SCIENTIFIC AMERICAN

MARCH 1994
$3.95

Visiting yourself in the past.

Rewriting the genes.

Information highwaymen.

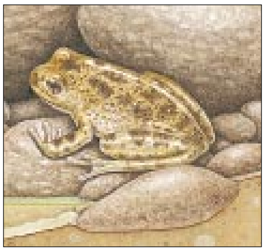
NAUTILE
IFREMER
IFREMER

In the deep Atlantic, Nautile *hunts for clues to the forces that make continents drift.*

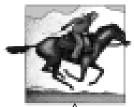
SCIENTIFIC AMERICAN

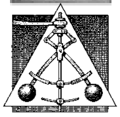
March 1994 Volume 270 Number 3

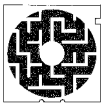
Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 1994 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in 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. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111, or fax: (212) 355-0408.

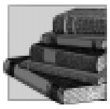


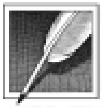
DEPARTMENTS











THE COVER painting depicts the *Nautile* as it skims along the Mid-Atlantic Ridge, the huge north-south scar bisecting the seafloor. The submersible, built by the French oceanographic institute IFREMER, can reach a depth of six kilometers. It houses three people in a 1.8-meter-diameter titantium sphere, whose portholes allow for external viewing. The *Nautile* collects rock samples that investigators use to determine how convection in the mantle affects the earth's surface features (see "The Earth's Mantle below the Oceans," by Enrico Bonatti, page 44).

THE ILLUSTRATIONS

Cover painting by George Retseck

Page	Source
37	Steve Vidler/Leo de Wys, Inc.
38	Johnny Johnson
39	Patricia J. Wynne
40	Johnny Johnson
41	Ed Kashi/J. B. Pictures
42	Courtesy of Ag-Chem Equipment Co., Inc.
44–45	Jack Harris/Visual Logic
46	Gabor Kiss (*top*), Dimitry Schidlovsky (*bottom*)
47	William F. Haxby, Lamont-Doherty Earth Observatory
48–49	Dimitry Schidlovsky (*top*), Jack Harris/Visual Logic (*bottom*)
50–51	Dimitry Schidlovsky
52	Mario R. Capecchi
53	Tomo Narashima
54–57	Jared Schneidman Design
58	Mario R. Capecchi (*top*), Tomo Narashima (*bottom*)
63	Courtesy of International Business Machines Corporation
64–65	Ian Worpole
66	IBM (*left*), Johnny Johnson (*right*)
67	Ian Worpole
68–69	Patricia J. Wynne
70–73	Dimitry Schidlovsky
74	© 1991 by Bill Watterson, Andrews and McMeel (a Universal Press Syndicate Company)
76–77	Yechiam (Eugene) Gal
78–79	Jared Schneidman Design
80	Steven Rubin/J. B. Pictures
81	D. Aubert/Sygma
82–83	Vaughan H. Shoemaker
84–85	Roberto Osti; Jason Küffer (*maps*)
86	Roberto Osti
87	Arthur Gloor/Animals Animals (*top left*), Paul Freed/Animals Animals (*top right*), Rodolfo Ruibal (*bottom*)
88	Johnny Johnson
90–91	Patricia J. Wynne
92	Jason Goltz
93	Jared Schneidman Design
94	Stephanie Rausser
98	Jared Schneidman Design
99	Chris Usher/Black Star
100	Jared Schneidman Design
101	Stephanie Rausser
110–111	Patricia J. Wynne

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Michelle Press, *Managing Editor;* John Rennie, *Associate Editor;* Timothy M. Beardsley; W. Wayt Gibbs; Marguerite Holloway; John Horgan, *Senior Writer;* Philip Morrison, *Book Editor;* Madhusree Mukerjee; Corey S. Powell; Ricki L. Rusting; Gary Stix; Paul Wallich; Philip M. Yam

ART: Joan Starwood, *Art Director;* Edward Bell, *Art Director, Graphics Systems;* Jessie Nathans, *Associate Art Director;* Johnny Johnson, *Assistant Art Director, Graphics Systems;* Nisa Geller, *Photography Editor;* Lisa Burnett, *Production Editor*

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Molly K. Frances; Daniel C. Schlenoff

PRODUCTION: Richard Sasso, *Vice President, Production;* William Sherman, *Production Manager; Managers:* Carol Albert, *Print Production;* Janet Cermak, *Quality Control;* Tanya DeSilva, *Prepress;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Eric Marquard, *Special Projects;* Leo J. Petruzzi, *Manufacturing & Makeup;* Ad Traffic: Carl Cherebin

CIRCULATION: Lorraine Leib Terlecki, *Associate Publisher/Circulation Director;* Katherine Robold, *Circulation Manager;* Joanne Guralnick, *Circulation Promotion Manager;* Rosa Davis, *Fulfillment Manager*

ADVERTISING: Kate Dobson, *Associate Publisher/Advertising Director.* OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager;* William Buchanan, *Manager, Corporate Advertising;* Peter Fisch, Randy James, Elizabeth Ryan. Michelle Larsen, *Director, New Business Development.* CHICAGO: 333 N. Michigan Ave., Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager.* WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025; Lisa K. Carden, *Advertising Manager;* Tonia Wendt. 235 Montgomery St., Suite 724, San Francisco, CA 94104; Lianne Bloomer. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

MARKETING SERVICES: Laura Salant, *Marketing Director;* Diane Schube, *Promotion Manager;* Mary Sadlier, *Research Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; Vivienne Davidson, Linda Kaufman, Intermedia Ltd., Paris; Karin Ohff, Groupe Expansion, Frankfurt; Barth David Schwartz, *Director, Special Projects,* Amsterdam. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.; SINGAPORE: Hoo Siew Sai, Major Media Singapore Pte. Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie M. Beaumonte, *General Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue, New York, NY 10017-1111
(212) 754-0550

CHAIRMAN AND CHIEF EXECUTIVE OFFICER: John J. Hanley

CO-CHAIRMAN: Dr. Pierre Gerckens

CORPORATE OFFICERS: *President,* John J. Moeling, Jr.; *Chief Financial Officer,* R. Vincent Barger; *Vice Presidents,* Robert L. Biewen, Jonathan Piel

DIRECTOR, ELECTRONIC PUBLISHING: Martin Paul

CHAIRMAN EMERITUS: Gerard Piel

PRINTED IN U.S.A.

LETTERS TO THE EDITORS

Free-for-All on Trade

I was hopeful that "The Case for Free Trade," by Jagdish Bhagwati, and "The Perils of Free Trade," by Herman E. Daly [SCIENTIFIC AMERICAN, November 1993], would help clarify the policy confusion gripping this issue. Instead, despite some insightful analysis, these two men were talking past each other like seasoned political rivals.

International trade agreements like NAFTA and GATT do not allow countries to restrict the import of products based on how those products are made. Why? Because a country could theoretically block all imports with its environmental, health and labor laws—throwing traditional notions of sovereignty and comparative advantage into a tailspin. The answer, easier said than done, is to define concepts such as national sovereignty more precisely through new trading rules that account for the myriad threats to the global environment. No one has yet given a coherent reason why the U.S. must accept dolphin-deadly tuna as GATT desires. Bhagwati's "Pandora's box" response begs the question.

To his credit, Daly has identified sustainable resource scale, full-cost internalization and migratory capital as concepts that could advance the integration of trade and the environment. But Daly has a problem, too: How are the developing countries going to react to a "no growth" mandate? The challenge facing both Daly and environmental organizations is to define explicitly what is meant by sustainable development—an appealing but factually ambiguous concept.

WILLIAM J. SNAPE III
Defenders of Wildlife
Washington, D.C.

Bhagwati repeats the largely unsubstantiated dogma that only rich individuals and nations express concern about environmental values. A recent Health of the Planet survey by the Gallup Organization challenges that dogma. It finds that in nine out of 12 developing nations surveyed, a majority of the respondents considered environmental protection to be a higher priority than economic growth.

The author also skates on very thin ice when he cites the Grossman and Krueger study as evidence that "environmentalists are in error when they fear that trade, through growth, will necessarily increase pollution." That study focuses only on sulfur dioxide, particulate matter and smoke, which one would expect to see diminish as economies turn to less immediately hazardous means of generating energy. Yet developed economies produce far more toxic chemicals, far more radioactive wastes, far more carbon dioxide and far more ozone depletors. The adverse environmental impacts of those pollutants are much more worrisome in the long run.

Bhagwati is correct in one sense: many of the differences between economists and environmentalists can be attributed to misconceptions. As his article indicates, however, environmentalists are not always the ones missing the essential concepts.

TOM E. THOMAS
Environmental Management Program

JAMES R. KARR
Institute for Environmental Studies
University of Washington

Bhagwati writes that the Grossman and Krueger study found that sulfur dioxide levels fell as per capita income rose. He notes that "the only exception was in countries whose per capita incomes fell below $5,000" and implies that those exceptions are rare. But according to the data in Daly's pie chart, 85 percent of the world's population earns only $1,000 annually per capita. Either Bhagwati has not elucidated his case properly, or it is his argument, not the environmentalists', that is in error.

SEAN ALLEN-HERMANSON
Dartmouth, Nova Scotia

Bhagwati replies:

Snape asserts that no "coherent" defense of the GATT panel's tuna-dolphin decision has yet been given by anyone. Rubbish; my article certainly does so. He then shifts ground and says instead that it "begs the question." What question? Why? His conclusions are more obvious than his arguments.

Thomas and Karr are no better. Concerns over the environment can and do crisscross per capita rankings: the GATT Report on Trade and the Environment stated that clearly. As I wrote: "Rich countries today have more groups worrying about environmental causes than do poor countries." That is both correct and wholly different from the Thomas-Karr assertion that I believe "only rich individuals and nations express concern about environmental values"! Allen-Hermanson infers from my writing what I do not argue or believe. The implication is his error, not mine.

Fortunately, not all environmentalists are so careless or contemptuous of reasoned argument. I continue to believe that a bridge can be built between their concerns and those of economists.

Daly replies:

I would not support a no-growth mandate for the developing world, at least not yet. Sustainable development must begin in the North and spread rapidly to the South. But the current model is far from sustainable, and the North should not preach what it does not even try to practice.

Defining sustainable development is not so hard: it is qualitative improvement without quantitative expansion—specifically without growth in resource throughput beyond nature's regenerative capacity or beyond its capacity to absorb or recycle wastes. Nonrenewable resources are depleted no faster than renewable substitutes are developed. All important concepts have some ambiguities, but I submit that this definition of sustainable development is no more ambiguous than economists' definitions of money.

Letters selected for publication may be edited for length and clarity. Manuscripts will not be returned or acknowledged unless accompanied by a stamped, self-addressed envelope.

ERRATA

Contrary to an implication in "Diamond Film Semiconductors" [October 1992], the group of Boris V. Spitsyn was not involved with research on polywater.

A news story on page 18 of the December 1993 issue erroneously stated that Targeted Genetics is using adeno-associated virus in its gene therapy for HIV infection. That virus is being used to develop a cystic fibrosis therapy; the HIV therapy uses a different virus.

50 AND 100 YEARS AGO

MARCH 1944

"A radically new form of 'lighthouse' radio relay station will make relaying of television programs a relatively simple matter after the war, according to Ralph R. Beal, Research Director of RCA Laboratories. He envisages that these unattended stations, located 20 to 50 miles apart, not only will link television stations into a national network but will open a new era in international communications. The relay transmitters will operate on microwaves with the energy concentrated almost in a bee line."

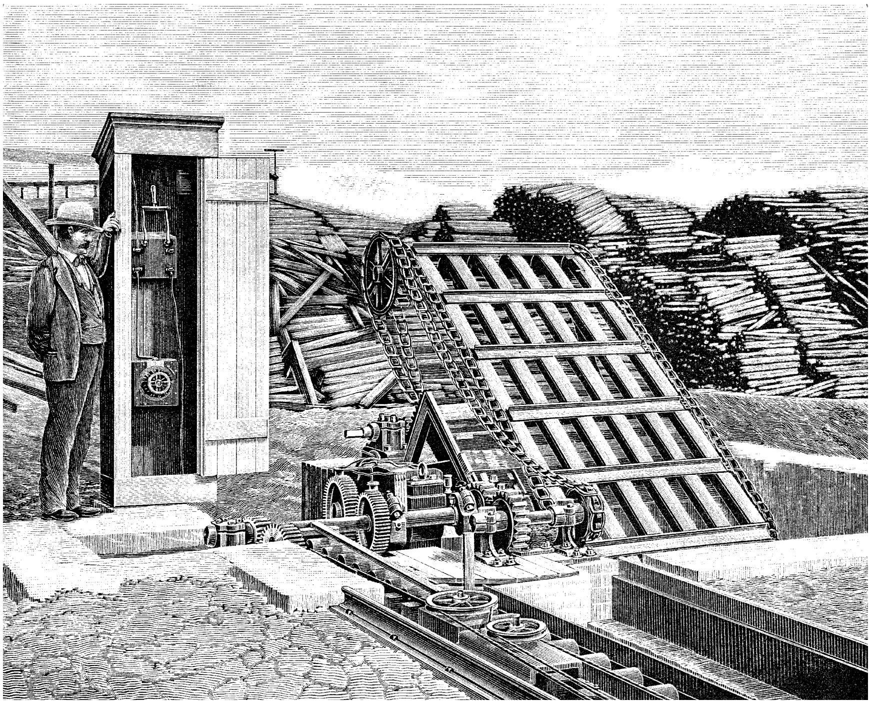
"Look upon natural gas as a raw material source for the chemical industry in the near future. Ninety-five percent of production is currently for industrial and household fuel. It is entirely probable, however, that more and more of this gas will be diverted to other purposes. Butadiene, glycerine, carbon tetrachloride, gasoline, sulfa drugs, and fertilizers are some of the products available directly or indirectly from natural gas."

"The recently completed State Street subway in Chicago is proving its worth in that city's vast network of transit lines. Although conceived originally as an aid to relieving the badly congested traffic conditions in the famous downtown 'Loop' section of the elevated rapid-transit lines, this modern transportation facility incorporates many conveniences for its patrons. For example: Escalators furnish effortless access to and from the loading platforms, and automatic ventilators provide fresh air within the subway. This 4.9-mile section is the first of four proposed units to be completed."

SCIENTIFIC AMERICAN

MARCH 1894

"Mr. F. Corkell, writing to the *Mining and Scientific Press,* says: On the night of Feb. 1, at Candelaria, Nevada, a brilliant meteor appeared. It made a tremendous illumination suddenly; the light was a dazzling electric blue, like many arc lights had shot into existence for about four seconds. Thirty seconds later a terrific explosion occurred, shaking the hills and echoing through the rocky caverns. There followed a boiling, sizzling roar, like an immense mass of red hot iron cooling in water. This lasted about fifteen seconds. None who saw or heard this meteor will forget it; they will relate it as a great event."



Handling steel billets by electrical power

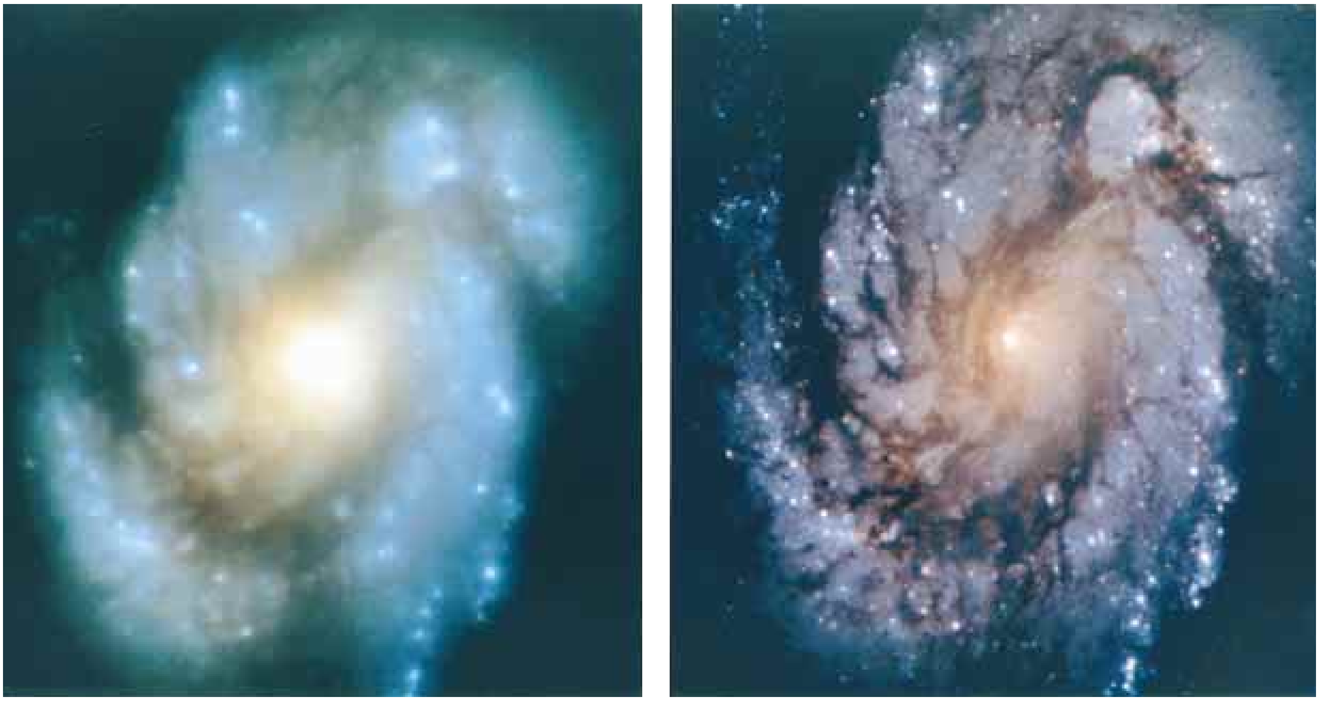
"Paul Bert has found by experiment that oxygen, this gas, vital above all others, is a violent poison, for the plant as for the animal, for the cellule as for the complete organism; and, if found in the air in certain proportions, immediately becomes an instrument of death. This is one of the most curious of recent discoveries. No oxygen, no life; too much oxygen, equally no life.—*Public Opinion, from Revue des Deux Mondes.*"

"One afternoon this winter, though walking briskly along, I was uncomfortably cold; my ears were so chilled as frequently to require the application of my gloved hands. I then began taking deep, forced inspirations, holding the air as long as possible before expulsion. After a few inhalations, the surface of my body grew warmer. The next to feel the effects were my ears. Within the time required to walk three blocks, hands and feet partook of the general warmth and I felt as comfortable as if the time had been passed by a glowing fire.—*E. B. Sangree, M.D., American Therapist.*"

"The camels now running wild in Arizona are the descendants of a small herd originally imported to Virginia City, Nevada. They were wanted for use in packing salt across the desert. Eventually they were sent to Arizona for packing ore. But they became footsore and useless and were turned adrift to shift for themselves.—*San Francisco Chronicle.*"

"Our illustration represents an electrical apparatus employed at the Illinois Steel Company, at Joliet, to load steel billets on flat cars with the minimum amount of manual labor. Billets to be shipped are delivered from the yard to a long line of rollers, partly shown at the left in the illustration, and are thus carried along until they strike a deflecting plate, by which they are conveyed to an endless moving apron, set at an incline, as prominently shown. This apron first elevates and then drops the billets to a car, which lies on a depressed railroad track on the farther side."

SCIENCE AND THE CITIZEN



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

CORE OF SPIRAL GALAXY M100 as seen by the Hubble Space Telescope *before refurbishment* (left) *and after* (right). *The Wide Field and Planetary Camera that took the photo at the left was replaced to correct the error in* Hubble'*s main mirror.*

Image Enhancement

Hubble *repairs create euphoria and burnish NASA's reputation*

"A small change for a mirror, a giant leap for astronomy," was how Christopher J. Burrows of the Space Telescope Science Institute epitomized the feelings of the ecstatic astronomers who in January proudly showed off the first, brilliantly sharp images from the newly refurbished *Hubble Space Telescope.* A jubilant National Aeronautics and Space Administration wasted no time capitalizing on the success of December's shuttle mission during which astronauts corrected the blurry vision of the orbiting observatory. "We believe *Hubble* is fixed," declared administrator Daniel S. Goldin. NASA's own shaken reputation enjoyed some fixing as well.

The agency has suffered a series of ignominious setbacks in recent years, culminating in the loss of the *Mars Observer* last August. *Hubble* had been an orbiting embarrassment since two months after its launch in 1990, when NASA realized that the telescope's primary mirror had been manufactured to the wrong shape. As a result, *Hubble*'s performance had fallen far short of expectations.

The fix should enable the $1.5-billion telescope to fulfill its original promise. *Hubble* has a resolution at least 10 times better than that of any ground-based instrument, so it can see clearly throughout a volume of space 1,000 times larger. "Beyond our wildest expectations" was the verdict of Ed Weiler, *Hubble* program scientist. New gyroscopes, solar arrays and magnetometers also installed during the mission have improved *Hubble*'s stability and introduced backup capability for pointing.

Ever mindful of the need for friends on Capitol Hill, NASA invited Senator Barbara A. Mikulski of Maryland, chair of the senate subcommittee that oversees the agency's appropriations, to help announce the success. "The repair of *Hubble* is a benchmark," Mikulski said, flourishing pictures of a star taken with the telescope's Faint Object Camera before and after the refit. "There is now a confidence that the space station can be built. There will be the technical and astronaut capability to do it."

Most astronomers could not care less about the planned space station, but, like a Chagall bridegroom, they are over the moon about the wealth of data now likely to come from *Hubble.* Two major changes in the telescope enhance its capacities. One is COSTAR, the corrective optics package, which carries 10 button-size mirrors that remedy the error in the primary mirror for three *Hubble* instruments: the Faint Object Camera, the Goddard High Resolution Spectrograph and the Faint Object Spectrograph. Another instrument was sacrificed to make room for COSTAR. The other important fix is the new Wide Field and Planetary Camera (WFPC-2), which corrects the fault in the primary mirror without COSTAR's help.

