbivorous dinosaurs, but it was seldom considered to be true of the very group to which the book is devoted, namely the predatory dinosaurs. One has only to look at the magnificent reconstructions prepared by Charles R. Knight in the late 1800's for the American Museum of Natural History—the small theropod *Ornitholestes* leaping to snatch a meal, say, or a dueling pair of *Dryptosaurus*—to dispel the view. Similarly, Gerhard Heilmann's 1927 renditions of such bipedal dinosaurs as a sprinting *Compsognathus* or an alert, ostrichlike *Struthiomimus* counter the alleged orthodox image of sluggish predatory dinosaurs.

Following in Bakker's footsteps, Paul has adopted the practice of claiming that dinosaurs have a mammalian-avian physiology. There is simply no such thing. What he is really referring to is only one aspect of their physiology: the fact that both mammals and birds are capable of maintaining a constant body temperature and have a high metabolic rate. He constantly, uncritically and indiscriminately attempts to force dinosaurs into either a mammalian or an avian mold, depending on the scenario he wants to draw. Mammals serve extensively as analogues for dinosaur behavior, whereas almost all other characteristics are patterned after birds.

By assuming that all dinosaurs have an avian physiology, Paul blurs the transition between early dinosaurs and their ancestors at one end and dinosaurs and true birds at the other. Consequently one is left with the impression that the majority of predatory dinosaurs and their ancestors had feathers and a resting metabolic rate comparable to that of modern birds, and that anatomically and physiologically there is little difference between the dinosaur group from which birds arose and true birds.

Between sweeping generalizations Paul makes a number of novel suggestions. Within the dinosaurs, it is argued, flight evolved and was lost more than once. This is contrary to the established notion that flapping flight evolved only three times in vertebrates: with birds, bats and ptero-

saur. Paul states that flight possibly occurred first among the earliest dinosaurs. He bases the claim on a recently discovered form called *Protoavis* from late Triassic rocks of Texas. Until the description of the fossil is complete, it is difficult to judge that contention, but the characters Paul uses to relate this form to the contemporary heavy, bipedal dinosaurs are not convincing.

The famous *Archaeopteryx* from the lithographic limestone of Germany is considered only a second experiment with flight. *Archaeopteryx* is usually thought to have been the first bird, and Paul agrees with John H. Ostrom's well-accepted contention that it is also very similar to theropod dinosaurs; Paul accepts a close relation between *Archaeopteryx* and small, sickle-clawed theropods such as *Deinonychus*. In some features, according to Paul, these dinosaurs and some others (termed protobirds) are more "birdy" than *Archaeopteryx*, which compels him to infer that they are secondarily flightless.

In fact, he thinks the flying protobirds "may have spawned a series of ever more advanced protobird theropods that separately lost the ability to fly." Similarly, the so-called ostrich-mimics and several other theropods are speculatively considered descendants of flying ancestors. That creates problems, since these forms have features "less birdlike" than their presumed ancestors. Paul invokes a series of reversals to account for the inconsistency, with tails being reenlarged, teeth redeveloping serrations and hip elements swinging back to a primitive position. Current evolutionary thinking finds such a scenario untenable and would support instead the simpler view that such features have merely been retained from flightless ancestors.

It is not clear whether Paul thinks true birds came from *Archaeopteryx*-like forms, from his secondarily flightless dromaeosaurs (a group of agile dinosaurs that includes the sickle-clawed theropods) or even from an unknown stock that developed flight independently. In any case, it is clear that he considers flight a relatively minor achievement. He states that many Jurassic theropods were "ready for avian-style flight"—that "all they had to do was elongate their forelimbs, modify them a little, and increase their power, and they could have flown." Most paleontologists and evolutionary biologists would consider this fanatically overstated at best. The prerequisites for the demands of

aerial locomotion are considered to be among the most complex of locomotor adaptations, and Paul's trivialization of the acquisition of flight is quite shocking.

When it does not suit his view, Paul contends that negative evidence has no meaning, but when such evidence is consistent with his opinions, it is adduced to substantiate critical ideas. For example, in order to bolster his claim that the late Triassic *Protoavis* was not ancestral to true birds, he cites the absence of feathers in strata laid down after *Protoavis*, where feathers should abound if an adaptive radiation of birds had taken place. On the other hand, he argues that the lack of evidence for feathers on predatory dinosaurs (feathers are known only from *Archaeopteryx* and true birds) or in pre-Cretaceous strata does not mean that small and medium-size dinosaurs were unfeathered; the feathers simply were not preserved. Lack of the same type of fossil evidence is cited in opposite ways to support preconceived ideas.