As of late January, the spectrograph mirrors on COSTAR had not all been tested, but NASA officials were confident. The Faint Object Camera mirrors of COSTAR, as well as WFPC-2, both worked as soon as they were activated, needing little adjustment to achieve almost perfect performance.

WFPC-2's performance is now "very close to the theoretical limit," according to Burrows. Between 60 and 70 percent of the light from a point source imaged with the camera falls within a circle 0.2

arc second across. Because of spherical aberration caused by the defect in the primary mirror, the old WFPC could put only 12 percent of the light from a point in the same area. At the American Astronomical Society meeting in January, J. Jeffrey Hester of Arizona State University presented dramatic images made with the new camera of the stellar nursery known as R 136.

The first WFPC revealed that R 136, a cluster in the nebula 30 Doradus, comprised several hundred stars; now WFPC-2 sees more than 4,000. WFPC-2 has also made vivid images of the giant star Eta Carinae.

But even such prizes pale before the prospect that the refurbished telescope will enable astronomers finally to determine the value of a key cosmological parameter: the Hubble constant. The Hubble constant, named—like the telescope—for the American astronomer Edwin Hubble, is a number that relates the velocity of an astronomical object to its distance. It thus leads straight to an estimate of the age of the universe. At present, astronomers disagree by a factor of two over the size of the Hubble constant. Consequently, the age of the universe cannot be calculated with any precision beyond that of a hand wave (the number is thought to be between 10 and 20 billion years).

NASA

PART OF THE GREAT NEBULA in Orion, a region of recent star formation, is seen in unprecedented detail in an image from Hubble*'s new camera. Colors correspond to different gases.*

To resolve the argument, it will be necessary to bring into focus variable stars called Cepheids in galaxies as much as 50 million light-years away. As Hubble realized, the fact that the absolute brightness of a Cepheid can be inferred from its periodicity makes them useful as cosmic milestones; stars of the same brightness look dimmer the farther away they are. The old WFPC could resolve Cepheids that lay only 12 million or so light-years away. But WFPC-2 easily resolves individual stars in the galaxy M100, which float at a distance between 35 million and 80 million light-years. Some of those stars

Chaotic Chaos

Students of chaos have clung to the notion that chaotic systems retain some shreds of order. The shreds manifest themselves in the form of an attractor, a pattern of behavior toward which the system periodically settles. Identifying the attractor enables one to predict the final behavior of a chaotic system, at least in a qualitative, statistical sense. That comforting notion has been damaged by Edward Ott of the University of Maryland and John C. Sommerer of Johns Hopkins University and their colleagues. They have shown that for certain systems that have more than one attractor, even qualitative predictions are impossible. “The repeatability of an experiment gets thrown into question,” Ott says.

The problem is rooted in the way a chaotic system determines which attractor to follow. The initial conditions that control the choice are said to be located in a basin of attraction. Ott and Sommerer have spoiled the party by showing that a basin may be rather leaky: it may have “holes” that make it impossible to predict which attractor the system will follow.

Building on earlier mathematical work, the physicists used a computer to conduct numerical experiments in which a particle moving on a frictional surface is occasionally pushed. Consequently, the particle could begin moving either periodically or sporadically. The researchers found that even for this fairly simple system they could not determine which of the two attractors the particle would chase, because one basin is riddled with pieces of the other basin. In fact, every area in one basin, no matter how small, contained pieces of the other basin within it. “Hence, arbitrarily small changes can cause the system to go to a completely different attractor,” Ott remarks. The only way to guarantee an outcome is not to have any error or noise whatsoever—a practical impossibility for real systems. And, anyway, what kind of chaos would that be?

Ott points out that the results differ from other forms of chaos in which the starting point straddles the boundary between two basins of attraction. In such borderline situations, one might be able to move the starting point away from the boundary so that the attractor can be predicted. The same cannot be done for systems that have riddled basins, because no region is free of holes. “You’re always on the borderline,” Ott explains.

Although riddled basins appear only in situations that have certain spatial symmetries, they are probably not rare. “A lot of physics is based on conservation laws, which are based on symmetries,” Sommerer observes. Currently the workers are looking for real physical phenomena that have riddled basins. They suspect that turbulent fluids, chemical mixtures and lasers may be among such systems. Sommerer even speculates that experimentalists have already encountered this kind of chaos. Projects that went awry the second time around could have been a result of the mischievous property of riddled basins. “I have a sneaking suspicion this might be the case for some,” he intones.

—Philip Yam

may well be Cepheids. “We appear to have a camera that should be capable of that fundamental task,” says Jon Holtzman of Lowell Observatory.

Images from the Faint Object Camera, now seeing sharply for the first time thanks to COSTAR, are no less impressive. Peter Jakobsen of the European Space Agency, which built the Faint Object Camera, drew spontaneous applause at the January meeting when he showed an image of supernova SN1987A from the instrument. The photograph clearly resolved the central fireball of the exploding star.

Robert Jedrzejewski of the Space Telescope Science Institute elicited the same reaction when he exhibited a just-drawn diagram of the brightness and temperature of stars in the globular cluster 47 Tucanae. Perhaps the most spectacular image was displayed by F. Duccio Macchetto, also of the Space Telescope Science Institute, who presented a view of the fiery heart of a Seyfert type 2 galaxy, NGC 1068. The core, believed to contain a black hole, shines a billion times brighter than the sun. Although infalling matter obscures the core, the new photograph shows detailed structure around the inferno where before there was only a blur. Edwin Hubble would have been proud. —*Tim Beardsley*

Down the Green

As Ras grabs headlines, workers find a short signaling pathway

The Ras pathway, one route by which DNA is turned on and off by signals arriving at the cell membrane, has been keeping cell biologists busy for the past year. If molecular biology were billiards, the Ras pathway (so named because a key element in it is the Ras protein) would be an epically complex combination shot using every ball and cushion to angle the target ball, a growth signal, toward the pocket.

As the Ras story unfolded in a rapidly building series of papers, other scientists were quietly uncovering a much simpler pathway, a kind of straight shot down the green. The control sequence they describe carries chemical information from the cell membrane to the nucleus via only two key families of proteins, Janus kinases (JAK) and signal transducers and activators of transcription proteins (STAT), without relying on secondary messengers. The sequence begins when an occupied membrane receptor phosphorylates a JAK kinase, which in turn calls STAT proteins into action. The STAT proteins then journey to the nucleus, where alone or in tandem with other DNA binding proteins they stimulate transcription.

“The Ras pathway is a much more complex, sensitive interplay of proteins than what we’re looking at,” explains James E. Darnell, Jr., of the Rockefeller University, one of the discoverers of the new pathway. “I don’t believe the Ras pathway is the decisive pathway for transcriptional signals, but it is critical in growth control.” Darnell first noticed the role STAT proteins play in cells responding to signals from two interferons, IFN-alpha and IFN-gamma, that dock at different membrane receptors. Both signals cause antiviral reactions as well as restrained growth in many cell types, but they were presumed to use independent pathways. It turns out that both could engage the same protein, Stat91.

Meanwhile biologists at the Imperial Cancer Research Fund in London were developing an additional line of evidence. The English investigators selected mutant cells incapable of responding to IFN-alpha or IFN-gamma, or both. But the group found that the IFN response could be restored by adding to the cells genetic instructions for the production of Stat91. The various cell lines showed that not only was the activation of Stat91 required for a cell to respond to the interferons but that separate sets of JAK proteins were needed to interact with the STAT proteins in order to initiate either reply.

Several laboratories have since demonstrated that the JAK-STAT pathway is involved in cell responses to other growth factors and cytokines. "We do know that the JAK kinases decorate the STAT proteins, but we do not yet know who phosphorylates whom," Darnell adds. The JAK-STAT pathway may well facilitate a vast number of cellular responses. Much like Lego blocks, these proteins may snap together in a number of configurations to activate many different genes. Furthermore, the distinct protein arrangements bind with varying affinities to their related gene sites. STAT proteins may thus enable cells to distinguish degrees of urgency in the extracellular signals they receive.

"We believe different extracellular signals probably trigger a different profile of gene responses," Darnell says. "But we don't know how many separable response elements there are in the genome or how many different permutations of transcription factors will be required." For a straight shot down the green, this setup is beginning to look fairly complicated. *—Kristin Leutwyler*

Spinning Out

Physicists cannot agree on the origin of proton spin

Just how much of a proton's spin comes from that of its quarks? Ask an experimenter, and the answer is 10, 55 or, most recently, 35 percent. If this isn't confusing enough, ask a theorist. Predictions range all the way from 0 to 100 percent; a good number of theorists come in at about 65. Others argue that this percentage, called sigma (Σ), is simply incalculable.

That our best-loved subatomic particle should have come to such a pass is perplexing. The proton's composition is seemingly clear-cut: two up quarks and one down quark held together by gluons. Like many other elementary particles, the proton—and the quark—carries a built-in angular momentum, known as spin, that has a magnitude of $^1/_2$ (of a quantum unit). But because the proton is made of quarks, its spin is plausibly dissectable into that of its quarks. The debate on how to implement this dissection continues while the proton, so to speak, waits on the operating table. Alongside lies its significant other, the neutron; both have the same Σ.

In 1988 experimenters at CERN, the European laboratory for particle physics, announced that Σ is roughly 10 percent. This finding, conflicting as it did with most theoretical expectations, provoked a swirl of activity. In 1993 a group at the Stanford Linear Accelerator Center (SLAC) found Σ to be 55 percent, whereas the Europeans came back with a new measurement: again 10 percent. Theorists and experimenters went right back to their desks and labs. In early January the CERN collaboration declared its latest result: about 35 percent. The Stanford group expects to reveal *its* new result by this summer.

So what is the value of Σ? It will be some time before the dust settles: the measurements have large errors (of tens of percents), so the results quoted are actually quite fuzzy. Meanwhile the confusion on the experimental side is mirrored by theoretical uncertainty about just how to slice up the proton's spin.

The hitch is the intricacy of the real-life proton. Its quarks and gluons interact with one another in myriad ways prescribed by the theory of quantum chromodynamics (QCD). These interactions are so hard to calculate that theorists try to abstract the essence of QCD in simpler models, which they then use to make predictions.

In the “naive” quark model, one of the three quarks spins in a direction opposite to the other two; when we sum the spins of all three, we get $\frac{1}{2} + \frac{1}{2} - \frac{1}{2} = \frac{1}{2}$—which is simply the proton’s spin. In this model *all* the proton’s spin comes from the quarks’: Σ is 100 percent. More complex models allow the quarks to orbit one another; some of the proton’s spin then comes from the quarks’ orbital angular momentum and only about 65 percent from their spin. John Ellis of CERN and Robert L. Jaffe of the Massachusetts Institute of Technology have predicted a Σ of 60 percent. They use an elegant formulation of QCD that takes the up, down and strange quark masses to be equal, while ignoring the contribution to Σ from (spontaneously created) strange quarks.

All the above calculations are questioned by Alfred H. Mueller of Columbia University and his collaborators. They argue that gluons mix with quarks so intimately that it is impossible to predict the spin contribution of the (unglued) quarks. At the far side of the debate lies the Skyrme model, which sees the proton as a hedgehoglike kink in a quantum field; it gives a Σ of 0 percent.

“The problem,” points out Xiangdong Ji of M.I.T., “is that we really don’t have a good model for the proton.” Theorists do agree, however, on one aspect: an up quark’s contribution to Σ should be quite similar to that of a down quark. If the experiments rule otherwise, violating a 1966 prediction by James D. Bjorken of SLAC, then QCD itself will be called into question. The divergence of CERN and SLAC data has threatened just that. Looks like the proton will remain on the operating table for a while. *—Madhusree Mukerjee*

Molecular Mischief

Spectroscopic studies may point to a cause of schizophrenia

In recent years, investigators looking for physiological abnormalities in the brain that might be associated with schizophrenia have focused on a region known as the prefrontal cortex. Diverse clues suggest that it is a site of crucial events. One is the observation that schizophrenics have below-average blood flow in their prefrontal cortices, which indicates depressed activity there. Another clue, found by Patricia S. Goldman-Rakic, Charles J. Bruce and Martha G. MacAvoy of Yale University, is that cuts at specific locations in the prefrontal cortices of rhesus monkeys make the animals prone to errors on tests designed to use “working memory.” Schizophrenics do poorly on the same type of test. Lesions at other sites in the simian prefrontal cortex cause jerky eye movements when a fast-moving object is tracked—also a characteristic feature of schizophrenia.

Jay W. Pettegrew of the University of Pittsburgh has pressed on to the molecular level—in human beings. Pettegrew uses nuclear magnetic resonance spectroscopy to measure what he calls the “molecular mischief” of the disease. In a study that compared 24 schizophrenics who had never received antipsychotic medication with 29 healthy and matched control subjects, Pettegrew found that the patients had markedly lower levels of chemicals called phosphomonoesters in their prefrontal cortices.

At the annual meeting of the Society for Neuroscience in Washington, D.C., late last year, Pettegrew presented evidence that this chemical abnormality has relevance to symptoms. Patients who were more sick, as assessed by tests of verbal fluency and other measures, had lower levels of phosphomonoesters than those who were less sick.

Phosphomonoesters are building blocks for the phospholipids found in

membranes surrounding neurons. Pettegrew thinks schizophrenics have an impaired ability to synthesize the membranes. Other compounds known as phosphodiesters are also present in elevated amounts in schizophrenics, a finding that could indicate that in such patients the breakdown of neural membranes is accelerated.

The results, which Pettegrew says have been replicated, seem to fit in with the finding that the cells in the prefrontal cortex of a schizophrenic individual are typically smaller and more densely packed than those in normal brains. Pettegrew proposes that in schizophrenics a "pruning" of neurons that normally occurs during adolescence is exaggerated.

If healthy children who have unusually low phosphomonoester levels are more likely than others to show symptoms of schizophrenia later—a big "if"—then, Pettegrew suggests, giving such children drugs designed to stimulate the growth of neurons might forestall the development of the disease. "We should start to think about schizophrenia as something we can prevent," he declares.

First, such drugs must be found, however. Work on this disease has failed to redeem its promise many times before. *—Tim Beardsley*

Gene Rich, Cash Poor

The genome project has plenty of findings but not dollars

By all the short-term measures, the Human Genome Project is succeeding beyond its planners' dreams. Four years ago it was launched as a 15-year effort to read and decipher the DNA in human cells. But within two years researchers will have fairly detailed maps of all the chromosomes and may even know where nearly all the genes are. Those discoveries are ushering in a new age in biology. With genetic decoders in hand, investigators will soon be finding molecular solutions to long-standing puzzles of development and cellular function.

At the same time, however, geneticists are also worrying about whether the program has the technical and financial resources to keep the party going. "It is very difficult to look at the budget we have and see how we're going to get it done by 2006," laments Francis Collins, director of the genome project at the National Institutes of Health.

Researchers unanimously agree that the compilation of genetic linkage and physical maps, which indicate where genes appear on chromosomes, are proceeding on or ahead of schedule. Just before Christmas, in fact, the physical mapping project received a gift from Daniel Cohen of the Centre d'Étude du Polymorphisme Humain (CEPH) and Généthon in Paris, who released a map of more than 90 percent of the human genome. Cohen is a pioneer in the use of large pieces of yeast DNA, called megaYACs, as mapping tools. His group had dissected chromosome 21 by that method in 1992. But Cohen decided that handling the human chromosomes one by one was too inefficient, so his team changed tactics and analyzed all of them simultaneously.

If the CEPH map covers virtually the entire human genome, why isn't that part of the project finished? The resolution of the CEPH map is low: the genetic landmarks it charts are millions of nucleotide base pairs apart. Geneticists usually need to be within 100,000 bases or so of a marker to find and sequence a specific gene. Collins believes a map with a 300,000-base resolution could be available in 1995.

The ultimate blueprint, and the goal of the sequencing effort, will be the readout of all three billion base pairs that make up human DNA. But this leg of the genome project is looking rickety.

Simple arithmetic shows why: even the best laboratories can now sequence only about two million base pairs a year, and only four or five laboratories in the U.S. can work that fast. At that rate, sequencing the entire genome would take more than 300 years. The sequencing timetable was always built on the assumption that technological improvements would keep the rate of sequencing rising exponentially. But basic research into developing sequencing technologies has suffered from neglect.

The good news, Collins says, is that meeting the 2006 deadline "isn't going to require a blue-sky breakthrough. We're not going to have to depend on something we can't think of yet." Researchers are already raising the speed and efficiency of the electrophoretic gel equipment they use to analyze DNA. The biggest jump, most investigators think, will come from automating repetitive tasks now done manually.

Molecular geneticists are also looking hopefully to improvements in sequencing techniques such as primer walking. Researchers can make primer molecules of DNA about 18 bases long that will bind to a unique location in the genome. With enzymes, they can extend a bound primer by several hundred more bases complementary to the genomic DNA. By sequencing the elongated primer, they can then determine the genome sequence. A sequence from the far end can serve as a primer for the next "step" along the DNA. Unfortunately, primer walking in this way is labor intensive: a new 18-base primer must be synthesized for each round of walking, and that typically takes a day.

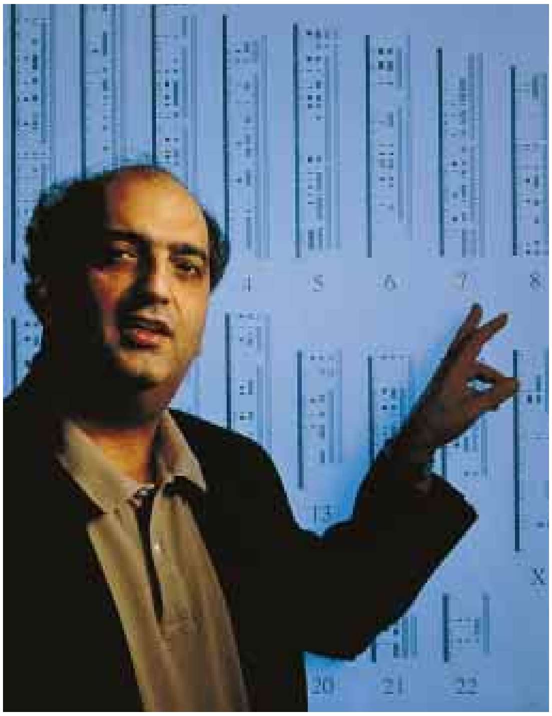
F. William Studier of Brookhaven National Laboratory and his colleagues have found a way to simplify primer walking. Their approach uses a library of hexamers (six-base primers) and a protein that binds to single-strand genomic DNA. The binding protein prevents individual hexamers from pairing stably with the DNA. But three end-to-end hexamers—the equivalent of an 18-base chain—reinforce one another enough to muscle the protein aside. The advantage of the technique is that there are only 4,096 different types of hexamers, as opposed to more than 68 billion 18-base primers. All the necessary hexamers can therefore be prepared in advance as off-the-shelf reagents.

One aspect of the sequencing effort—finding the genes—is moving ahead at astonishing speed with existing technology. By most estimates, less than 3 percent of the billions of bases in the genome are parts of genes: the rest consists of regulatory sequences and junk DNA. Several years ago, while he was a researcher at NIH, J. Craig Venter discovered how to find the gene needles in the DNA haystacks. Venter isolates the messenger RNA molecules transcribed from active genes in cells, then reverse-transcribes them into DNA. He identifies a few hundred bases from these DNAs and uses computers to look for similar strings of bases in the data banks of known sequences. In this way, he is able to flag those sequences as genetic, even though the actual function of the gene may remain obscure.

Venter was soon identifying thousands of genes every month. Today the Institute for Genomic Research, which he founded in Gaithersburg, Md., is reportedly identifying about 600 genes a day. If the institute meets its announced target, it will have labeled half of all the human genes by April of this year. Other laboratories have also adopted his methods. "Collectively, through the worldwide effort, the majority of genes should be known by the end of 1995," Venter predicts.

Still, researchers emphasize that Venter's gene tagging does not replace comprehensive sequencing. "You've probably heard the claims about being able to identify essentially all the genes in the genome within a few years," cautions David Galas, former head of the Department of Energy's genome research program. "Regardless of whether that's literally true, you're certainly going to be able to find a lot of genes. But how you use that information to probe the organization and expression of genes is still unclear." Sequencing therefore remains essential.

In short, the ideas for how to speed up sequencing are already on the table. The challenge will be to translate them into practical tools in dozens of laboratories. And that is why Collins says more funding for technology development is necessary. He notes that federal funding for the project has leveled off at about 60 percent of its inflation-adjusted $200-million annual need. "Right now is a very critical time," Galas says. "There are important advances that need to be developed further. It would be a good time to get a boost in funding." —*John Rennie*



STEVE MUREZ *Rapho, Black Star*

DANIEL COHEN of Généthon shows off his latest prize, the best map yet of human chromosomes.

Cold Confusion

Assault on the link between CO_2 and global climate

For those who worry about climatic change, the terms "carbon dioxide" and "global warming" often seem as inseparable as "yin" and "yang." Since the 1980s several studies of ice cores drilled from the thick glaciers on Greenland and Antarctica have offered evidence of a correlation between carbon dioxide and global climate. Those cores showed that carbon dioxide levels in the atmosphere were much lower during ice ages than during comparatively warm periods such as the present. The finding has amplified the ominous implications of the huge quantities of carbon dioxide that humans continue to dump into the air.

Now the ice core data on atmospheric carbon dioxide have come under assault. At the December 1993 meeting of the American Geophysical Union, Alex T. Wilson of the University of Arizona asserted that current measurements of prehistoric carbon dioxide levels are considerably in error. In particular, Wilson finds that the levels during recent ice ages were only marginally lower than in modern times—and far higher than most scientists have believed.

The source of the error, according to Wilson, is the technique used to deduce what the composition of the air was thousands of years ago. In the conventional approach, workers crush a sample of ancient ice to release the pockets of air trapped inside and then measure the gas that emerges. Although the process seems simple enough, Wilson perceives "a pretty surprising assumption" lurking below the surface.

The ice-crushing technique works only if the freed air has the same composition as the air originally trapped, millennia ago, under layers of overlying snow. Wilson notes, however, that deep in the ice layers the pressure is so great that air dissolves into the surrounding ice, and bubbles disappear. When brought to the surface, the ice decompresses, and the air reappears in bubbles or voids in the ice. Wilson claims that about one quarter of the carbon dioxide remains trapped in the ice itself and so never shows up in the laboratory measurements.

Working with Austin Long, also at the University of Arizona, Wilson is utilizing an alternative method for extracting air from the archaic ice. In essence, they evaporate the ice in a vacuum chamber (a process known as sublimation) and then analyze everything that comes out. Their results look quite a bit different from those of their colleagues. A 35,000-year-old ice sample from the Greenland Ice-Sheet Project 2 (GISP2) yielded 250 parts per million of carbon dioxide, only slightly below the modern but preindustrial levels of about 270 parts per million. For comparison, conventional techniques give a value of roughly 180 parts per million—a considerable discrepancy.

Many of Wilson's colleagues question his technique. Martin Wahlen of the Scripps Research Institute, who also performs carbon dioxide measurements on the GISP2 ice cores, maintains that "from our experiments and tests, we have no clue that he might be right." Bernhard Stauffer of the University of Bern is more direct: "Wilson is definitely wrong with his arguments." Stauffer is concerned that the sublimation technique could be measuring contaminants in the ice or in the apparatus itself that give the impression of artificially high carbon dioxide concentrations.

Wilson counters that his tests show negligible signs of contamination. He also notes that his results disagree with those from ice crushing only for deep core samples—those in which air once dissolved into the ice. "There is no doubt that 180 parts per million is far too low," he says. Stauffer, Wahlen and other climate researchers complain that Wilson has not been terribly open about his methodology; in particular, they worry that he has not shown other researchers the dry runs of his apparatus.

Even if Wilson and Long's numbers hold up, they do not silence those who believe global warming is a genuine danger. Curt Covey of Lawrence Livermore National Laboratory notes that smaller variations in carbon dioxide between glacial periods and warmer eras could mean that climate may actually be more sensitive to changing levels of carbon dioxide than scientists have thought. On the other hand, it could underscore the considerable influence of other factors that affect global climate. As Covey observes, "You need more than carbon dioxide changes to get ice ages."

Indeed, the relation between carbon dioxide and ice ages is still far from clear. Paul A. Mayewski of the University of New Hampshire explains that a crucial piece of information is whether the changes in carbon dioxide concentrations precede or follow the onset of ice ages. In other words, climatologists cannot yet determine whether those changes are a symptom or a cause of the wholesale environmental changes that occur during ice ages. As Mark A. Chandler of the Goddard Institute for Space Studies wryly observes, "Watching what happens over the next 50 years will be a great experiment" for clarifying the influence of carbon dioxide on global temperatures.

Studies of ice cores are also uncovering evidence of surprisingly erratic behavior in the earth's climate—behavior that cannot all result from the action of carbon dioxide and other greenhouse gases. Researchers have been stunned by recent reports by Kendrick C. Taylor of the Desert Research Institute in Reno, Nev., and his colleagues that the temperatures recorded in the Greenland ice cores fluctuated rapidly during the last ice age, warming and cooling over the course of a decade or less. Just a few months ago Willi Dansgaard of the University of Copenhagen and his coworkers added to the excitement when they announced evidence that similar climate swings occurred during the last warm period. That controversial finding could indicate that global temperatures might take another violent swing during the current warm spell.

The short-term climate fluctuations "clearly result from changes in atmospheric circulation patterns," Mayewski reports. The mechanisms responsible for that altered circulation remain highly speculative. Mayewski cites variations in the brightness of the sun as a likely culprit. "People have shied away from the idea of solar variability because they lacked the proper long-range records," he says. The ongoing analysis of atmospheric gases, dust and other components trapped in the ice cores could settle the matter, he believes.

TIM FULLER

***INNOVATIVE APPARATUS** for measuring carbon dioxide in ice cores was developed by Alex T. Wilson* **(standing)** *and Austin Long of the University of Arizona.*

Mayewski hopes better insight into the inconstant nature of the sun will enable researchers to determine whether the present, human-generated increases in carbon dioxide are negating a natural global cooling or enhancing a global warming. Either way, he says, the findings "will not eradicate the importance of carbon dioxide."

Data from the various ice cores should eventually enable theorists to develop a comprehensive model of terrestrial climate for the past tens of thousands of years. The first step in that endeavor entails collecting accurate measurements of all the parameters that influence the global environment. The present dispute over carbon dioxide adds an unwelcome uncertainty to the effort. "The finger-pointing is part of the process. Ultimately we'll sort it all out, and we'll have a much stronger program," Taylor says cheerfully. Moments later, reflecting the mood of a field that has been progressing at breakneck speed, he adds, "It's just time to stop talking and start doing." *—Corey S. Powell*

Fermat's Theorem Fights Back

"Problems worthy of attack," quoth the physicist-poet Piet Hein, "prove their worth by hitting back." That is certainly the case with Fermat's Last Theorem, which after being apparently knocked out last summer has bounced off the mat for another round.

The deceptively simple theorem states that the equation $X^N + Y^N = Z^N$ has no positive, integral solutions for exponents greater than 2. Posed some 350 years ago by the French polymath Pierre de Fermat, who claimed in the margin of a book that he had found a proof but did not have room to write it down, it became perhaps the most famous problem in mathematics.

Last June, Andrew J. Wiles of Princeton University electrified his field by announcing that he had discovered a proof of the theorem. Based largely on Wiles's solid reputation and on his outline of an approach that had previously seemed promising, a number of leading lights declared the proof to be almost certainly correct. The finding was trumpeted on the front page of the *New York Times*—and favorably reported in the pages of this magazine.

Shortly after his announcement, Wiles submitted a 200-page manuscript to *Inventiones Mathematicae,* and the journal's editor, Barry Mazur of Harvard University, sent it to six reviewers. Wiles quickly fixed several minor problems identified by the reviewers, but one problem proved less tractable. In December, Wiles released a statement that he was working on a "calculation" that was "not yet complete." He reassured his audience, "I believe that I will be able to finish this in the near future."

Karl Rubin of Ohio State University, who as a reviewer is one of the few people who has actually read Wiles's manuscript, is optimistic that Wiles will succeed. But he concedes that only Wiles knows exactly where the proof stands, and since his December statement Wiles has remained incommunicado.

Indeed, his reticence, and his refusal to make his manuscript more widely available, has reportedly annoyed some colleagues. Kenneth A. Ribet of the University of California at Berkeley notes that it is customary for mathematicians, once they have submitted a manuscript to a journal, to disseminate it freely so that it can be "ripped apart in seminars." A proof by Ribet himself, which helped to convince Wiles to take on Fermat's theorem in 1986, was refined in this way. But pointing out that Wiles worked on his proof in virtual isolation for seven years before revealing it, Ribet suggests that Wiles "feels he has the right to finish it by himself."

In his December statement, Wiles said he would discuss the proof further at a graduate seminar beginning in February. But some observers are skeptical about just how revealing Wiles will be, given his penchant for caution and privacy. Wiles has said he would reveal details of his proof twice before—once at the end of the summer and again in November. Ronald L. Graham of AT&T Bell Laboratories speculates that even if Wiles does begin discussing his proof during his class, he might take months to arrive at the part now causing him trouble.

James Propp of the Massachusetts Institute of Technology thinks the Wiles affair raises an interesting "sociological" question: "When is a theorem deemed to be true?" Joseph J. Kohn, chairman of the Princeton mathematics department, espouses a true-until-proved-otherwise position toward Wiles's proof. Wiles should still have the benefit of the doubt, Kohn argues, because he has "an extraordinarily good track record."

Gerd Faltings of Princeton turns Kohn's argument on its head. The very fact that Wiles is so competent, Faltings points out, means that he must be facing an extremely difficult and perhaps insurmountable problem. "If it were easy, he would have solved it by now," says Faltings, whose work helped Wiles to construct his proof. "Strictly speaking," Faltings comments, Wiles's recent travails suggest that "it wasn't a proof when it was announced."

Alan Baker of the University of Cambridge agrees. He was one of the few prominent mathematicians openly to voice skepticism toward Wiles's proof from the start. According to one source, Baker even offered to bet 100 bottles of wine against a single bottle that within a year the proof would be shown to be invalid.

Baker denies that report, but he admits he did express a "healthy skepticism" toward the proof. After all, Fermat's theorem is notoriously difficult, and Wiles's proof drew on work that was less than a decade old and thus perhaps not thoroughly vetted. Baker, like Faltings, emphasizes that he hopes Wiles completes the proof, but he adds, "I think the prospects are lower now." *—John Horgan*

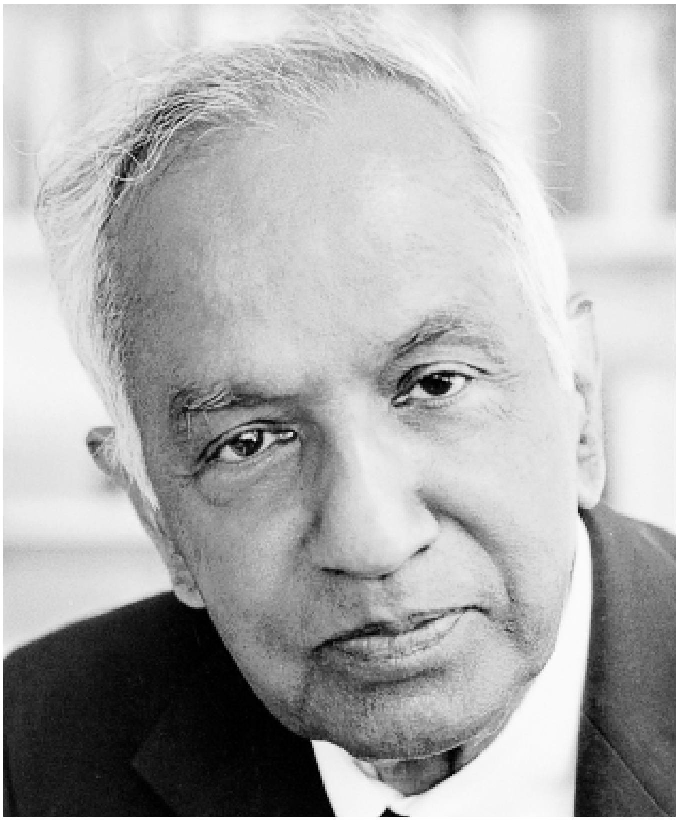
WOLFRAM RESEARCH

COMPLEX CURVE represents a set of nonintegral solutions to the equation $X^N + Y^N = Z^N$.

PROFILE: SUBRAHMANYAN CHANDRASEKHAR

Confronting the Final Limit

Clad in a dark, classically tailored suit and black shoes, Subrahmanyan Chandrasekhar approaches with a slow but fluid gait. He shakes my hand firmly, unsmiling; he has no need to ingratiate. Easing his lean frame into a chair, he slouches sideways and cocks his head, as if from this oblique angle his obsidian eyes can bore in on me better. What, precisely, do I want to talk about? he inquires. His voice still bears an Indian lilt, although he came here to the University of Chicago more than half a century ago.


COURTESY OF UNIVERSITY OF CHICAGO

CHANDRASEKHAR calls Newton's **Principia**, *which he has been studying, an achievement with "no parallel in science at any time."*

I reply that I am interested in all aspects of his career, including his demonstration in the 1930s that stars above a certain mass—now known as the Chandrasekhar limit—undergo a catastrophic collapse. The finding, for which Chandrasekhar received, belatedly, the 1983 Nobel Prize, remains a cornerstone of modern astrophysics. I am also eager to hear his views on his latest object of study, Isaac Newton's *Philosophiae Naturalis Principia Mathematica* (*Mathematical Principles of Natural Philosophy*), the opus that laid the foundation for modern science.

Chandrasekhar says he is completing a book on the *Principia,* and he is not sure he wants to preview it. I assure him that since my article will be only two pages long, it cannot discuss the *Principia* in detail. His eyes grow darker still. "You think you can summarize Homer's *Odyssey* in two pages?" he snaps, jabbing first one, then both, impossibly long forefingers at me. "You think you can write about the Sistine Chapel in two pages?" His voice quavers with incredulity, disgust. "If you write only two pages, I don't think it matters very much if you talk to me."

Somehow the interview lurches forward, and Chandrasekhar, whom friends call Chandra, slips into the charming persona that colleagues had described. He dispenses jokes, anecdotes and aphorisms, as well as smiles and laughter, generously. But in that moment of anger, he has revealed the incompressible passion—not only for scientific truth but for beauty, which in Chandrasekhar's mind are fused—at his core. It is this quality that helped Chandrasekhar overcome an enormous blow early in his career to become one of the world's most distinguished and productive physicists.

The trait may also explain why Chandrasekhar, who at 83 is still legendary for his work habits, exudes a certain restlessness. In *Chandra,* a biography published in 1991, the physicist Kameshwar C. Wali suggests that a clue to Chandrasekhar's character can be found in a striking photograph hanging in his office. It shows a man climbing a ladder that leans against some vast, abstract structure. Like the ascending man, Wali says, Chandrasekhar is "constantly aware of how much more there is to know" and of his own inadequacies.

Chandrasekhar was nurtured on ambition. His mother, in addition to raising 10 children, found time for such pursuits as translating Henrik Ibsen's *A Doll's House* into Tamil. His father was a government official whose younger brother, the physicist C. V. Raman, received the 1930 Nobel Prize. Not surprisingly, then, Chandrasekhar became a star student of physics and mathematics at the Presidency College in Madras.

In 1930 he left India for the University of Cambridge, and since then he has returned to his native land only for visits. Chandrasekhar admits he sometimes wonders how his career would have unfolded had he remained in India. Like Raman, his uncle, he might someday have presided over his own institute, but he then would have become enmeshed in the arcane politics of India's scientific establishment. "I have one advantage here" in the U.S., Chandrasekhar says. "I have enormous freedom. I can do what I want. Nobody bothers me."

At Cambridge, Chandrasekhar began applying his already broad knowledge of quantum mechanics and relativity to the question of how stars evolve. Among his mentors was Sir Arthur Eddington, whose influential text on astrophysics had lured Chandrasekhar to that subject. Chandrasekhar's theoretical forays soon led him to an unsettling conclusion. Most astronomers believed that when stars exhausted their store of nuclear fuel, they settled into interminable old age as small, dense white dwarfs. Chandrasekhar's calculations revealed that in stars whose masses were more than 1.4 times that of the sun, gravity would overcome the outward, repulsive pressure of electrons and trigger a collapse into states of matter even denser than that of white dwarfs.

Astronomers eventually unraveled the

strange destinies of stars whose masses transcend the Chandrasekhar limit: after erupting into supernovae, their cores implode into spheres of compacted neutrons called neutron stars (one cup of which outweighs Mount Everest) or into infinitely dense black holes. But acceptance of Chandrasekhar's insight was slow in coming. The reason was that in 1935, immediately after the 24-year-old Chandrasekhar presented his theory before the Royal Astronomical Society, Eddington himself stood to ridicule it as self-evidently wrong, an example of reductio ad absurdum. Eddington had previously given his protégé no inkling of his views.

Chandrasekhar insists that at the time he harbored no ill feelings toward Eddington; they even remained friends. Eddington's repudiation of Chandrasekhar's theory nonetheless played a role in his decision in 1937 to leave England for the University of Chicago, where he has remained. He also left behind the subject of collapsing stars, but not before he had written a book. "I simply decided, well, I will write a book and present my idea, leave the subject and go on to other things. And that's all happened, you see."

Although brought on by trauma, this pattern—total immersion in a subject followed by an abrupt swerve toward "other things"—was to become characteristic of Chandrasekhar. After his stellar evolution phase, he spent five years considering the motion of stars within a galaxy, demonstrating that stars exert a kind of friction on one another through their gravitational interactions. From 1943 through 1950 he contemplated the transfer of radiation within stellar and planetary atmospheres. Then came periods devoted to the properties of fluids and magnetic fields and to ellipsoids, geometric objects whose properties have proved useful for understanding galaxies. Between 1974 and 1983 he explored black holes, coming back full circle, in a sense, to the work that had launched his career.

The books that Chandrasekhar wrote at the close of each period were instant classics, praised for their breadth and clarity. Chandrasekhar says he has always sought to present his findings in as elegant, even literary, a form as possible. "I select some writers in order to learn," he confides. "For example, I read Henry James or Virginia Woolf, and I don't simply read the text as a novel; I see how they construct sentences, how they construct paragraphs, how one paragraph goes into another and so on."

Too few scientists write well or even carefully, according to Chandrasekhar: "You take any volume of the *Astrophysical Journal* or the *Physical Review,* turn to the middle of it, put your hand on a paragraph. You are sure to find a mistake, either in style or grammar or something." Chandrasekhar sought to encourage good writing during the 20 years he served as editor of the *Astrophysical Journal,* the premier publication of his field. "I will tell you a malicious statement I used to make" to authors, he remarks, grinning. "I would say, 'Your paper is scientifically correct, but I wish you would ask your colleague in the English department to read it.' "

Chandrasekhar's latest epoch began when he was invited to contribute a paper to a meeting held in 1987 to celebrate the 300th birthday of the *Principia.* Chandrasekhar had long hoped to delve into the *Principia;* he bought an English translation of the book (which Newton wrote in Latin) decades ago. But he had always been too busy staking out his own territory—and, he now believes, too intellectually immature for serious study of the difficult work. He notes that in order to understand Newton's somewhat "secretive" and elliptical style, "you must read line by line."

He decided early on that rather than assessing Newton secondhand, through commentaries, he would absorb the *Principia* unmediated. More specifically, he would read a proposition and then, before going on to Newton's proof, would try to derive his own. Chandrasekhar points out that although he has 300 extra years of knowledge at his disposal, in virtually every case his proofs fell short of Newton's.

Reading Newton became for Chandrasekhar a sustained epiphany. "The view of science that he exhibits, the clarity with which he writes, the number of new things he finds, manifest a physical and mathematical insight of which there is no parallel in science at any time." It is common knowledge that Newton invented calculus as well as seminal theories of gravity and optics. But Chandrasekhar argues that the *Principia* contains other achievements that have been overlooked. For example, Newton set forth a theory of gyroscopes, which were not invented for another 200 years. He was the first scientist to note that knowledge of the initial conditions of a system should provide one with knowledge of its entire future, an insight usually credited to Laplace. He invented a theory of image formation generally ascribed to Lord Kelvin.

Chandrasekhar is as entranced by the style of the *Principia* as he is by its substance. He compares Newton's prose to that of Henry James, who was similarly fond of long, complex sentences. To demonstrate his point, Chandrasekhar fetches his massive, black copy of the *Principia* and reads: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes." Chandrasekhar looks up and exclaims, his voice cracking, "Isn't that a beautiful sentence? Absolutely!"

Chandrasekhar likens reading Newton to what were for him equally awe-evoking experiences: gazing at the ceiling of the Sistine Chapel, watching Sir John Gielgud play Hamlet or hearing Arturo Toscanini conduct Beethoven's Ninth Symphony. Indeed, as great as Newton's reputation is, it is not great enough to satisfy Chandrasekhar. "Newton is not one of the two or three greatest scientists. He is one of the two or three greatest intellects, ever, in any subject. If you want to compare Newton to anybody, you have to go outside science."

Chandrasekhar has already sent more than 20 chapters of his planned 30-chapter book to his publisher, and he hopes to complete it this spring. Has he given thought to some new project beyond that? "No, that's the end," he says abruptly. "I don't expect to do science after I finish work on the *Principia.*" When I express surprise that someone who has been so consistently productive could simply cease working, he says heatedly, "Obviously I can go on doing work of a quality that is below my standards, but why do that? So the time must come when I say, 'Stop.' "

I am reminded of an essay, published in *Nature* in 1990, in which Chandrasekhar describes the creative life as a constant striving against "one's inherent and often insurmountable limitations." He concludes the essay with lines from a poem by T. S. Eliot: "It is strange, isn't it/That a man should have a consuming passion/To do something for which he lacks the capacity?"

Yet there are consolations, even for a seeker past his prime. Chandrasekhar recollects that G. H. Hardy, in his classic memoir *A Mathematician's Apology,* called an old mathematician whose ideas have run dry "a pathetic person." Hardy consoled himself, particularly when forced to endure boring, second-rate colleagues, with the knowledge that he had once communed with some of the greatest intellects of his age. Chandrasekhar confesses that he has cultivated a similar habit when he finds himself in "tiresome" situations: "I think to myself, 'I have been in the company of Newton.' " *—John Horgan*

SCIENTIFIC
AMERICAN

Can the Growing Human Population Feed Itself?

As human numbers surge toward 10 billion, some experts are alarmed, others optimistic. Who is right?

by John Bongaarts

Demographers now project that the world's population will double during the next half century, from 5.3 billion people in 1990 to more than 10 billion by 2050. How will the environment and humanity respond to this unprecedented growth? Expert opinion divides into two camps. Environmentalists and ecologists, whose views have widely been disseminated by the electronic and print media, regard the situation as a catastrophe in the making. They argue that in order to feed the growing population farmers must intensify agricultural practices that already cause grave ecological damage. Our natural resources and the environment, now burdened by past population growth, will simply collapse under the weight of this future demand.