It quickly becomes clear that Paul's work suffers from the absence of a phylogenetic framework. In reviewing systematic methods he describes cladistics, which is generally accepted as being the best available methodology for assessing relationship. Cladistics is an approach to classification that relies on the identification of unique features to determine monophyletic groups (groups that have a common ancestor) and reconstruct genealogies. Yet later he eschews monophyletic groups—a central tenet of cladistics—and opts for a "commonsense" approach.

Paul thinks a system should contain as much information as possible and yet be "simple enough to grasp in a single look." He dismisses the cladistic approach of Jacques Gauthier as "really as arbitrary as any other" because only major branching points are named, and chooses instead to construct a grade-based taxonomy. (A grade-based taxonomy is a classification that reflects general level of organization rather than genealogy. For example, as currently defined, the class Reptilia is a grade, because it includes all scaly-skinned vertebrates that lay hard-shelled eggs. In a cladistic classification, though, the class Reptilia—which excludes birds—does not exist. The reason is that birds share a more recent common ancestor with crocodilians than crocodilians do with lizards, turtles or any other "rep-

tile"; cladists therefore place birds and crocodiles in the same monophyletic group: the Archosauria.)

Rather than being simple, Paul's "modern assessment" is almost uninterpretable, lacking, as it does, any reflection of phylogeny. Predatory dinosaurs are "those that lack herbivorous teeth, or ancestors who had them." This systematist's nightmare is made up of all dinosaurs except prosauropods, brontosaurs and ornithischians. The archosaurs (dinosaurs, birds, crocodylians, pterosaurs and some others) are grouped in a major taxonomic unit, a class. (Other examples of classes are amphibians, reptiles and mammals.) The archosaur class is defined by the presence in all its members of a large "preorbital opening" on both sides of the snout, between the openings for the eyes and the mouth, and by "bird-type bone histology" (not defined) and "more erect gaits" in most members. These last two characters are meant to suggest the presence of rapid growth rates and increased metabolisms. Class rank appears to satisfy Paul because "dinosaurs and their relatives are no longer either 'lower' vertebrates or reptiles, they are archosaurs, equal to mammals in status."

The class Dinosauria is redefined to include forms such as *Lagosuchus* (which, with the ornithosuchians, form a perplexing group called protodinosaurians) and is then divided into four new superorders, the interrelations of which remain unclear. In addition Paul erects four new orders, five suborders, three families, four subfamilies and two new species.

With no clear hypothesis of relationship for many groups, he freely describes evolutionary trends for animals that are not grouped by hierarchically arranged characters, leaving almost any conclusion or scenario open. In the light of the recent resurgence of systematics and recognition of its value to evolutionary studies, this is most disappointing. We find groups being described variously as "birdy," "birdlike," "increasingly birdlike," "more birdlike" and "even more birdlike," but we never find characters organized in a way that can be evaluated critically.

The relentless reference to avian attributes in dinosaurs serves Paul's goal of supporting a theropod ancestry for birds, as well as giving us a model to help visualize extinct animals in our mind. He is correct in pointing out that many features of birds first appeared in theropods, but

if that is the case, only those changes distinguishing birds from theropods are truly avian features. One does not refer to evolutionary novelties that distinguished the first mammals as being "humanlike" or "batlike" or "whalelike"; they are simply "mammalian." For lack of a hierarchy in his systematics, Paul loses track of the levels at which unique or derived characters appear; he uses a terminal taxon (birds) to tell us which characters are derived even if they are not retained in birds. To state, for example, that the remarkably primitive dinosaur *Herrerasaurus* had "birdlike lower hips" is meaningless.

A chapter on fossil tracks (ichnites) is a refreshingly short and interesting description of the value of footprints in dinosaur paleontology. Trackways are solid evidence of actual dinosaur behavior, and Paul reviews the benefits and limitations of their study. Although tracks can be difficult to categorize and compare because of the differences in substrate, preservation and foot movement, they are often the best indicator of the presence of dinosaurs in areas where fossil bones are poorly represented. A restudy of the alleged "man tracks" alongside dinosaur tracks in Texas has led to their recognition as prints made by small theropods walking flat-footed rather than up on their toes. Trackways have been analyzed to estimate the speed of the track maker; multiple trails give evidence of group behavior. Ichnology (the study of fossil tracks) is experiencing a revival, and new insights will surely follow.