The optimists, on the other hand, comprising many economists as well as some agricultural scientists, assert that the earth can readily produce more than enough food for the expected population in 2050. They contend that technological innovation and the continued investment of human capital will deliver high standards of living to much of the globe, even if the population grows much larger than the projected 10 billion. Which point of view will hold sway? What shape might the future of our species and the environment actually take?

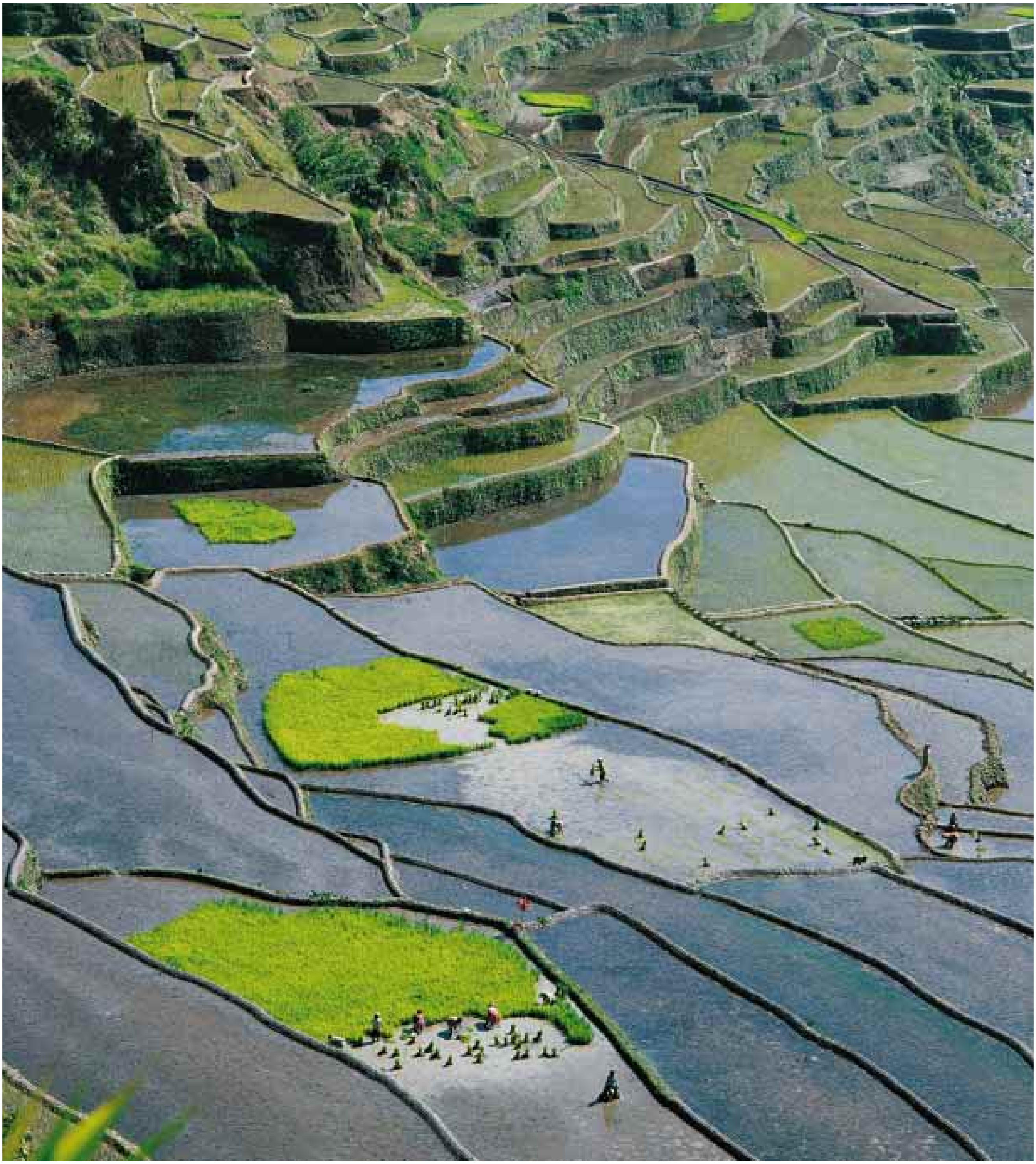
Many environmentalists fear that world food supply has reached a precarious state: "Human numbers are on a collision course with massive famines.... If humanity fails to act, nature will end the population explosion for us—in very unpleasant ways—well before 10 billion is reached," write Paul R. Ehrlich and Anne H. Ehrlich of Stanford University in their 1990 book *The Population Explosion.* In the long run, the Ehrlichs and like-minded experts consider substantial growth in food production to be absolutely impossible. "We are feeding ourselves at the expense of our children. By definition farmers can overplow and overpump only in the short run. For many farmers the short run is drawing to a close," states Lester R. Brown, president of the Worldwatch Institute, in a 1988 paper.

Over the past three decades, these authors point out, enormous efforts and resources have been pooled to amplify agricultural output. Indeed, the total quantity of harvested crops increased dramatically during this time. In the developing world, food production rose by an average of 117 percent in the quarter of a century between 1965 and 1990. Asia performed far better than other regions, which saw increases below average.

Because population has expanded rapidly as well, per capita food production has generally shown only modest change; in Africa it actually declined. As a consequence, the number of undernourished people is still rising in most parts of the developing world, although that number did fall from 844 million to 786 million during the 1980s. But this decline reflects improved nutritional conditions in Asia alone. During the same period, the number of people having energy-deficient diets in Latin America, the Near East and Africa climbed.

Many social factors can bring about conditions of hunger, but the pessimists emphasize that population pressure on fragile ecosystems plays a significant role. One specific concern is that we seem to be running short on land suitable for cultivation. If so, current efforts to bolster per capita food production by clearing more fertile land will find fewer options. Between 1850 and 1950 the amount of arable land grew quickly to accommodate both larger populations and greater demand for better diets. This expansion then slowed and by the late 1980s ceased altogether. In the developed world, as well as in some developing countries (especially China), the amount of land under cultivation started to decline during the

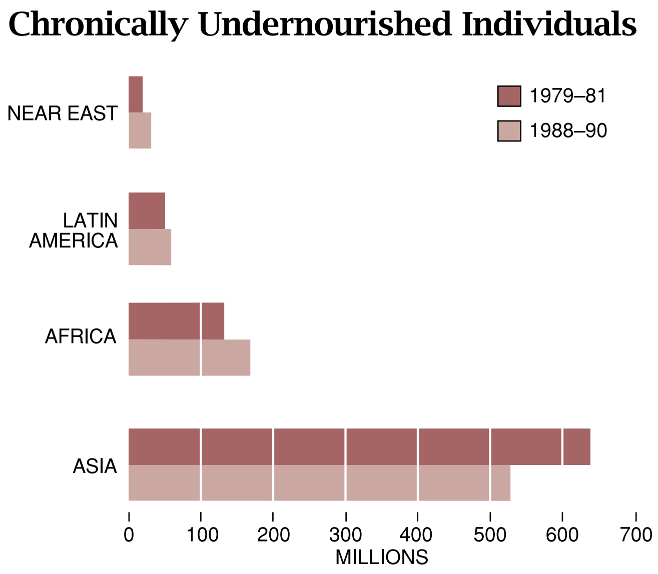
JOHN BONGAARTS has been vice president and director of the Research Division of the Population Council in New York City since 1989. He is currently a member of the Johns Hopkins Society of Scholars and the Royal Dutch Academy of Sciences. He won the Mindel Sheps Award in 1986 from the Population Association of America and the Research Career Development Award in 1980–85 from the National Institutes of Health.

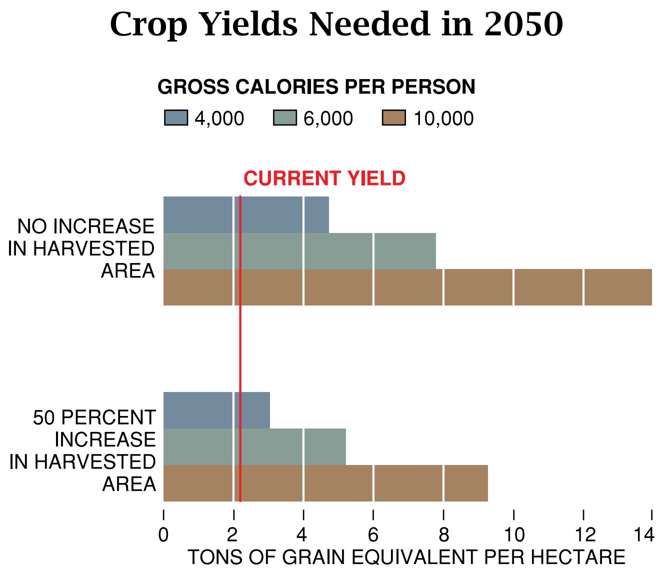


RICE PADDIES (these are in Indonesia) provide the principal food for more than half the world's population. In many parts of Asia the terrain prevents farmers from using mechanized farm equipment; to grow and harvest a single acre of rice can demand more than 1,000 man-hours. Still, Asian countries now produce more than 90 percent of all rice grown.

1980s. This drop is largely because spreading urban centers have engulfed fertile land or, once the land is depleted, farmers have abandoned it. Farmers have also fled from irrigated land that has become unproductive because of salt accumulation.

Moreover, environmentalists insist that soil erosion is destroying much of the land that is left. The extent of the damage is the subject of controversy. A recent global assessment, sponsored by the United Nations Environment Program and reported by the World Resources Institute and others, offers some perspective. The study concludes that 17 percent of the land supporting plant life worldwide has lost value over the past 45 years. The estimate includes erosion caused by water and wind, as well as chemical and physical deterioration, and ranks the degree of soil degradation from light to severe. This degradation is least prevalent in North





INCIDENCE OF CHRONIC UNDERNUTRITION fell in the developing world from an estimated 844 million sufferers in 1979 to 786 million in 1990, showing evidence of dramatic nutritional improvements in Asia (*left*). Agricultural productivity must improve to continue this trend (*right*). Even if more land is harvested in 2050, the average yield must rise sharply as well to offer the projected Third World population of 8.7 billion the current diet of 4,000 gross calories per day.

America (5.3 percent) and most widespread in Central America (25 percent), Europe (23 percent), Africa (22 percent) and Asia (20 percent). In most of these regions, the average farmer could not gather the resources necessary to restore moderate and severely affected soil regions to full productivity. Therefore, prospects for reversing the effects of soil erosion are not good, and it is likely that this problem will worsen.

Despite the loss and degradation of fertile land, the “green revolution” has promoted per capita food production by increasing the yield per hectare. The new, high-yielding strains of grains such as wheat and rice have proliferated since their introduction in the 1960s, especially in Asia. To reap full advantage from these new crop varieties, however, farmers must apply abundant quantities of fertilizer and water.

Environmentalists question whether further conversion to such crops can be achieved at reasonable cost, especially in the developing world, where the gain in production is most needed. At the moment, farmers in Asia, Latin America and Africa use fertilizer sparingly, if at all, because it is too expensive or unavailable. Fertilizer use in the developed world has recently waned. The reasons for the decline are complex and may be temporary, but clearly farmers in North America and Europe have decided that increasing their already heavy application of fertilizer will not further enhance crop yields.

Unfortunately, irrigation systems, which would enable many developing countries to join in the green revolution, are often too expensive to build. In most areas, irrigation is essential for generating higher yields. It also can make arid land cultivable and protect farmers from the vulnerability inherent in natural variations in the weather. Land brought into cultivation this way could be used for growing multiple crop varieties, thereby helping food production to increase.

Such advantages have been realized since the beginning of agriculture: the earliest irrigation systems are thousands of years old. Yet only a fraction of productive land in the developing world is now irrigated, and its expansion has been slower than population growth. Consequently, the amount of irrigated land per capita has been dwindling during recent decades. The trend, pessimists argue, will be hard to stop. Irrigation systems have been built in the most affordable sites, and the hope for extending them is curtailed by rising costs. Moreover, the accretion of silt in dams and reservoirs and of salt in already irrigated soil is increasingly costly to avoid or reverse.

Environmentalists Ehrlich and Ehrlich note that modern agriculture is by nature at risk wherever it is practiced. The genetic uniformity of single, high-yielding crop strains planted over large areas makes them highly productive but also renders them particularly vulnerable to insects and disease. Current preventive tactics, such as spraying pesticides and rotating crops, are only partial solutions. Rapidly evolving pathogens pose a continuous challenge. Plant breeders must maintain a broad genetic arsenal of crops by collecting and storing natural varieties and by breeding new ones in the laboratory.

The optimists do not deny that many problems exist within the food supply system. But many of these authorities, including D. Gale Johnson, the late Herman Kahn, Walter R. Brown, L. Martel, the late Roger Revelle, Vaclav Smil and Julian L. Simon, believe the world’s food supply can dramatically be expanded. Ironically, they draw their enthusiasm from extrapolation of the very trends that so alarm those experts who expect doom. In fact, statistics show that the average daily caloric intake per capita climbed by 21 percent (from 2,063 calories to 2,495 calories) between 1965 and 1990 in the developing countries. These higher calories have generally delivered greater amounts of protein. On average, the per capita consumption of protein rose from 52 grams per day to 61 grams per day between 1965 and 1990.

According to the optimists, not only has the world food situation improved significantly in recent decades, but further growth can be brought about in various ways. A detailed assessment of climate and soil conditions in 93 developing countries (excluding China) shows that nearly three times as much land as is currently farmed, or an additional 2.1 billion hectares, could be cultivated. Regional soil estimates indicate that sub-Saharan Africa and Latin America can exploit many more stretches of unused land than can Asia, the Near East and North Africa.

Even in regions where the amount of potentially arable land is limited, crops could be grown more times every year than is currently the case. This scenario is particularly true in the tropics and subtropics where conditions are such—relatively even temperature throughout the year and a consistent distribution of daylight hours—that more than one crop would thrive. Nearly twice as many crops are harvested every year in Asia than in Africa at present, but further increases are possible in all regions.

In addition to multicropping, higher yields per crop are attainable, especially in Africa and the Near East. Many more crops are currently harvested per hectare in the First World than elsewhere: cereal yields in North America and Europe averaged 4.2 tons per hectare, compared with 2.9 in the Far East (4.2 in China), 2.1 in Latin America, 1.7 in the Near East and only 1.0 in Africa.

Such yield improvements, the enthusiasts note, can be achieved by expanding the still limited use of high-yield crop varieties, fertilizer and irrigation.

The Potential Impact of Global Warming on Agriculture

The scientific evidence on the greenhouse effect indicates that slow but significant global warming is likely to occur if the emission of greenhouse gases, such as carbon dioxide, methane, nitrogen oxide and chlorofluorocarbons, continues to grow. Agriculture is directly or, at least in some cases, indirectly responsible for releasing a substantial proportion of these gases. Policy responses to the potentially adverse consequences of global climatic change now focus primarily on hindering emissions rather than on halting them. But considering the present need to improve living standards and produce more food for vast numbers of people, experts doubt that even a reduction in global emissions could occur in the near future.

In a 1990 study the Intergovernmental Panel on Climate Change estimated that over the next century the average global temperature will rise by three degrees Celsius. The study assumes that agriculture will expand considerably. This forecast of temperature change is uncertain, but there is now broad agreement that some global warming will take place. All the same, the effect that temperature rise will have on human society remains an open question.

Global warming could either enhance or impede agriculture, suggest Cynthia Rosenzweig of Columbia University and Martin L. Parry of the University of Oxford. Given sufficient water and light, increased ambient carbon dioxide concentrations absorbed during photosynthesis could act as a fertilizer and facilitate growth in certain plants. In addition, by extending the time between the last frost in the spring and the first frost in the fall, global warming will benefit agriculture in cold regions where the growing season is short, such as in Canada and northern areas of Europe and the former Soviet Union. Moreover, warmer air holds more water vapor, and so global warming will bring about more evaporation and precipitation. Areas where crop production is limited by arid conditions would benefit from a wetter climate.

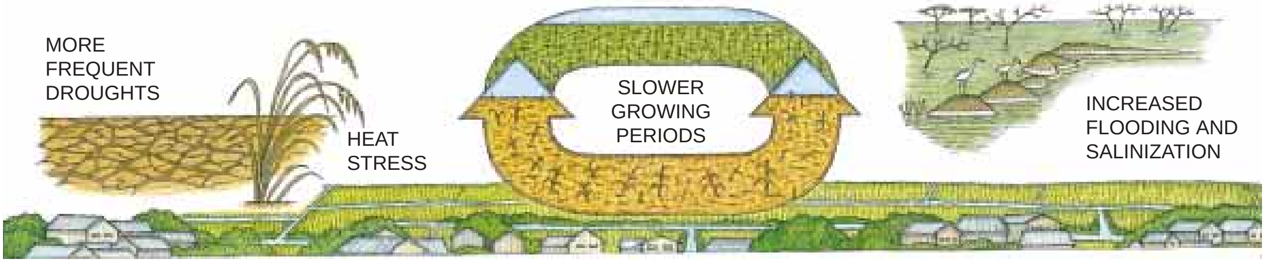
If increased evaporation from soil and plants does not coincide with more rainfall in a region, however, more frequent dry spells and droughts would occur. And a further rise in temperature will reduce crop yields in tropical and subtropical areas, where certain crops are already grown near their limit of heat tolerance. Furthermore, some cereal crops need low winter temperatures to initiate flowering. Warmer winters in temperate regions could therefore stall growing periods and lead to reduced harvests. Finally, global warming will precipitate a thermal swelling of the oceans and melt polar ice. Higher sea levels may claim low-lying farmland and cause higher salt concentrations in the coastal groundwater.

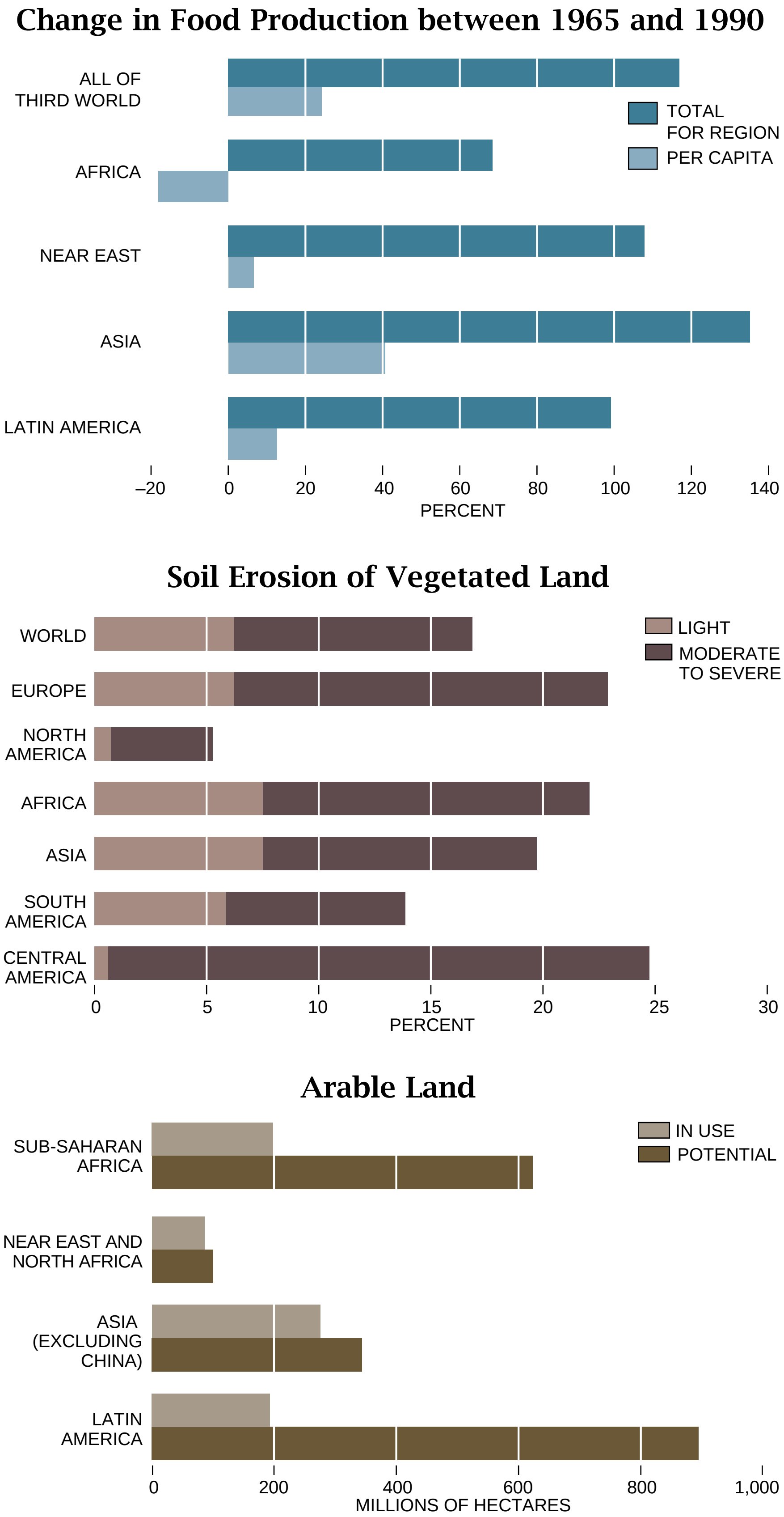
Techniques used to model the climate are not sufficiently advanced to predict the balance of these effects in specific areas. The most recent analysis on the impact of climatic change on the world food supply, by Rosenzweig and Parry in 1992, concludes that average global food production will decline 5 percent by 2060. And they anticipate a somewhat larger drop in the developing world, thus exacerbating the problems expected to arise in attempts to feed growing populations. In contrast, their report predicts a slight rise in agricultural output in developed countries situated at middle and high latitudes.

POSSIBLE BENEFITS OF GLOBAL WARMING ON AGRICULTURE

CO_2
CARBON DIOXIDE FERTILIZATION
LONGER GROWING SEASONS
INCREASED PRECIPITATION

POSSIBLE DRAWBACKS OF GLOBAL WARMING ON AGRICULTURE





TOTAL FOOD PRODUCTION rose nearly 120 percent between 1965 and 1990 in the developing world. Per capita food production showed little change in regions outside Asia (*top*). Soil erosion has debased much of the land worldwide on which that food was produced (*middle*). But many Third World nations have vast holdings that could be farmed successfully if given more water and fertilizer (*bottom*).

In *World Agriculture: Toward 2000,* Nikos Alexandratos of the Food and Agriculture Organization (FAO) of the United Nations reports that only 34 percent of all seeds planted during the mid-1980s were high-yielding varieties. Statistics from the FAO show that at present only about one in five hectares of arable land is irrigated, and very little fertilizer is used. Pesticides are sparsely applied. Food output could drastically be increased simply by more widespread implementation of such technologies.

Aside from producing more food, many economists and agriculturalists point out, consumption levels in the developing world could be boosted by wasting fewer crops, as well as by cutting storage and distribution losses. How much of an increase would these measures yield? Robert W. Kates, director of the Alan Shawn Feinstein World Hunger Program at Brown University, writes in *The Hunger Report: 1988* that humans consume only 60 percent of all harvested crops, and some 25 to 30 percent is lost before reaching individual homes. The FAO, on the other hand, estimates lower distribution losses: 6 percent for cereals, 11 percent for roots and 5 percent for pulses. All the same, there is no doubt that improved storage and distribution systems would leave more food available for human nutrition, independent of future food production capabilities.

For optimists, the long-range trend in food prices constitutes the most convincing evidence for the correctness of their view. In 1992–93 the World Resources Institute reported that food prices dropped further than the price of most nonfuel commodities, all of which have declined in the past decade. Cereal prices in the international market fell by approximately one third between 1980 and 1989. Huge government subsidies for agriculture in North America and western Europe, and the resulting surpluses of agricultural products, have depressed prices. Obviously, the optimists assert, the supply already exceeds the demand of a global population that has doubled since 1950.

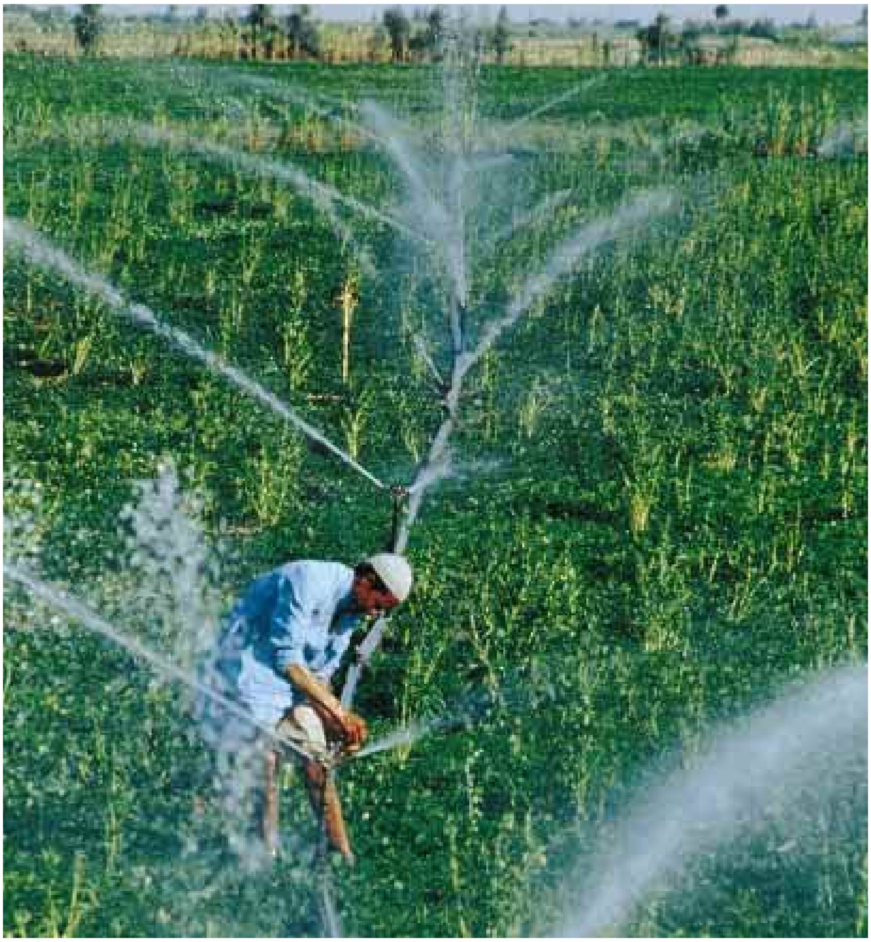
Taken together, this evidence leads many experts to see no significant obstacles to raising levels of nutrition for world populations exceeding 10 billion people. The potential for an enormous expansion of food production exists, but its realization depends of course on sensible governmental policies, increased domestic and international trade and large investments in infrastructure and agricultural extension. Such improvements can be achieved, the optimists believe, without incurring irreparable damage to global ecosystems.

Proponents of either of these conflicting perspectives have difficulty accepting the existence of other plausible points of view. Moreover, the polarity between the two sides of expert opinion shows that neither group can be completely correct. Finding some common ground between these seemingly irreconcilable positions is not as difficult as it at first appears if empirical issues are emphasized and important differences in value systems and political beliefs are ignored.

Both sides agree that the demand for food will swell rapidly over the next several decades. In 1990 a person living in the developing world ate on average 2,500 calories each day, taken from 4,000 gross calories of food crops made available within a household. The remaining 1,500 calories from this gross total not used to meet nutritional requirements were either lost, inedible or used as animal feed and plant seed. Most of this food was harvested from 0.7 billion hectares of land in the developing world. The remaining 5 percent of the total food supply came from imports. To sustain this 4,000-gross-calorie diet for more than twice as many residents, or 8.7 billion people, living in the developing world by 2050, agriculture must offer 112 percent more crops. To raise the average Third World diet to 6,000 gross calories per day, slightly above the 1990 world average, food production would need to increase by 218 percent. And to bring the average Third World diet to a level comparable with that currently found in the developed world, or 10,000 gross calories per day, food production would have to surge by 430 percent.

A more generous food supply will be achieved in the future through boosting crop yields, as it has been accomplished in the past. If the harvested area in the developing world remains at 0.7 billion hectares, then each hectare must more than double its yield to maintain an already inadequate diet for the future population of the developing world. Providing a diet equivalent to a First World diet in 1990 would require that each hectare increase its yield more than six times. Such an event in the developing world must be considered virtually impossible, barring a major breakthrough in the biotechnology of food production.

Instead farmers will no doubt plant more acres and grow more crops per year on the same land to help augment crop harvests. Extrapolation of past trends suggests that the total harvested area will increase by about 50 percent by the year 2050. Each hectare will then have to provide nearly 50 percent more tons of grain or its equivalent to keep up with current dietary levels. Improved diets could result only from much larger yields.



EGYPTIAN FARMERS, advised by Israeli agronomists, have converted more than 400,000 acres of desert soil into rich cropland by implementing irrigation systems. Farms in Nubariya now produce ample harvests of fruit.

The technological optimists are correct in stating that overall world food production can substantially be increased over the next few decades. Current crop yields are well below their theoretical maxima, and only about 11 percent of the world's farmable land is now under cultivation. Moreover, the experience gained recently in a number of developing countries, such as China, holds important lessons on how to tap this potential elsewhere. Agricultural productivity responds to well-designed policies that assist farmers by supplying needed fertilizer and other inputs, building sound infrastructure and providing market access. Further investments in agricultural research will spawn new technologies that will fortify agriculture in the future. The vital question then is not how to grow more food but rather how to implement agricultural methods that may make possible a boost in food production.

A more troublesome problem is how to achieve this technological enhancement at acceptable environmental costs. It is here that the arguments of those experts who forecast a catastrophe carry considerable weight. There can be no doubt that the land now used for growing food crops is generally of better quality than unused, potentially cultivable land. Similarly, existing irrigation systems have been built on the most favorable sites. Consequently, each new measure applied to increase yields is becoming more expensive to implement, especially in the developed world and parts of the developing world such as China, where productivity is already high. In short, such constraints are raising the marginal cost of each additional ton of grain or its equivalent. This tax is even higher if one takes into account negative externalities—primarily environmental costs not reflected in the price of agricultural products.

The environmental price of what in the Ehrlichs' view amounts to "turning the earth into a giant human feedlot"

DASHBOARD COMPUTER on a tractor, carrying maps compiled via satellite, can now guide farmers in performing soil analysis and applying site-specific amounts and blends of fertilizer. Such technology saves money and increases efficiency.

could be severe. A large inflation of agriculture to provide growing populations with improved diets is likely to lead to widespread deforestation, loss of species, soil erosion and pollution from pesticides, and runoff of fertilizer as farming intensifies and new land is brought into production. Reducing or minimizing this environmental impact is possible but costly.

Given so many uncertainties, the course of future food prices is difficult to chart. At the very least, the rising marginal cost of food production will engender steeper prices on the international market than would be the case if there were no environmental constraints. Whether these higher costs can offset the historical decline in food prices remains to be seen. An upward trend in the price of food sometime in the near future is a distinct possibility. Such a hike will be mitigated by the continued development and application of new technology and by the likely recovery of agricultural production and exports in the former Soviet Union, eastern Europe and Latin America. Also, any future price increases could be lessened by taking advantage of the underutilized agricultural resources in North America, notes Per Pinstrup-Andersen of Cornell University in his 1992 paper “Global Perspectives for Food Production and Consumption.” Rising prices will have little effect on high-income countries or on households possessing reasonable purchasing power, but the poor will suffer.

In reality, the future of global food production is neither as grim as the pessimists believe nor as rosy as the optimists claim. The most plausible outcome is that dietary intake will creep higher in most regions. Significant annual fluctuations in food availability and prices are, of course, likely; a variety of factors, including the weather, trade interruptions and the vulnerability of monocropping to pests, can alter food supply anywhere. The expansion of agriculture will be achieved by boosting crop yields and by using existing farmland more intensively, as well as by bringing arable land into cultivation where such action proves economical. Such events will transpire more slowly than in the past, however, because of environmental constraints. In addition, the demand for food in the developed world is approaching saturation levels. In the U.S., mounting concerns about health have caused the per capita consumption of calories from animal products to drop.

Still, progress will be far from uniform. Numerous countries will struggle to overcome unsatisfactory nutrition levels. These countries fall into three main categories. Some low-income countries have little or no reserves of fertile land or water. The absence of agricultural resources is in itself not an insurmountable problem, as is demonstrated by regions, such as Hong Kong and Kuwait, that can purchase their food on the international market. But many poor countries, such as Bangladesh, cannot afford to buy food from abroad and thereby compensate for insufficient natural resources. These countries will probably rely more on food aid in the future.

Low nutrition levels are also found in many countries, such as Zaire, that do possess large reserves of potentially cultivable land and water. Government neglect of agriculture and policy failures have typically caused poor diets in such countries. A recent World Bank report describes the damaging effects of direct and indirect taxation of agriculture, controls placed on prices and market access, and overvalued currencies, which discourage exports and encourage imports. Where agricultural production has suffered from misguided government intervention (as is particularly the case in Africa), the solution—policy reform—is clear.

Food aid will be needed as well in areas rife with political instability and civil strife. The most devastating famines of the past decade, known to television viewers around the world, have occurred in regions fighting prolonged civil wars, such as Ethiopia, Somalia and the Sudan. In many of these cases, drought was instrumental in stirring social and political disruption. The addition of violent conflict prevented the recuperation of agriculture and the distribution of food, thus turning bad but remediable situations into disasters. International military intervention, as in Somalia, provides only a short-term remedy. In the absence of sweeping political compromise, hunger and malnutrition will remain endemic in these war-torn regions.

Feeding a growing world population a diet that improves over time in quality and quantity is technologically feasible. But the economic and environmental costs incurred through bolstering food production may well prove too great for many poor countries. The course of events will depend crucially on their governments’ ability to design and enforce effective policies that address the challenges posed by mounting human numbers, rising poverty and environmental degradation. Whatever the outcome, the task ahead will be made more difficult if population growth rates cannot be reduced.

FURTHER READING

POVERTY AND HUNGER: ISSUES AND OPTIONS FOR FOOD SECURITY IN DEVELOPING COUNTRIES. World Bank, 1986.

ENERGY, FOOD, ENVIRONMENT: REALITIES, MYTHS, OPTIONS. Vaclav Smil. Clarendon Press, 1987.

WORLD AGRICULTURE: TOWARD 2000. Nikos Alexandratos. New York University Press, 1988.

WORLD RESOURCES 1992–93. World Resources Institute. Oxford University Press, 1992.

The Earth's Mantle below the Oceans

Samples collected from the ocean floor reveal how the mantle's convective forces shape the earth's surface, create its crust and perhaps even affect its rotation

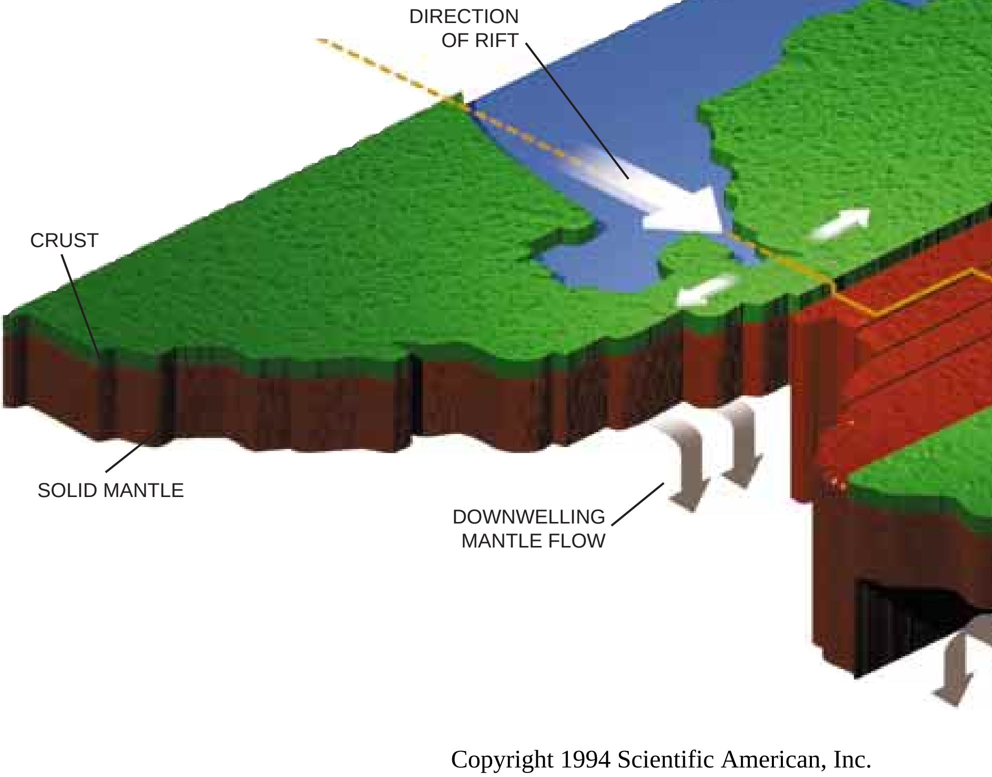
by Enrico Bonatti

Looking at a globe, one can easily imagine the continents and oceans as eternal, unchanging aspects of the earth's surface. Geophysicists now know that the appearance of permanence is an illusion caused by the brevity of the human life span. Over millions of years, blocks of the earth's rigid outer layer, the lithosphere, move about, diverging at midocean ridges, sliding about at faults and colliding at the margins of some of the oceans. Those motions cause continental drift and determine the global distribution of earthquakes and volcanoes.

Although the theory of plate tectonics is well established, the engine that drives the motion of the lithospheric plates continues to defy easy analysis because it is so utterly hidden from view. To confront that difficulty, several investigators and I have focused our research on the midocean ridges. The ridges are major, striking locations where the ocean floor is ripping apart. Examination of the composition, topography and seismic structure of the region along the midocean ridges is yielding results that often run contrary to conventional expectations. More complicated and fascinating than anyone had anticipated, the chemical and thermal processes in the mantle below midocean ridges dictate how new oceanic crust forms. Mantle activity may also cause different types of islands to emerge in the middle of oceans and some deep trenches to form at their edges. In fact, these processes may be so potent that they may even subtly affect the rotation of the planet.

The idea that the earth incorporates a dynamic interior may actually have its roots in the 17th century. René Descartes, the great French philosopher, made one of the first attempts to speculate scientifically about the earth's interior. In his 1644 treatise *Principles of Philosophy,* Descartes wrote that the earth had a central nucleus made of a primordial, sunlike fluid surrounded by a solid, opaque layer. Succeeding concentric layers of rock, metal, water and air made up the rest of the planet.

Geophysicists still subscribe to the notion of a layered earth, although their thinking has evolved considerably since the time of Descartes. In the current view, the earth possesses a solid inner core and a molten outer core. Both consist of iron-rich alloys. The earth's composition changes abruptly about 2,900 kilometers below the surface, where the core gives way to a mantle made of solid magnesium-iron silicate minerals. Another significant discontinuity, locat-



ENRICO BONATTI holds degrees in geology from the University of Pisa and the Scuola Normale Superiore in Pisa. After coming to the U.S. in 1959, he spent several years as a research scientist in petrology and marine geology at the University of California's Scripps Institution of Oceanography and as a professor at the University of Miami's Rosenstiel School of Marine Sciences. Since 1975 he has been with Columbia University's Lamont-Doherty Earth Observatory. Recently he has been teaching and researching in his native country. He has led or participated in expeditions in all the major oceans and in some remote but geologically intriguing lands, most recently in the polar Ural region of Russia.

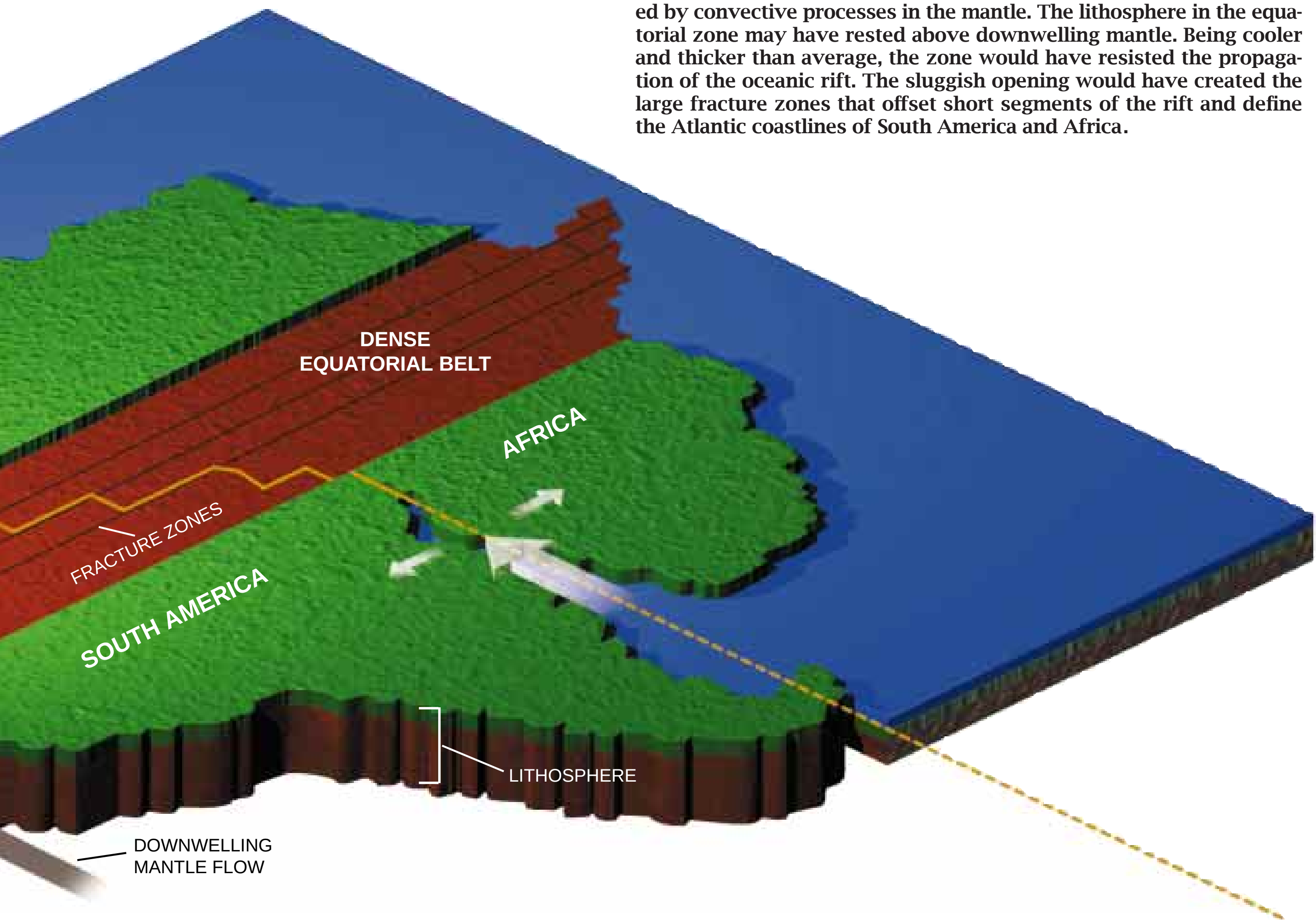
ed 670 kilometers below the surface, marks the boundary between the upper and lower mantle (the lattice structure of the mantle minerals changes across that boundary because of high pressure). An additional major transition known as the Mohorovicic discontinuity, or Moho, separates the dense mantle from the crust. The Moho lies 30 to 50 kilometers below the surface of the continents and less than 10 kilometers below the seafloor in the ocean basins. The lithosphere, which includes the crust and the upper part of the mantle, behaves like rigid plates lying above a hotter, more pliable lower part of the mantle called the asthenosphere.