The locomotor system of the animals producing the tracks is discussed in other chapters: "The Nuts and Bolts of Predatory Dinosaur Anatomy and Action" and "Predatory Dinosaur Speed." Paul's analysis of theropod locomotor anatomy and function is somewhat oversimplified and quite controversial. His major premise is that theropod hind limbs are structurally and functionally like those of large, flightless birds such as ostriches.

Many early reconstructions of dinosaurs were similar to sprawling reptiles, with limbs splayed out from the body. It is now evident, however, that dinosaurs (like birds) had a fully erect posture, with the hind-limb bones held close to the body and operating essentially back and forth in the direction of travel. Paul's recognition that many of the features we see in modern bird limbs, particularly the foot,

evolved in theropods and were passed along to their avian descendants is sound. He goes further, however, to insist that the hind limbs of theropods of all sizes (even the huge tyrannosaurs) were positioned and used like those of birds.

Confusion arises when one tries to characterize the movement of even modern bird limbs, much less those of extinct theropods. When standing, birds keep the upper leg bone (femur) relatively horizontal. This horizontal orientation is considered to be a way to move the feet forward under the center of mass of the animal, which has shifted forward owing to massive reduction of the tail. The short femur and knee are often hidden under the wings and body feathers, leading to the impression that the "knee" (which is actually the ankle) is bending backward. X-ray and light films made and analyzed in our laboratory at Harvard show that birds depend only slightly on movement of the hip joint at low and moderate speeds; knee flexion produces most foot movement. At higher speeds the femur is moved through a progressively larger arc, although the knee is still flexed much more. Bird locomotion is therefore characterized by a speed-dependent recruitment of the hip joint: the faster the bird moves, the more it moves its leg at the hip.

Paul asserts correctly that, contrary to the traditional view, birds do move the femur, and he describes a film sequence showing an ostrich at high speed, with its femur moving through a substantial arc. Regrettably, he then relies on an ostrich running at full speed as his definition of bird locomotion. And he writes that "long-tailed theropods did not have the same balance problem [as birds], and the femora are so long that they must have been in full use at all speeds." Why then does he say that theropods used their limbs like birds? In spite of his admission that there are distinct differences between theropods and birds in the tail, hips and limbs, he is adamant about the "birdlike" nature of theropod locomotion.

The evolution of the avian lung is also addressed by Paul in a frustratingly simplistic way. Most reptiles have saclike lungs that are filled and emptied by expansion and contraction of the thoracic cavity, resulting in a bidirectional flow of air with each breath. In birds, however, there is an extensive system of interconnecting nonrespiratory air sacs throughout the body, with a pair of small lungs

high up near the vertebral column. The lung is relatively stiff; air is not pumped by its alternate expansion and contraction but rather is circulated through it unidirectionally by the air sacs. The bird lung is designed to meet the high aerobic cost of sustained flight. If it could be shown convincingly that Paul's advanced theropods (termed avetheropods) had such a lung, that would certainly suggest very high activity levels in these forms. He maintains that "a look at complete avetheropod rib cages shows they were avian in design." What is the evidence for this statement?

Paul states that typical animals ventilate the lung with the front ribs, but because birds have short front ribs, they instead use ribs in the belly to force air between the air sacs and ventilate the lung. He then goes on to claim that he can see the avian system evolving within theropods. One entire group of theropods is said to possess front ribs that were "too short to have ventilated large normal lungs" and long belly ribs that "can be explained only if they were ventilating large abdominal air sacs that in turn fed unidirectional air flow lungs." Paul's ability to detect from the structure of fossil ribs the evolution of a system that is not even fully understood in living birds must be met with a great deal of skepticism.

Paul uncritically assumes that what is thought to be true for other dinosaurs and birds is also true for predatory dinosaurs. John R. Horner has carefully documented evidence from Cretaceous nest sites indicating that herbivorous dinosaurs formed breeding colonies and practiced parental care. Without justification, Paul assumes that this holds for predatory dinosaurs as well.