This ordered, layered structure might seem to imply that the earth's interior is static. On the contrary, the deep earth is quite dynamic. Thermal energy left over from the time of the earth's formation, augmented by energy released through the radioactive decay of elements such as uranium and thorium, churns the material within the earth. The heat travels across the earth's inner boundaries and sets into motion huge convection currents that carry hot regions upward and cold ones downward. These processes ultimately cause many of the broad geologic phenomena on the surface, including mountain building, volcanism and the motions of continents.

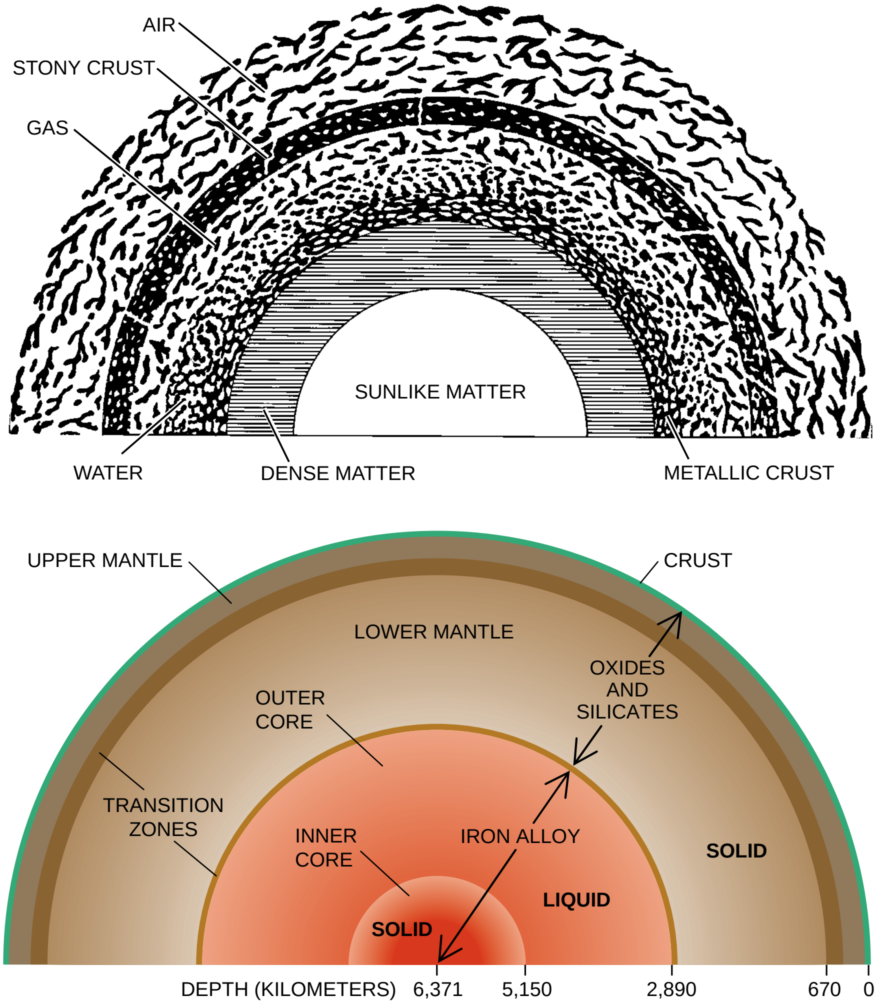
Among the regions offering the best access to the earth's insides are midocean ridges. These ridges dissect all the major oceans. They actually make up a system that winds around the globe like the seams of a baseball, stretching a total of more than 60,000 kilometers. The Mid-Atlantic Ridge is a part of that global ridge system. A huge north-south scar in the ocean floor, it forms as the eastern and western parts of the Atlantic move apart at a speed of roughly one centimeter per year. In addition to the frequent earthquakes that take place there, the summit of the Mid-Atlantic Ridge spews out hot magma during frequent volcanic eruptions. The magma cools and solidifies, thus forming new oceanic crust. The ridge is higher than the rest of the Atlantic basin. At progressively farther distances from the ridge, the seafloor deepens with respect to sea level, presumably because the lithospheric plate that forms the bottom of the Atlantic contracts as it gradually cools with age.

The magma that rises at the Mid-Atlantic Ridge obviously originates in the upper mantle. Geologists have known for years, however, that the material that surfaces at midocean ridges differs considerably from that composing the mantle. Magma at ocean ridges forms a common kind of rock known as basalt. But researchers have found that seismic waves travel through the upper mantle at a rate of more than eight kilometers per second, far faster than they would pass through basalt.

The only material that could possibly allow such a high velocity of sound is a type of dense, dark-green rock called peridotite. Peridotite consists mostly of three silicon-based minerals: olivine, a dense silicate containing magnesium and iron; orthopyroxene, a similar but less dense mineral; and clinopyroxene, which incorporates some aluminum



BIRTH OF THE ATLANTIC 100 million years ago may have been affected by convective processes in the mantle. The lithosphere in the equatorial zone may have rested above downwelling mantle. Being cooler and thicker than average, the zone would have resisted the propagation of the oceanic rift. The sluggish opening would have created the large fracture zones that offset short segments of the rift and define the Atlantic coastlines of South America and Africa.



EARTH'S INTERIOR was imagined by the French philosopher René Descartes in the 17th century (*top*). He viewed the earth as having a nucleus made of a hot, sunlike fluid covered by a dense, opaque solid. Succeeding layers consisted of metal, water, gas, stone and air. In the modern view (*bottom*), a solid inner core is cloaked by a molten outer core; both are made of iron alloy. The mantle is composed mostly of solid silicates and oxides of iron and magnesium.

and more than 20 percent calcium. Peridotites also have small quantities of spinel, an oxide of chromium, aluminum, magnesium and iron.

How can basaltic magma be produced from a mantle made of peridotite? More than 20 years ago experimental petrologists such as Alfred E. Ringwood and David H. Green and their colleagues at the Australian National University exposed samples of peridotite to elevated temperatures (1,200 to 1,300 degrees Celsius) and high pressures (more than 10,000 atmospheres). These values duplicate the temperature and pressure that exist in the suboceanic upper mantle roughly 100 kilometers below the seafloor. The workers showed that gradual decompression of peridotite at those high temperatures melts up to 25 percent of the rock. The melt had a basaltic composition similar to that of melts in midocean ridges.

These experimental results support the view that hot, peridotitic material rises under the midocean ridges from depths exceeding 100 kilometers below the seafloor. As it moves upward, the mantle peridotite decompresses and partially melts. The melted part takes on the composition of a basaltic magma and separates from the periodotite that did not melt. It rises rapidly toward the surface. Part of the melt erupts on the seafloor along the crest of the midocean ridge, where it cools and solidifies and adds to the ridge crest. The remainder cools and solidifies slowly below the surface, giving rise to new oceanic crust.

If the model outlined above happened all along the Mid-Atlantic Ridge, the summit of the ridge would roughly be at the same depth below sea level along its length. This depth would mark an equilibrium level determined by the temperature and initial composition of the upper mantle below the ridge.

In the real world such consistency is highly unlikely. Small variations in mantle temperature along the ridge would cause the summit to settle at varying elevations. Regions of suboceanic mantle where temperatures are higher have lower densities. As a result, the ridge summits there will be higher. In addition, a hotter mantle would melt more and produce a thicker basaltic crust.

The summit of the Mid-Atlantic Ridge shows just such variations in depth below sea level. For instance, along the ridge between about 35 and 45 degrees north latitude lies an area of abnormally high topography. Earth-orbiting satellites have detected in the same region an upward swell in the level of the geoid (the equilibrium level of the earth's surface, roughly equivalent to the average sea level).

Researchers generally attribute this swell to the influence of a so-called hot spot centered on the Azores island group. Hot spots are zones that have high topography and excess volcanism. They are generally ascribed to unusually high mantle temperatures. Most oceanic islands, including the Hawaiian Islands and Iceland, are thought to be the surface expressions of hot spots. The source of the heat is thought to lie in the boundary zones deep inside the earth, even as deep as the core-mantle boundary [see "The Core-Mantle Boundary," by Raymond Jeanloz and Thorne Lay; SCIENTIFIC AMERICAN, May 1993].

My colleagues and I set out to test that theory by exploring how the topography along the Mid-Atlantic Ridge relates to the temperature, structure and composition of the underlying mantle. One way to collect such information is to examine the velocities of seismic waves passing through the mantle under the ridge. Another approach involves searching for local variations in the chemistry of basalts that erupted along the axis of the ridge. Those variations can be used to infer the extent of melting and the physical nature of the mantle source from which they derived.

I followed a third approach. I attempted to collect rock samples of mantle peridotite. Some peridotite is left as a solid residue after the basaltic magma component melts out of the upper mantle rocks. Mantle rocks normally lie buried under several kilometers of ocean crust, but in some cases blocks of upper mantle peridotite are accessible. They are typically found where the axis of the midocean ridge is offset laterally by transform faults or where the mantle rocks have been transported close to the seafloor, so that they can be sampled by drilling or dredging or retrieved directly through the use of a submersible.

In 1989, during a mostly French expedition organized by Jean-Marie Auzende of the oceanographic institution IFREMER in Plouzané, France, we used a small submersible to gather samples of a section of upper mantle at the Vema transform zone in the Atlantic, 10 degrees north of the equator. Here a transform fault, cutting a deep valley through the oceanic crust, offsets the axis of the Mid-Atlantic Ridge by about 320 kilometers. We planned to descend to the seafloor—more than five kilometers down—in the submersible *Nautile* to explore the walls of that transform valley. We hoped to find an exposed, pristine section of mantle and crust. Most of our colleagues viewed our task with skepticism: the prevalent opinion was that the normal sequence of upper mantle and crust is completely disrupted near a transform fault.

Nevertheless, we pressed on. We began a series of dives that started at the base of the transform valley wall and moved up the slope. Each dive lasted about 12 hours, about half of which was spent descending to the seafloor and returning to the surface. The cramped quarters of the *Nautile* accommodate two pilots and one scientist, who lies face down for the duration of the trip.

On our first dive we verified that the base of the section consists of mantle peridotite. On the second day we discovered a layer of gabbros—rocks that form below the seafloor when basaltic melts cool slowly—resting above the peridotite. According to widely accepted geophysical models, gabbros are the main component of the lower part of the oceanic crust.

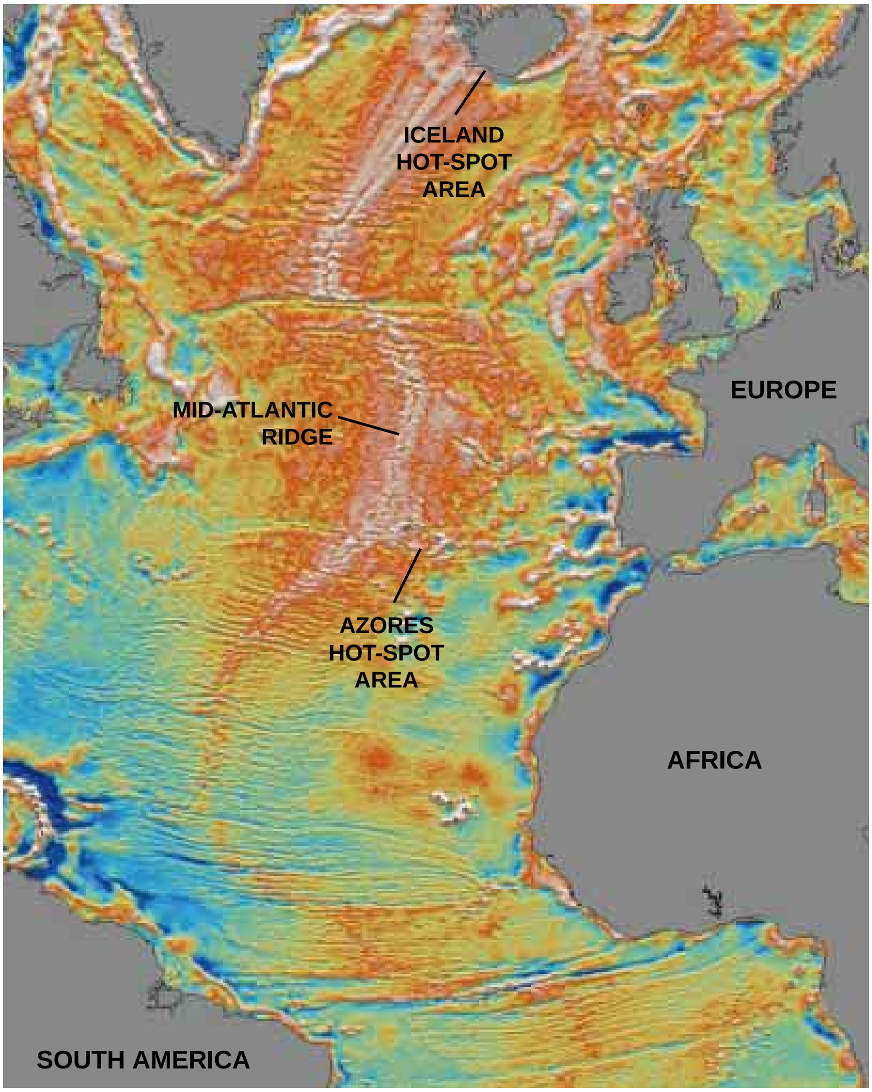
The next day I took the *Nautile* on a dive that started from the level reached by the submersible the previous day. As I progressed along the slope, skimming the seafloor, a spectacular rock formation called a dike complex gradually revealed itself to my eyes. Theory holds that dike complexes form where hot molten material from the mantle squirts upward toward the seafloor through many narrow fissures in the crust. Never before had a dike complex been observed on the seafloor.

The dike complex, about one kilometer thick, was topped by a layer of pillow basalt, the form taken by basaltic magma when it cools and solidifies rapidly on eruption to the seafloor. During the next several days, we explored a different section and confirmed our previous findings. We were quite excited because no one had ever before observed a complete and relatively undisturbed section of oceanic upper mantle and crust. We immediately documented our discovery in a short paper that we mailed to *Nature* as soon as we docked a few weeks later.

During the dives, we had used the *Nautile*'s mechanical arm to grab a number of samples of mantle peridotite. Those samples, along with many others I and other researchers collected along the ridge, enabled us to search for regional heterogeneities in the chemistry of the upper mantle.

To analyze the mantle minerals in the Atlantic peridotite samples, my colleagues Peter J. Michael and Monique Seyler, then at the Lamont-Doherty Geological Observatory, and I used an electron microprobe. This instrument focuses a beam of electrons only a few microns in diameter onto a slice of rock. In response, the mineral emits x-rays of characteristic wavelengths. An analysis of the wavelengths and intensities of these x-rays allows a determination of the chemical composition of the mineral. Collaborating with Nobumichi Shimizu of the Woods Hole Oceanographic Institution, we also used a different instrument—an ion microprobe—to determine the concentration of trace elements such as titanium, zirconium and rare-earth elements. The ion probe focuses a beam of ions onto a sample, which dislodges other ions in the sample for measurement. The method enabled us to determine the concentrations of trace elements down to a few parts per billion.

Such analyses reveal much about the



SATELLITE MAP of the North Atlantic reveals the topography of the seafloor. The satellite used radar to measure variations in sea level, which correlate with the bumps and depressions underwater. The Mid-Atlantic Ridge is clearly visible. The ridge swells into broad platforms above the hot spots associated with Iceland and the Azores. A large fracture zone breaks the ridge between the hot spots.

conditions in the mantle where the sample rocks formed, because the temperatures and pressures there produce distinct compositions in the peridotites. Petrologists, including Green and A. Lynton Jaques of the Australian Geological Survey Organization, have shown that partial melting modifies the relative abundances of the original minerals in the peridotite. Some minerals, such as clinopyroxene, melt more easily than do others and hence rapidly decrease in abundance during the melting. Moreover, the partial melting process changes the composition of the original minerals: certain elements in them, such as aluminum and iron, tend to follow the melt. Their concentration in the minerals decreases as melting proceeds. Other elements, such as magnesium and chromium, tend to stay behind, so that the solid residue becomes enriched with them. Thus, as a result of partial melting, olivine (a silicate of iron and magnesium) becomes more magnesium-rich and iron-poor; orthopyroxene and clinopyroxene lose some of their aluminum; the ratio of chromium to aluminum in spinel increases; and so on.

Our data showed that substantial regional variations exist in the composition of the mantle. For instance, the chromium-to-aluminum ratio of spinel is highest in peridotites sampled from a broad area between about 35 degrees and 45 degrees north latitude. The ratio suggests that the degree of melting of the upper mantle lying below this region may reach as high as 25 percent. In most parts, about 10 to 20 percent of the mantle melts during the trip upward. This area of above-average melting corresponds to the Azores hot-spot region, lending credibility to the theory that hot spots result from unusually hot mantle plumes upwelling deep within the earth. Other findings support that idea, including work by Henry J. B. Dick of Woods Hole, who also studied oceanic peridotites, and by Emily M. Klein working with Charles H. Langmuir of Lamont-Doherty, who independently examined the chemistry of basalts along the Mid-Atlantic Ridge.

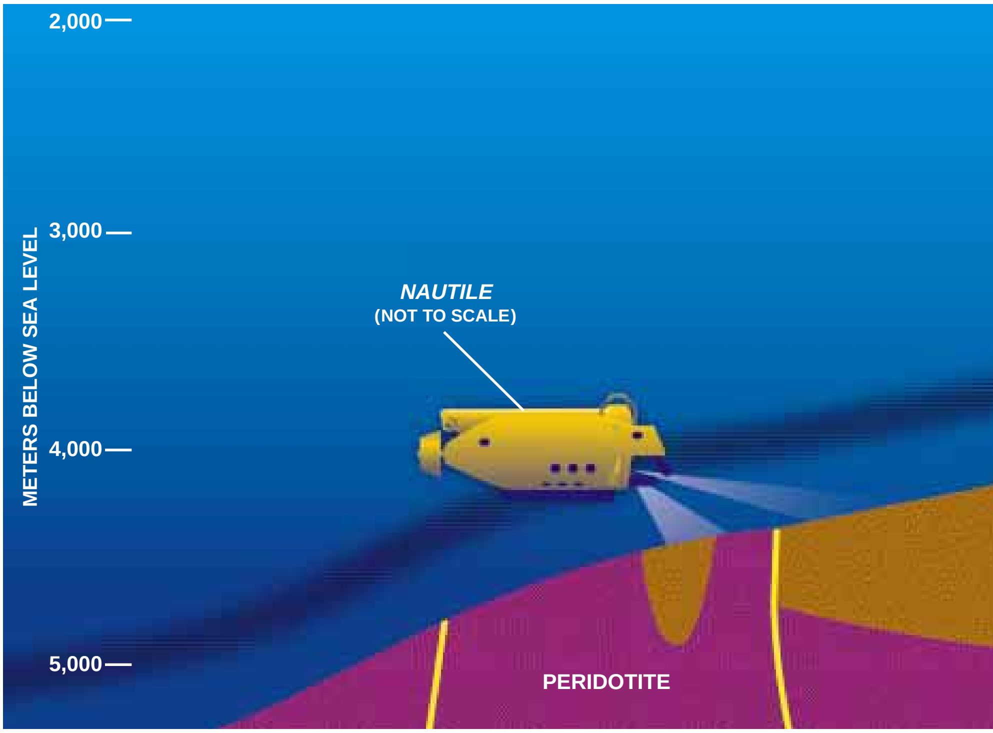
Clearly, a hot spot would seem to be the cause of so much melting. In fact, assuming that temperature alone causes the melting in the Azores hot-spot region, we calculated that the hot-spot mantle would need to be about 200 degrees C warmer than the mantle from elsewhere below the ridge.

Is there a way of testing the validity of this temperature estimate and its underlying assumption? A number of geothermometers have been proposed. They are based on the observation that certain mineral pairs that coexist in equilibrium in the mantle undergo temperature-dependent chemical reactions. For instance, the orthopyroxene and clinopyroxene in a mantle peridotite react with each other until they reach an equilibrium composition that depends on temperature. Laboratory experiments have calibrated that relation. Thus, determining the composition of the coexisting mineral pair can indicate the temperature at which the members of the pair reached equilibrium.

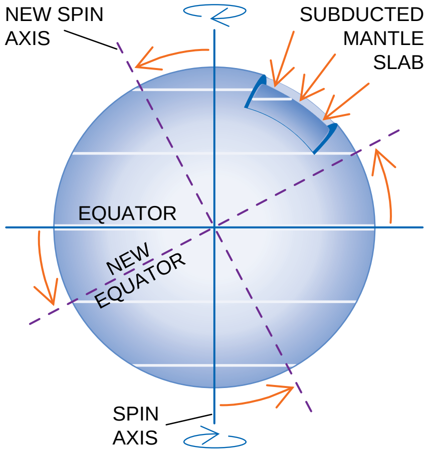
I applied two geothermometers, one devised by Donald H. Lindsley of the State University of New York at Stony Brook and the other by Peter R. A. Wells of the University of Oxford, to the Mid-Atlantic Ridge peridotites. The results were surprising. They did not show higher temperatures in the hot-spot region; if anything, the region gives temperatures that are slightly lower.