In Paul's reconstructions the lifestyles of dinosaurs come to be interchangeable with those of herding mammals of the African savanna. Since dinosaurs are considered to be so much like mammals, they are said to have had secondary palates so that they could breathe and chew at the same time (this is true of only some mammals), even though birds have lost this unique feature. Why dinosaurs would need a secondary palate is a mystery, since "predators gulp their food" and theropods had a tongue that was "stiff, like a lizard's" (lizards have tongues that are mobile and muscular). Paul tells us dinosaurs and mammals arose at the same time and developed an erect posture at the same rate, metabolic rates "flow over"

the cross-sectional areas of an animal's body and "muscles produce power that elastic bones store and return..." Statements such as these are more than misleading, they are simply incorrect.

The disappearance of the dinosaurs at the end of the Cretaceous receives scant treatment in this book. Amidst the flurry of publications dealing with dinosaur extinction that is something of a relief; by concentrating on the diversity and lifestyles of predatory dinosaurs Paul tries to stress the success of these animals rather than their eventual demise. Since he incessantly downplays differences between dinosaurs and birds, however, he is left with a dilemma: how to explain why his fully warm-blooded theropods vanish, whereas birds do not. Contrary to the popular slant toward extraterrestrial impacts, Paul follows Bakker's more traditional view of dinosaur extinction. He addresses the possibility of a decline in dinosaur diversity well before an impact, the implications of new land connections due to falling sea levels and the possibility that an impact was the last straw for an already waning dinosaur population. In the end Paul is uncharacteristically straightforward in admitting that he has no good explanation for the extinction of theropods and the survival of birds.

It is not clear for whom this book is intended. The first half deals with the life and evolution of predatory dinosaurs and the last half is a descriptive catalogue of each species. The level of terminology varies considerably, with many scientific terms being softened ("throat trachea" and "hip's pubis") in an effort to make them understandable to most readers. Other attempts to contrive descriptive terms are misleading, with mammalian limbs said to be "supple" whereas bird limbs are "stiff."

The second half of the book could be a useful catalogue, but it follows Paul's idiosyncratic system of classification and is therefore of little use to professional paleontologists. On the other hand, this last half assumes a familiarity with the anatomy of dinosaurs and the geological horizons in which they are found that is surely beyond the scope of the general reader. Sprinkled throughout are personal tidbits such as "This is my favorite velociraptor, and that is saying a lot" or "Dinosaurs look neat," along with sections of cheap adventure drama in which he describes imaginary

lethal combat between herds of multi-ton giants.

Paul is a superior artist, and he has taken great pains to reconstruct as accurately as possible the skeletons of all the known predatory dinosaurs, their ancestors and several descendants. Although the number of illustrations is excessive, this is the most useful aspect of the book and will be helpful to any artist trying to reconstruct predatory dinosaurs. Paul's illustrations are lively, and he relies on them to convey many of his ideas about dinosaur biology. His reconstructions of theropods give no indication of what is real and what is imaginary, however, and yet they will no doubt be reproduced many times and will continue to influence the public image of dinosaurs.

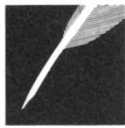
Predatory Dinosaurs of the World, like Bakker's *Dinosaur Heresies*, is a highly personal account of one author's view of dinosaurs. Controversies are glossed over, complexities are avoided and overgeneralizations abound. The reader should not expect an up-to-date synopsis of the current state of dinosaur studies and should be cautioned that many of Paul's ideas are extreme and that most of them have been disputed and will continue to be debated for years. This is not to say that there is no place for such individual views, but they need to be balanced in some way so that readers who are not familiar with the field will recognize that not all paleontologists would agree.

Paul is correct in stating that most dinosaur books written by nonscientists contain information and illustrations reworked from earlier publications. It should be possible for dinosaur specialists to create books that are authoritative and yet not overly one-sided, and in which old ideas are not prematurely branded as outdated by the erection of an artificial progressive vs. traditional dichotomy. *The Illustrated Encyclopedia of Dinosaurs*, by the paleontologist David Norman, is the best book we have found to fit this description.

Regrettably, in *Predatory Dinosaurs of the World* Paul adds little new information or insight and fails to exercise restraint when interpreting limited evidence. He tells us that by the 1930's "dinosaurs had become so popular with the public that the subject had taken on something of a circus air..." In this book Paul fails to give the curious reader a view of how the science of paleontology works; instead, anything goes.