Why did we not find higher mantle temperatures for a region that displays high melting? One possibility is that the upper mantle there has a composition that causes it to melt more easily. Water could be the main factor. Experiments by Peter J. Wyllie of the California Institute of Technology, Ikuo Kushiro of the University of Tokyo and the Carnegie Institution of Washington, and several others have demonstrated that trace amounts of water and other volatile elements in peridotite drastically decrease its melting temperature. So, if such a "wet" mantle upwelled under a stretch of midocean ridge, it would start melting more deeply in the earth than normal, "dry" mantle would. By the time the peridotite reached the surface, it would have undergone a degree of melting significantly greater than that of dry mantle under similar temperatures.

Is there any evidence that the upper mantle below the Azores hot-spot area is wetter than the mantle elsewhere below the Mid-Atlantic Ridge? Indeed there is. A few years ago Jean-Guy E. Schilling and his co-workers at the University of Rhode Island reported that basalts from the segment of the hot spot situated between 35 and 45 degrees north latitude contain three to four times more water than do normal midocean ridge basalts. The basalts also have abnormally high concentrations of other volatile elements such as chlorine and bromine. Moreover, Schilling found that the basalts from the hot-spot ridge segment contain a much greater abundance of several chemical elements (mostly light rare-earth elements) than do the normal midocean ridge basalts. The anomalously high concentration of those elements means that the parent mantle in the hot-spot



EXPLORATION OF THE SEAFLOOR by the *Nautile* occurred at the Vema transform fault, which lies in the northern section of the Mid-Atlantic Ridge. Along the southern wall, mantle peridotites were found to outcrop in the lower part of the slope. Above them were gabbros, rocks created by the slow cooling of basaltic melt (the



SHIFTING OF THE EARTH'S AXIS can be influenced by the sinking of cold, dense slabs of mantle. Such sinking occurs in subduction zones, such as those surrounding the Pacific Ocean. The earth's axis of rotation would tend to shift so that the equator would move closer to the dense slabs.

area harbors an enriched supply of these elements.

It seems, therefore, that the mantle below the Azores hot spot differs from the normal sub-Mid-Atlantic Ridge mantle not so much by being hotter as by having incorporated at some stage water and other fluids that changed its chemical composition and melting behavior. This chemical transformation of mantle peridotite by fluids is called metasomatism. It would explain why wet mantle near the surface would have experienced more melting than normal mantle would. It may also explain why the equilibrium temperatures estimated from peridotites at the Azores hot spot do not appear higher than average. Melting reactions consume heat, so that partial melting of upwelling mantle may actually have cooled the surrounding mantle. The higher the degree of melting, the greater the heat loss.

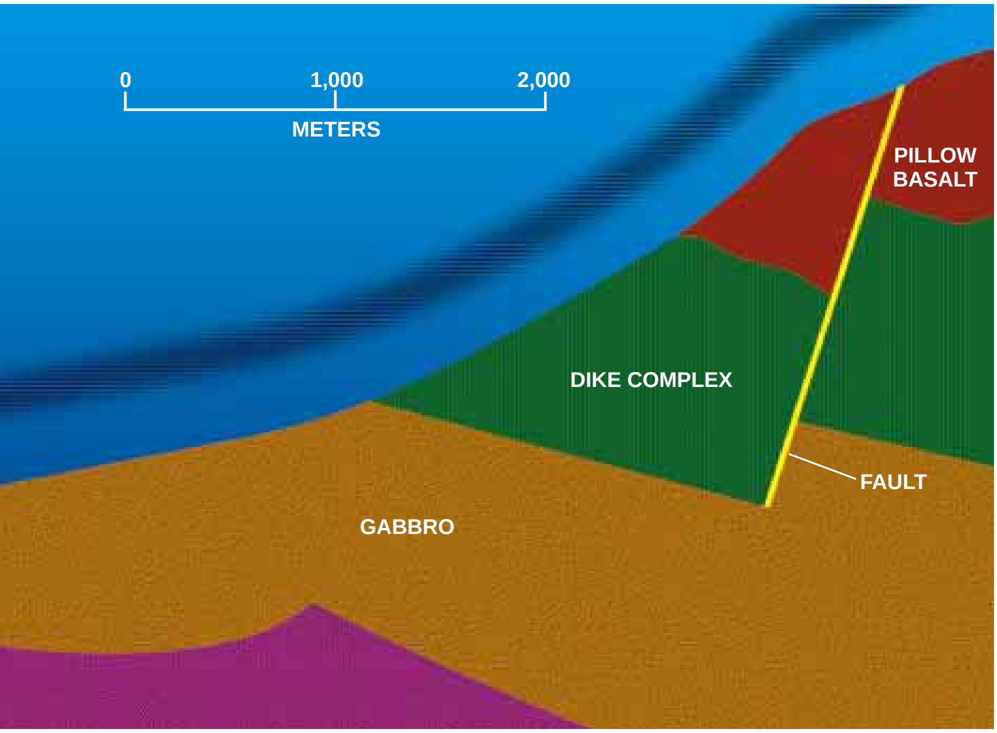
Where does the water that produces mantle metasomatism come from? One possible source is the sinking of slabs of old oceanic lithosphere in subduction zones at the margin of the oceans. This process recycles water into the mantle. Water could also be released in the upper mantle during degassing processes. For instance, methane, a gas that might be present in the deep mantle, could be oxidized once the upwelling reaches the upper mantle region. The reaction would yield water (plus carbon, either as diamond or graphite).

Because of its inferred below-average mantle temperature, the Azores hot spot clearly does not fit into the usual definition. How is one to distinguish the different types of hot spots (those that are really hot and those that are not so hot) and deduce their origins? Helium gas may lead us toward an answer. The element can form two stable isotopes: helium 3 and helium 4. Helium 4 is produced continuously in the earth's crust by the radioactive decay of uranium and thorium. Most investigators believe helium 3 stems from an incomplete escape of primordial gases that were incorporated within the earth in the early stages of its history. The ratio of helium 3 to helium 4 in the earth's atmosphere and in seawater is roughly one to one million.

Yet that ratio is different in rock samples retrieved from midocean ridges. Groups led by Harmon Craig of the Scripps Institution of Oceanography and Mark D. Kurz of Woods Hole have shown that the helium 3 to helium 4 ratio of basalts along midocean ridges is about eight times higher than the atmospheric ratio. The ratio at hot spots such as those under Hawaii and Iceland is even higher, perhaps reaching 30 times the atmospheric ratio. The large amount of helium 3 suggests that ancient gases are escaping at those sites. Thus, hot-spot areas with high ratios confirm the notion that they represent upwellings of hot plumes from deep within the earth.

A few hot spots—the Azores one among them—have basalts with a ratio of helium 3 to helium 4 lower than those of the midocean ridge basalts. The primordial component of those hot spots was somehow lost or diluted. The Azores hot spot may thus be a melting anomaly of relatively superficial origin in the mantle. It may not be linked to a thermal plume originating from the deep mantle or the core-mantle boundary. These hot spots may not be truly hot and perhaps are best classified as "wet spots," for the key role fluids may play in their formation.

Our studies of mantle peridotites from the Mid-Atlantic Ridge suggest that some areas with cooler mantle temperatures may represent the return strokes of the convection cycle in the mantle—that is, the downwelling regions. To understand the deduction, we must look south of the Azores region, to the equatorial zone of the Mid-Atlantic Ridge. The mineral composition of peridotites recovered from the equatorial Atlantic indicates that they underwent little or no melting, which implies that the mantle temperature was exceptionally low. Nadia Sushevskaya of the Vernadsky Institute of Geochemistry of the Russian Academy of Sciences reached similar conclusions in her study of basalts from the equatorial Atlantic. Moreover,



melted part of peridotite). The *Nautile* also discovered a dike complex, formed when basaltic melt cools and solidifies before reaching the seafloor. Above the dike complex lay pillow basalt, the form taken by basaltic melt that erupts on the seafloor and cools rapidly on contact with ocean water.

the crust of the equatorial Mid-Atlantic Ridge lies deeper below the geoid than that of the ridge at higher latitudes, and the velocity of the seismic waves is faster in the upper mantle below the equatorial Mid-Atlantic Ridge than at higher latitudes. Both these observations imply a denser, colder upper mantle below the equatorial region of the Atlantic. The temperature of the upper mantle there may be more than 150 degrees C lower than the mantle temperatures elsewhere below the ridge.

A plausible explanation for the relatively cool and dense equatorial upper mantle is that it results from downwelling mantle currents. Upwelling plumes from the northern and southern Atlantic mantle domains may meet here, give up their heat to their cooler surroundings and then sink.

Klein, Jeffrey Weissel and Dennis E. Hayes and their co-workers at Lamont-Doherty found a somewhat similar situation in a stretch of midocean ridge that runs between Australia and Antarctica. This ridge is exceptionally deep, and the basalts recovered from its crest give evidence of having been produced by extremely limited melting in the mantle. Their findings are consistent with the idea that broad mantle convection currents sweeping from the Pacific and the Indian Ocean converge and sink between Australia and Antarctica.

The equatorial position of the downwelling Atlantic mantle belt may not be arbitrary. It is possible that the earth's rotation and convection in the mantle are intimately connected phenomena. In the late 1800s George Darwin (the second son of Charles) pointed out that the distribution of large masses on the surface (such as continents) affects the position of the earth's axis of rotation. Several scientists since then have investigated how density inhomogeneities in the mantle cause true polar wander (that is, the shifting of the entire mantle relative to the earth's axis). The wander results from the natural tendency of a spinning object to minimize the energy spent for its rotation.

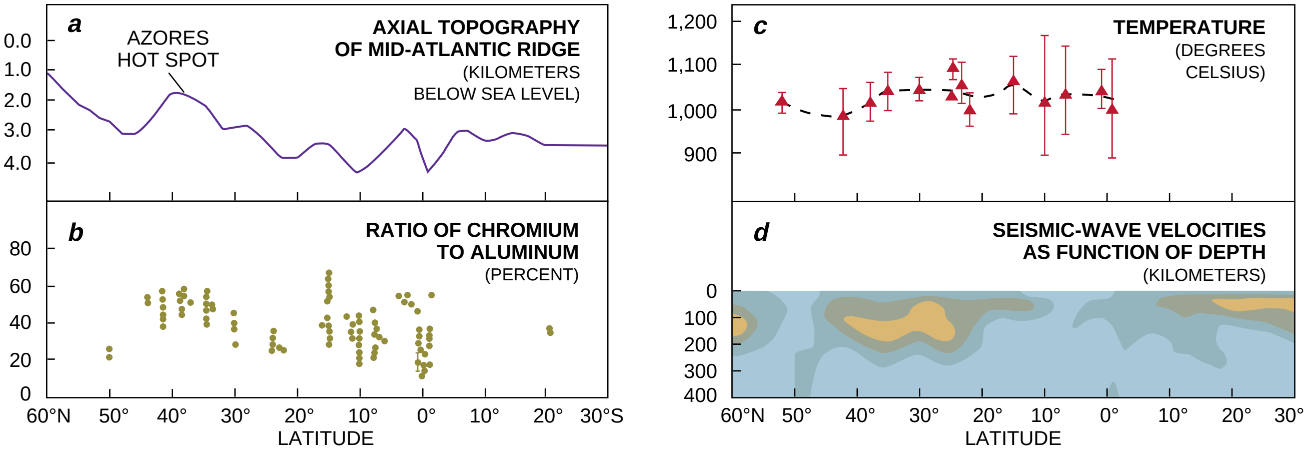
The redistribution of mass inside the earth may be recorded in the mantle. The late H. William Menard and LeRoy M. Dorman of Scripps suggested that the depth of midocean ridges generally depends on latitude: ridges become deeper toward the equator and shallower toward the poles. Moreover, gravity measurements revealed that an excess of mass sits below the equatorial areas. These data imply that abnormally cold and dense masses exist in the equatorial upper mantle.

The sinking of cold, dense slabs into the mantle appears to influence true polar wander. Evidence strongly suggests that the mantle is less viscous near the surface than it is deeper down. Any dense masses that find their way to the mantle, such as those that occur in subduction zones at the edge of some oceans, will affect the position of the rotation axis. The equator would tend to shift toward the dense masses. If high-density masses are near the equator, downwelling and cooler mantle spots are likely to prevail in the equatorial upper mantle. That phenomenon would explain at least qualitatively the cold upper mantle belt and resulting lack of normal melting in the equatorial zone of the Atlantic and probably the Pacific.

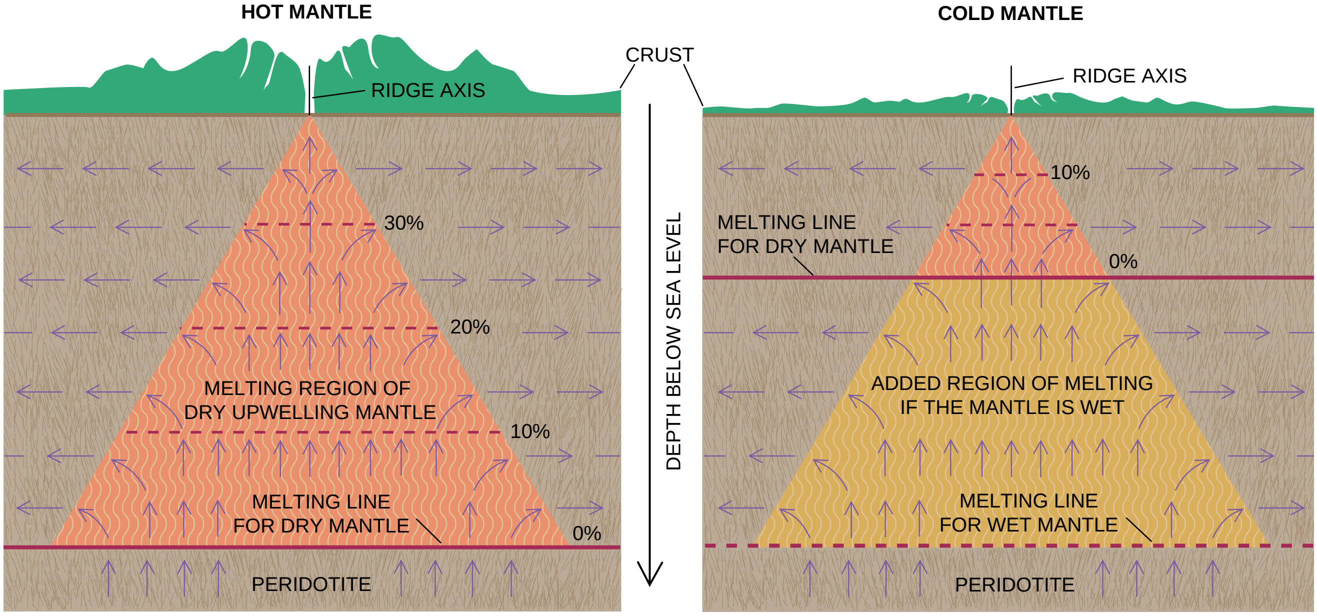
A downwelling mantle boundary could account for the peculiar geology of the equatorial region. In 1835, during his famous voyage with the H.M.S. *Beagle,* Charles Darwin landed on some desolate, small rocky islets that barely reached above sea level. The islands, now known as the St. Peter-Paul rocks, are in the center of the Atlantic, just a few miles north of the equator. Darwin described how nesting colonies of the seabirds called sulas compete with large red crabs for each parcel of available space on the rocks. The same contest can be observed today.

Darwin also noted that the islets are geologically different from most oceanic islands, insofar as they are not volcanic. This observation has been confirmed, most recently by William G. Melson of the Smithsonian Institution and Mary K. Roden of the State University of New York at Albany and their co-workers. The St. Peter-Paul rocks are in fact made of peridotites and represent an uplifted body of upper mantle.

The peridotites of the St. Peter-Paul rocks, however, differ from those collected elsewhere along the Mid-Atlantic Ridge. The chemistry of the St. Peter-Paul minerals indicates that they underwent little or no melting. The materials equilibrated in the mantle at a low temperature. They resemble peridotites from continental, or "preoceanic," rifts (such as those exposed in the island of



PROFILES ALONG THE AXIS of the Mid-Atlantic Ridge reveal the anomalous nature of the Azores area. Here the seafloor broadly swells (*a*). Measurements of the ratio of chromium to aluminum in spinel, a component of mantle peridotite, indicate that the mantle melted most here (*b*). These data suggest that the Azores region is a hot spot, an area of hot mantle. A discrepancy emerges, however, when temperature calculations are incorporated: the Azores region appears to be slightly cooler (*c*). The Azores area may have undergone much melting because the mantle material there is wet, as indicated by measurements of the velocities of seismic waves moving through the upper mantle (*d*). Wet areas have below-average densities, so seismic waves travel more slowly (*yellow*) through them. The equatorial area shows fast seismic velocities (*blue*), suggesting the presence of dense material and perhaps marking a site of mantle downwelling.


HOT MANTLE
RIDGE AXIS
CRUST
COLD MANTLE
RIDGE AXIS
30%
20%
10%
0%
MELTING REGION OF DRY UPWELLING MANTLE
MELTING LINE FOR DRY MANTLE
PERIDOTITE
DEPTH BELOW SEA LEVEL
10%
0%
MELTING LINE FOR DRY MANTLE
ADDED REGION OF MELTING IF THE MANTLE IS WET
MELTING LINE FOR WET MANTLE
PERIDOTITE

UPWELLING MANTLE melts to an extent that depends on whether the mantle is hot (*left*) or cold (*right*). The percentages indicate the amount of peridotite that melts. Melting proceeds until the peridotite stops rising and starts flowing horizontally. The hotter the mantle, the deeper the melting begins. As a result, more of the mantle melts, creating a thicker crust. Cold mantle melts less, unless it harbors fluids. In that case, it begins to melt much more deeply in the earth and may even melt more than hot mantle can. Wet mantle may explain why the Azores hot spot is rather cool.

Zabargad in the Red Sea) rather than those from ocean ridges. Moreover, they show signs of having been strongly affected in the mantle by metasomatism—more so than did the samples we collected from the Mid-Atlantic Ridge.

Hence, the St. Peter-Paul islets expose what appears to be a mantle typical of a continental rift rather than of a midocean ridge. Indeed, geochemistry work by Roden and her colleagues suggests that the metasomatism that affected the St. Peter-Paul mantle occurred about 150 million years ago; that time marks a rift stage that preceded the separation of Africa and South America in the equatorial Atlantic (that is, sometime during the breakup of Pangaea).

How could blocks of originally subcontinental mantle have been left in the center of the Atlantic Ocean? The answer may lie in the way Pangaea broke up in the face of a cold, dense upper mantle in the equatorial region.

A colder-than-normal equatorial mantle when the Atlantic first opened would imply a colder and thicker continental lithosphere along the equatorial belt. (The equator 100 million years ago crossed the future Atlantic coastlines of Africa and South America roughly along the same position as it does today.) The cold and thick equatorial lithosphere must have resisted the rift propagating from the south. The equatorial region may have behaved as a "locked zone" (in the sense used by French geologist Vincent E. Courtillot). As a result, the equatorial Atlantic opened sluggishly. This slow opening may have created the large equatorial fracture zones, visible today as east-west breaks that offset short segments of the midocean ridge.

During the opening of the equatorial Atlantic, these fracture zones were subjected to strong compressional stresses and intense vertical motions of lithospheric blocks. As a result, blocks of crust may periodically have sprung up through the ocean and sunk back down. Some slivers of continental lithosphere, however, might have been left behind in the middle of the ocean—such as that whose summit we identify as the St. Peter-Paul islets. Hence, just as hot, upwelling mantle regions create distinct types of volcanic islands, so too can cold, downwelling zones cause a different type of island to emerge.

It is interesting to speculate on how the rise and fall of such islands may have influenced life on the earth. One example is the migratory behavior of the green sea turtle (*Chelonia mydas*). These turtles live along the Brazilian coast but make an arduous 2,000-kilometer journey to Ascension island to breed. This curious act may be rooted in the behavior of their ancestors, which thrived 80 million years ago, when the equatorial Atlantic was narrow. The ancient turtles may have used islands that emerged close to the Brazilian coast as breeding grounds. As the Atlantic opened and some of the islands sank, their descendants were forced to extend their trek by hundreds of kilometers.

Much remains to be done before geologists develop a complete picture of mantle convection and its influence on surface geology. Because sending submersibles to the ocean floor is not always practical, other techniques, such as seismic tomography, must be further developed to distinguish wet spots from hot spots. Debate persists as to the origins of the mantle convection and whether it extends into the lower mantle. Indeed, symposia that include theoreticians, geophysicists, geochemists and petrologists invariably yield heated discussions and much dissent. On one point there is unanimity: the earth's mantle is very much alive and is an exciting region to study.

FURTHER READING

THEORY OF THE EARTH. D. L. Anderson. Blackwell Scientific Publications, 1989.

NOT SO HOT "HOT SPOTS" IN THE OCEANIC MANTLE. E. Bonatti in *Science*, Vol. 250, pages 107–111; October 5, 1990.

RIDGES, HOTSPOTS AND THEIR INTERACTION AS OBSERVED IN SEISMIC VELOCITY MAPS. Y. S. Zhang and T. Tanimoto in *Nature*, Vol. 355, No. 6355, pages 45–49; January 2, 1992.

A COLD SUBOCEANIC MANTLE BELT AT THE EARTH'S EQUATOR. E. Bonatti, M. Seyler and N. Sushevskaya in *Science*, Vol. 261, pages 315–320; July 16, 1993.

Targeted Gene Replacement

Researchers can now create mice bearing any chosen mutations in any known gene. The technology is revolutionizing the study of mammalian biology

by Mario R. Capecchi

Every cell of our bodies has within its nucleus an instruction manual that specifies its function. Although each cell carries the same manual, different cell types, such as liver or skin, use different parts of this manual to detail their unique functions. Perhaps most remarkable, the manual contains the information that allows a one-cell embryo, the fertilized egg, to become a fetus and then a newborn child. As the child matures physically and intellectually, he or she is still using the information within the instruction manual. We are each unique, and the manual is slightly different for each of us; it specifies most of the physical and many of the behavioral characteristics that distinguish us as individuals.