ESSAY

Needed: a free flow of information and ideas



by John Shattuck
and Muriel M. Spence

During the past decade the Federal Government has established a network of policies that restrict the availability, shape the content and limit the communication of information. This network includes an expanded classification system, restraints on the exchange of unclassified information, the use of export controls to restrict technical data, and limits on contacts between citizens of the U.S. and those of other countries. Federal information policy has also curtailed the role of the Government in both collecting and publishing scientific, technical and statistical information. When the Bush Administration takes office in January, it would do well to dismantle this bureaucratic apparatus. Otherwise the web of rules will continue to fetter intellectual freedom, weaken democratic decision making, cripple our economic competitiveness and—contrary to intent—erode national security. A few instances will show how the evolving new policies have impeded the free flow of information:

In 1982 President Reagan promulgated an Executive Order giving Federal officials unprecedented new authority to classify information, including information already in the public domain. The order erased a requirement that information must be declassified within a prescribed period.

In 1985 the Department of Defense ordered that the Society of Photo-Optical Instrumentation Engineers allow only U.S. citizens, Canadian citizens and permanent residents of the U.S. to attend a conference at which unclassified papers would be presented.

In 1986 the Administration initiated efforts to abolish or scale down the National Technical Information Service, a Government-sponsored clearinghouse for a wide range of publicly available scientific and technical data.

In 1987 the Federal Bureau of Investigation acknowledged the existence

of a "Library Awareness Program" to involve librarians in an effort to report to the bureau researchers and other library users who might be "hostile intelligence people."

In recent years the Federal Government has depended on two rationales to justify increased control of the flow of information. The first is the Government's duty to protect national security. The second is the Government's obligation to curtail deficit spending and to avoid excessive regulation.

Neither justification meets the objection that such policies can do substantial harm. The Soviet Union's experience illustrates the danger. During the recent series of breakthroughs in superconductivity, for example, the Soviets played no part. No one sought to exclude them, but Soviet scientists were burdened with travel restrictions and restraints on contacts with foreigners. In July, 1987, the Administration showed signs of emulating some of the restrictions by holding a White House conference on superconductivity that excluded some foreign scientists. Whether the rule ultimately deprived those individuals of access to information presented at the meeting is debatable.

There is no doubt, however, that restraints on the flow of scientific and technical information have already begun to hurt the U.S. economy. A 1987 National Academy of Sciences report indicates that every year controls on the export of information and manufactured goods cost the U.S. economy about 188,000 jobs and \$9 billion. The estimates are based on information supplied by exporters who compete in Western Europe and Japan.

Limits on the participation of foreign citizens in the U.S. economy have another harmful effect: they deprive the nation of needed foreign expertise in fields, such as engineering, where there is a chronic shortage of U.S. specialists. A further victim of controls is, ironically, likely to be U.S. national security. Our security depends on a strong technological and scientific base, which cannot be maintained if communication is impeded.

How can these trends be reversed? When President Bush takes office in January, the reform of Federal information policy should be a major element of his programs for science, the economy and national security. Within its first 100 days the new administration should signal a shift in policy by issuing an executive order revising the classification system. The Bush Administration should also make changes in the export-control system a mat-

ter of policy priority and place limits on the role of the Office of Management and Budget in controlling the information compiled and disseminated by Federal agencies. At the same time the president should deliver a message to Congress articulating his new approach to Federal information policy.

An executive order on information policy should be promulgated on the basis of two principles. One principle should presume as being normal the free and open communication of information generated both inside and outside the Federal Government. A second principle should hold that no information may be restricted solely on the ground that its speculative relation to other information might make it harmful if it were communicated.

The Executive Order should reverse the current presumption in favor of classification in all cases where officials are in doubt about whether secrecy is necessary. Furthermore, the order should raise the threshold standard for classification, require automatic declassification within a prescribed length of time, eliminate the authority of officials to reclassify information already in the public domain and lift restrictions on the communication of unclassified information between Americans and citizens of other countries.

Federal agencies should be precluded from exercising control by means of prepublication review over the content and conclusions of Federally sponsored research. Limits should be put on the power of the Office of Management and Budget to limit the collection and dissemination of data by Federal agencies.

The free flow of information and ideas is vital to the fabric of our society. The engines of innovation that drive our economy and guarantee our security are powered by open and unfettered communication. Government policy aimed at broadly controlling the communication of information and ideas may soon become irreparably damaging unless it is substantially reversed. Changes in Federal information policy should be essential elements on the Bush Administration's agenda and cannot be undertaken too soon.

JOHN SHATTUCK and MURIEL M. SPENCE are respectively vice-president of Harvard University for government, community and public affairs and director of policy analysis.

WHY
EAGLE
DARES.



WE HAVE THE LEADERSHIP, THE KNOW-HOW, AND THE CARS TO TAKE ON THE WORLD.

The last time a major domestic brand of automobile was introduced, Ike was in the White House. People were driving Packards, Studebakers. And, yes, the brand-new Edsel.

So why has the management of Chrysler dared to fly where others have feared to go?

Why now? Less than a year after the Bears roared and the Bulls ran for cover.

Why now? When Americans have almost 600 different car models to choose from. Produced by almost every country in the world capable of putting up an assembly line.

Because Chrysler's management is doing more than building cars. We're building cars that inspire *driver confidence*. We're so proud of the cars we're building, we're selling them alongside American classics—the legendary Jeep vehicles.

THERE IS NO SHORTCUT TO EXCELLENCE.

Every Eagle is built carefully, meticulously, from the inside out. Eagle engineering and design draw from a large pool of the world's most forward automotive thinking.

The result is a new line of sophisticated, aerodynamic, technologically advanced automobiles. Automobiles that represent the next step for the American driver and for Chrysler.

THE NEXT STEP. QUIET POWER THAT IS MAKING A LOT OF NOISE IN ENGINEERING CIRCLES.

Any car claiming to be a "driver's car" must have credentials under the hood. In the engine compartments of Eagles, you'll find things like a



SUMMIT 1.6L ENGINE*

hemispherical combustion chamber. Multi-port fuel injection. Dual overhead cams. Four valves per cylinder. These are what make Eagles fly. And what give drivers the confidence to handle the hundreds of driving decisions they must make day in and day out.

THE NEXT STEP. ROOM WITH A POINT OF VIEW.

The Eagle Premier has more *usable* room than any car in its class. And no car in its class surpasses the Eagle Medallion or the all-new Summit for roominess.** Every Eagle—Premier,

*Optional. **Interior comparisons based on EPA Interior Volumes and Ward's Intermediate Standard, Compact Standard, and Subcompact Classes. 1988 competitive model data used where 1989 data not yet available. †Optional leather seats with vinyl trim. Late availability. ††Protects engine and powertrain for 7 years or 70,000 miles and against outerbody rust-through for 7 years or 100,000 miles. See limited warranty at dealer. Restrictions apply.

All listed features not available on all Eagle models.

Jeep is a registered trademark of Jeep Eagle Corporation.

Buckle up for safety.



THE NEXT STEP. THE ALL-NEW EAGLE SUMMIT.

Medallion, and Summit—is an ergonomic masterpiece. Eagle's thoughtfully placed controls and adjustable seats reduce fatigue and keep you comfortable whether you drive around the block or around the country.



PREMIER ES LEATHER SEATING*

THE NEXT STEP.
THE SUSPENSION SYSTEM
WITH STREET SMARTS.

Our thinking on the whole issue of driver confidence goes something like this: A proper car is one that doesn't isolate you from the road, but rather one that keeps you in touch with it.

Eagles are designed to accomplish this by obtaining a meticulous balance between power, braking, and handling. A word about handling.



PREMIER ES.

We give Eagle cars things like a four-wheel independent suspension. Front and rear stabilizer bars. MacPherson struts. And a lot of little niceties that make the whole package sure-footed... athletic... confident.

A word about confidence. Every Eagle carries Chrysler's exclusive 7-year/70,000-mile Protection Plan††



THE NEXT STEP.

Enough of the rhetoric. Ultimately, what will make Eagle get off the ground is you, the American driver. To determine how successfully we've done our job, you must sit in an Eagle. You must drive an Eagle. If all goes as we expect, you will own or lease an Eagle.

For further information, call 1-800-JEEP-EAGLE.

Expect the Best.



New pitch in Sydney.

Your business trip is on a roll now. Of course, you had the advantage of getting off on the right foot. You called United.

To Sydney, Melbourne, and Auckland as well, United's all-747 fleet gives you all the advantages. Fine food, fine drink, and for First Class passengers, sleeper seats and our exclusive Concierge Service.

United. Rededicated to giving you the service you deserve. Come fly the friendly skies.



UNITED

A I R L I N E S

TOKYO • OSAKA • HONG KONG • SEOUL • TAIPEI • SYDNEY • MELBOURNE • BEIJING • SHANGHAI • AUCKLAND • SINGAPORE • MANILA • BANGKOK

© 1988 SCIENTIFIC AMERICAN, INC