This extraordinary manual, otherwise known as the genome, is written in the form of nucleotides, four of which constitute the entire alphabet—adenylate (*A*), cytidylate (*C*), guanylate (*G*) and thymidylate (*T*). It is the precise sequence of the nucleotides in DNA that conveys information, much as the sequence of letters in a word conveys meaning. During each cell division, the entire manual is replicated, and a copy is handed down from the mother cell to each of its two daughters. In humans and mice, the manuals each contain three billion nucleotides. If the letters representing the nucleotides were written down in order so that a page carried 3,000 characters, the manual would occupy 1,000 volumes, each consisting of 1,000 pages. Thus, a very complex manual is required to orchestrate the creation of a human or mouse from a fertilized egg.

Recently my colleagues at the University of Utah and I developed the technology for specifically changing a letter, a sentence or several paragraphs in the instruction manual within every cell of a living mouse. By rewriting parts of the manual and evaluating the consequences of the altered instructions on the development or the postdevelopmental functioning of the mouse, we can gain insight into the program that governs these processes.

The functional units within the instruction manual are genes. We specifically change the nucleotide sequence of a chosen gene and thereby alter its function. For instance, if we suspected a particular gene were involved in brain development, we could generate mouse embryos in which the normal gene was "knocked out"—that is, completely inactivated. If this inactivation caused newborn mice to have a malformed cerebellum, we would know that the gene in question was essential to forming that part of the brain. The process by which specified changes are introduced into the nucleotide sequence of a chosen gene is termed gene targeting.

Much of what is learned from gene-targeting experiments in mice should benefit humans, because an estimated 99 percent or more of the genes in mice and humans are the same and serve quite similar purposes. Application of the technology in mice is already clarifying not only the steps by which human embryonic development occurs but also the ways in which our immune system is formed and used to fight infection. Gene targeting should also go far toward explaining such mysteries as how the human brain operates and how defects in genes give rise to disease. In the latter effort the technique is being used to produce mouse models of human disorders—among them, cystic fibrosis, cancer and atherosclerosis.

Excitement over gene targeting stems from another source as well. It promises to expand on the knowledge generated by the genome project. This large-scale undertaking aims to determine the nucleotide sequence of every gene in the mouse and human genomes (approximately 200,000 genes in each). Currently we know the functions of only a minute percentage of the genes in either species. The nucleotide sequence of a gene specifies the amino acids that must be strung together to make a particular protein. (Proteins carry out most of the activities in cells.) The amino acid sequence of a protein yields important clues to its roles in cells, such as whether it serves as an enzyme, a structural component of the cell or a signaling molecule. But the sequence alone is not sufficient to reveal the particular tasks performed by the protein during the life of the animal. In

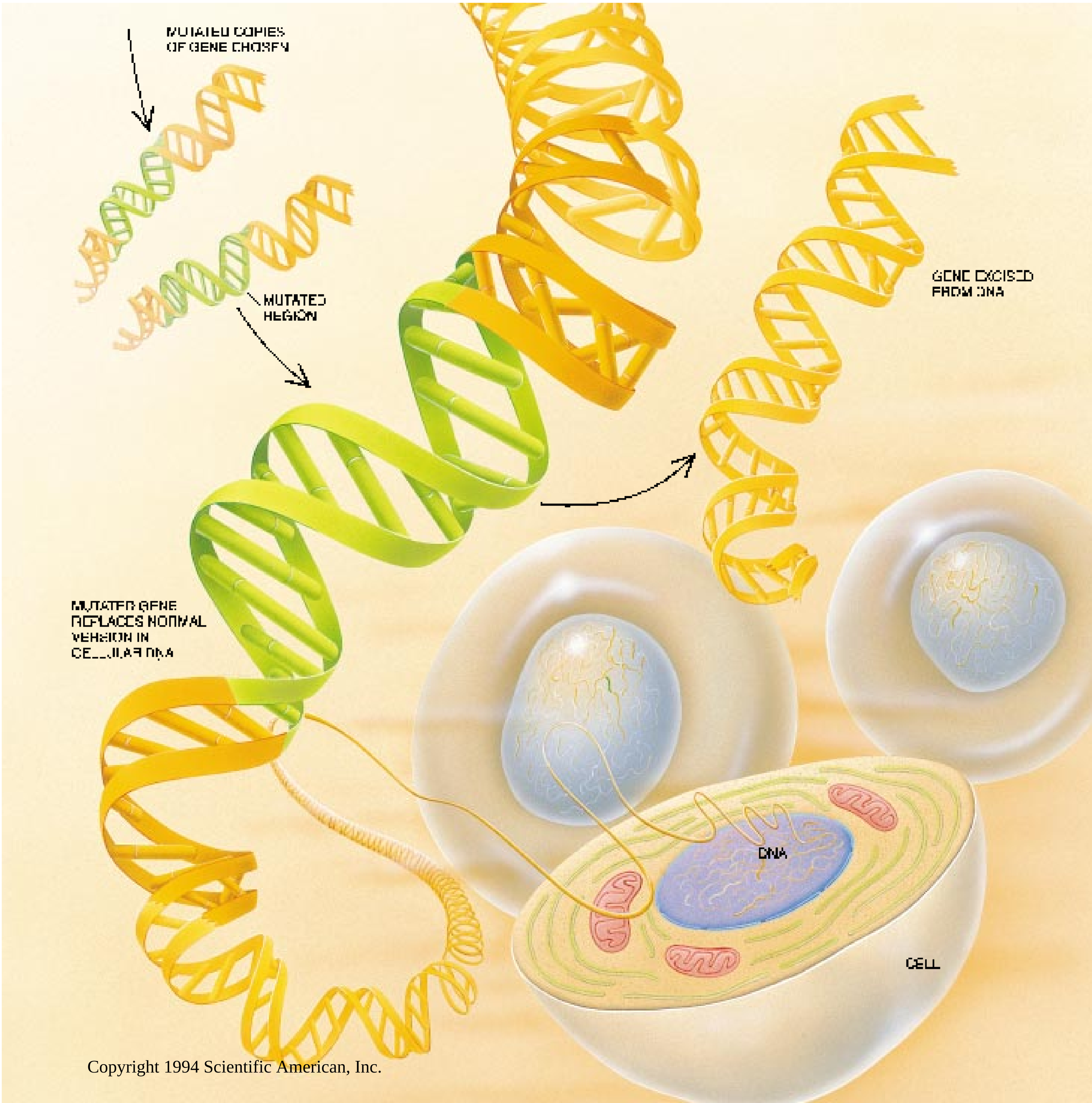
MARIO R. CAPECCHI, who was born in Verona, Italy, is an investigator at the Howard Hughes Medical Institute and professor of human genetics at the University of Utah School of Medicine. In addition to developing the techniques described in this article, Capecchi has helped elucidate the mechanism of protein synthesis. He has also contributed to the discovery of enhancers in DNA and to the development of a now widely used technique for directly injecting DNA into the nuclei of cells.

TARGETED MUTATION can be generated in a selected cellular gene by inserting mutated copies of the gene (*green-and-gold strips at far left*) into cells and allowing one copy to take the place of the original, healthy gene (*gold fragment at far right*) on a chromosome. Such altered cells are helping researchers to produce mice carrying specific genetic mutations. The finding of a curled tail and a balance-and-hearing disorder in one such mouse (*above*) led to the discovery that the affected gene, *int-2,* participates in development of the tail and the inner ear.

contrast, gene targeting can provide this information and thereby move our understanding of the functions of genes and their proteins to a much deeper level.

Gene targeting offers investigators a new way to do mammalian genetics—that is, to determine how genes mediate various biological processes. This technique was needed because the classical methods of genetics, which have been highly successful in analyzing biological processes in simpler organisms, were not readily adaptable to organisms as complex as mammals.

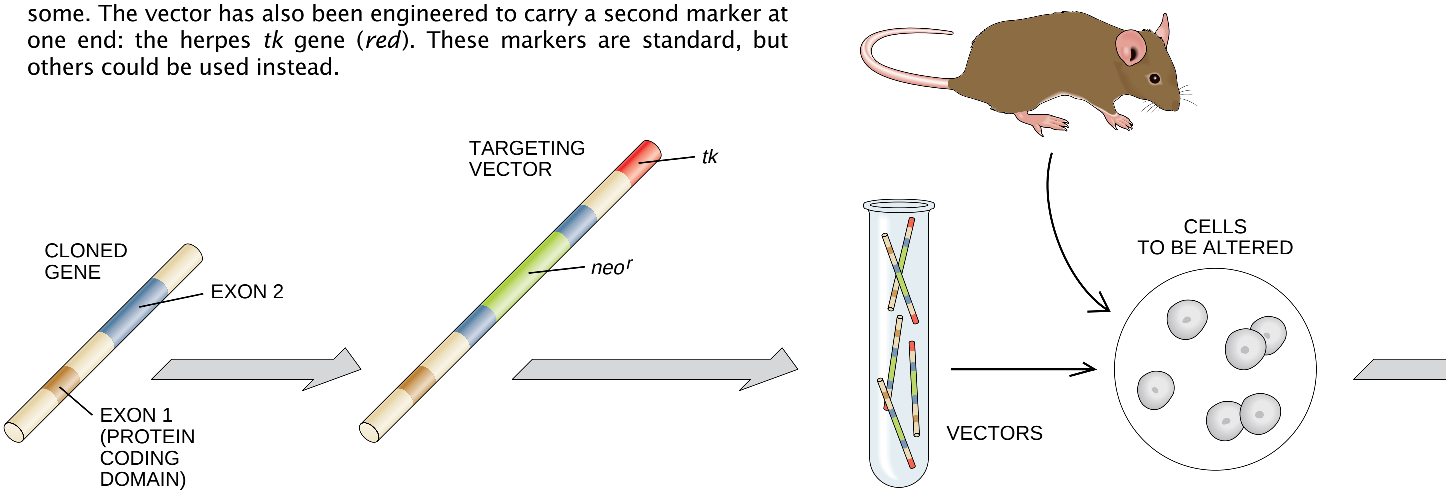
If geneticists want to learn, for example, how single-cell organisms, such as bacteria or yeast, replicate their DNA, they can expose a billion or more individuals to a DNA-damaging chemical (a mutagen). By choosing the right dosage of mutagen, they can ensure that each individual in that population carries a mutation in one or more genes. From this population of mutagenized bacteria or yeast, the geneticists can identify individuals not capable of replicating their DNA. The use of such a large mutagenized population makes it likely that separate individuals will be found with mutations in each of the genes required for DNA replication. (For a process as complicated as duplicating the bacterial or yeast genome, more than


MUTATED COPIES OF GENE CHOSEN
MUTATED REGION
GENE EXCISED FROM DNA
MUTATED GENE REPLACES NORMAL VERSION IN CELLULAR DNA
DNA
CELL

How Targeted Gene Replacement Is Accomplished in Cultured Cells

1. Workers alter copies of a gene (*strip at far left*) in the test tube to produce what is called the targeting vector (*lengthened strip*). The gene shown here has been inactivated by insertion of the *neo*^{r} gene (*green*) into a protein coding region (*blue*). The *neo*^{r} gene will serve later as a marker to indicate that the vector DNA took up residence in a chromosome. The vector has also been engineered to carry a second marker at one end: the herpes *tk* gene (*red*). These markers are standard, but others could be used instead.

2. Once a vector, with its dual markers, is complete, it is introduced into cells (*gray*) isolated from a mouse embryo.



100 genes are involved.) Once the individual genes are identified, their specific role in DNA replication, such as which genes control the decision to copy the DNA and which control the accuracy and rate of copying, can be determined.

Similar approaches have been applied to multicellular organisms, which are more complex. Two favorites of geneticists are *Caenorhabditis elegans,* a tiny, soil-dwelling worm, and *Drosophila melanogaster,* a common fruit fly. But even in these relatively simple forms of multicellular organisms, identifying all the genes involved in a specific biological process is more demanding.

A number of factors contribute to this increased difficulty. One is the size of the genome. The genome of the bacterium *Escherichia coli* includes only 3,000 genes, whereas that of *D. melanogaster* contains at least 20,000 genes; the mouse genome contains 10 times that number. With added genes comes added complexity, because the genes form more intricate, interacting networks. Tracing the effect of any one gene in such an involved network is a formidable task.

Moreover, the larger size of multicellular organisms places practical limits on the number of individuals that can be included in a mutagenesis experiment. It is fairly simple and inexpensive to search for specific kinds of mutants among more than a billion mutagenized bacteria or yeast. In contrast, screening even 100,000 mutagenized fruit flies would constitute a large experiment. By comparison, the practical limits on screening mice for a particular mutation would be reached at about 1,000 animals.

The logistical difficulties of identifying and studying genes in multicellular organisms are further increased by the fact that most are diploid—their cells carry two copies of most genes, one inherited from the father and a second from the mother. For survival purposes, having two copies of most genes is valuable. If one copy acquires a harmful mutation, the other copy can usually compensate, so that no serious consequences result. Such redundancy, however, means that a mutation will elicit anatomical or physiological defects in the organism only if both copies of the gene are damaged. Investigators produce such individuals by mating parents who each carry the mutation in one copy of the gene. Approximately one fourth of the offspring of such matings will bear two defective copies of the gene. The need for matings introduces delays in the analysis.

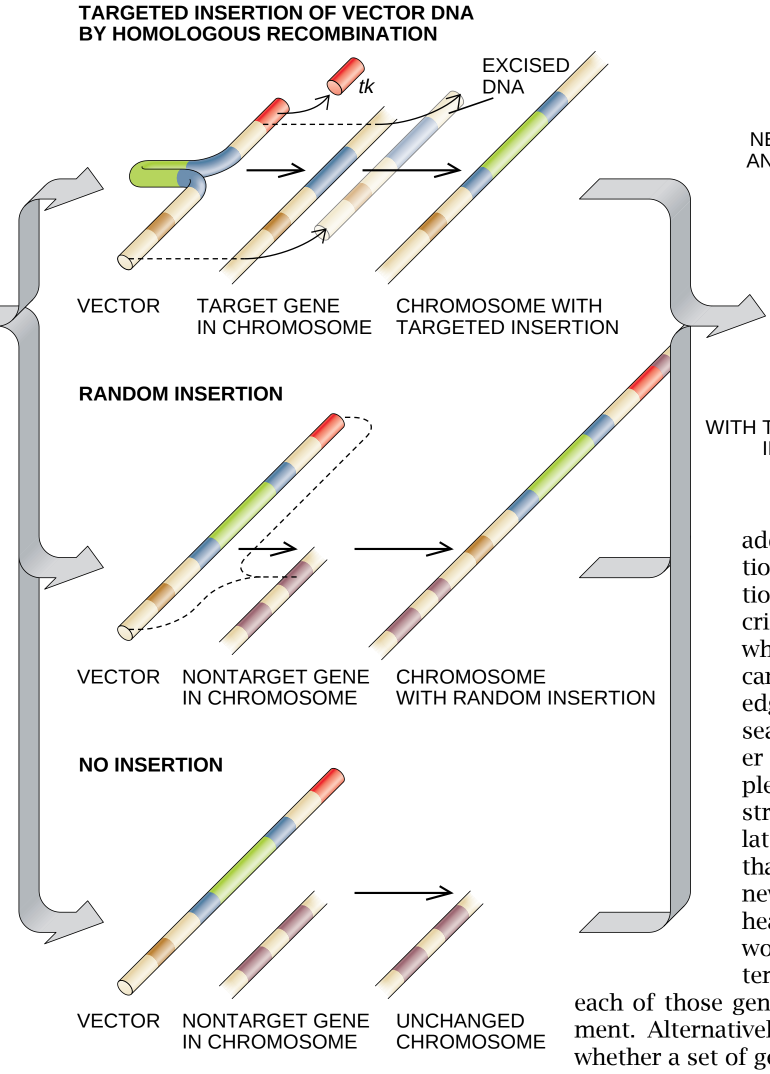
Despite the challenges, the identification of selected mutations in whole animals is unquestionably the most informative way to begin clarifying and separating the steps by which biological processes are carried out. Furthermore, if we want to understand processes that occur only in complex organisms, such as the mounting of a sophisticated immune response, such analysis must be pursued in those organisms. For these reasons, geneticists interested in mammalian development, neural function, immune response, physiology and disease have turned to the mouse. From a geneticist's point of view, the mouse is an ideal mammal. It is small and prolific and serves as a remarkably good analogue for most human biological processes.

On the other hand, the breadth of genetic manipulations that can be carried out in mice has been extremely limited relative to the operations that are possible in simpler organisms. Because of the obstacles I have already described, it is not practical to apply classical techniques to mice. To identify mutagenized mice carrying defects in the genes involved in some process of interest, researchers would have to screen 10,000 to 100,000 mice at a prohibitive cost. Instead mouse geneticists have historically studied mutant animals that arose spontaneously within their colonies. As a result of the keen observation and perseverance by such workers, the collection of existing mouse mutants is surprisingly large and is an invaluable resource for continued research.

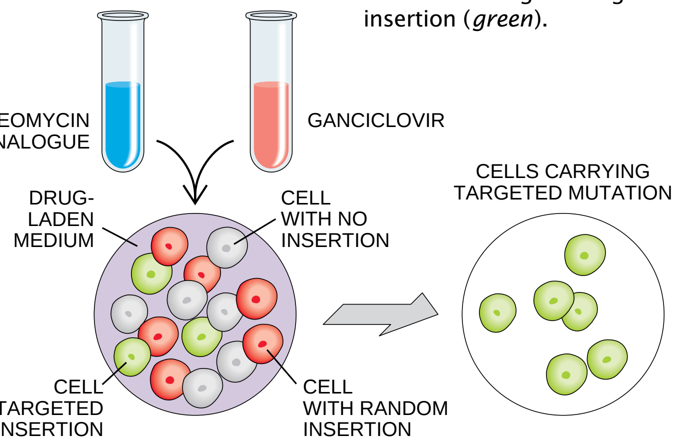
Yet even these hard-won animals have drawbacks. The existing collection of mutant mice does not harbor a random sampling of mutations in the mouse genome. Rather it contains a disproportionate number of mutations that result in readily observable abnormalities in physiology or behavior. In consequence, many mutations that affect coat color are present in this collection, whereas mutations that affect early development are underrepresented (since they often result in the undetected death of the embryo).

Further, the task of isolating the genes responsible for overt defects in

3. When all goes well, homologous recombination occurs (*top*): the vector lines up next to the normal gene (the target) on a chromosome in a cell, so that the identical regions are aligned; then those regions on the vector (together with any DNA in between) take the place of the original gene, excluding the marker at the tip (*red*). In many cells, though, the full vector (complete with the extra marker) fits itself randomly into a chromosome (*middle*) or does not become integrated at all (*bottom*).


TARGETED INSERTION OF VECTOR DNA BY HOMOLOGOUS RECOMBINATION
tk
EXCISED DNA
VECTOR
TARGET GENE IN CHROMOSOME
CHROMOSOME WITH TARGETED INSERTION
RANDOM INSERTION
VECTOR
NONTARGET GENE IN CHROMOSOME
CHROMOSOME WITH RANDOM INSERTION
NO INSERTION
VECTOR
NONTARGET GENE IN CHROMOSOME
UNCHANGED CHROMOSOME

4. To isolate cells carrying a targeted mutation, workers put all the cells into a medium containing selected drugs, here a neomycin analogue (G418) and ganciclovir. G418 is lethal to cells unless they carry a functional *neo^r^* gene, and so it eliminates cells in which no integration of vector DNA has occurred (*gray*). Meanwhile ganciclovir kills any cells that harbor a *tk* gene, thereby eliminating cells bearing a randomly integrated vector (*red*). Consequently, virtually the only cells that survive and proliferate are those harboring the targeted insertion (*green*).


NEOMYCIN ANALOGUE
GANCICLOVIR
DRUG-LADEN MEDIUM
CELL WITH NO INSERTION
CELLS CARRYING TARGETED MUTATION
CELL WITH TARGETED INSERTION
CELL WITH RANDOM INSERTION

mutant mice is very labor intensive, often taking years of concerted effort. Workers can deduce many steps involved in biological phenomena without ever finding the genes involved. But without isolating those genes, they cannot make progress at the molecular level. Notably, they cannot determine the nature of the proteins encoded by the mutated genes, nor can they identify the cells in which the genes are active.

Gene targeting allows investigators to circumvent such difficulties. Investigators now choose which gene to alter. They also have virtually complete control over how that gene is modified, so that the mutation can be tailor-made to address precise questions about the functions of the gene. The criteria for selecting which gene to mutate can be based on knowledge obtained from research on mice or other species. For example, it is now relatively straightforward to isolate a series of genes that are active in the newly forming mouse heart; gene targeting would then permit determining the role of each of those genes in heart development. Alternatively, we can ascertain whether a set of genes known to be involved in guiding the paths taken by developing neurons in *D. melanogaster* exist and serve a similar function in the mouse.

An initial approach often involves knocking out a gene in order to evaluate the consequences to the organism of not having the gene product. The consequences may be complex and may affect multiple pathways. Further insight into the gene's function can be obtained by introducing more subtle, defined mutations, which may affect only one of its multiple roles. Soon geneticists should be able to place genes under control of a switch. Such switches will allow researchers to turn a gene on and off during the embryonic or postnatal development of the mouse. For example, a hypothetical gene could be responsible for the creation and proper operation of a set of nerve cells. Knocking out the gene would result in the absence of those neurons during formation of the brain and preclude assessing the gene's activity in the adult. If the gene were under control of a switch, however, the switch could be left on during development, and the neurons would be formed. In the adult the switch could then be turned off, enabling workers to evaluate the function of this gene in adult neurons.

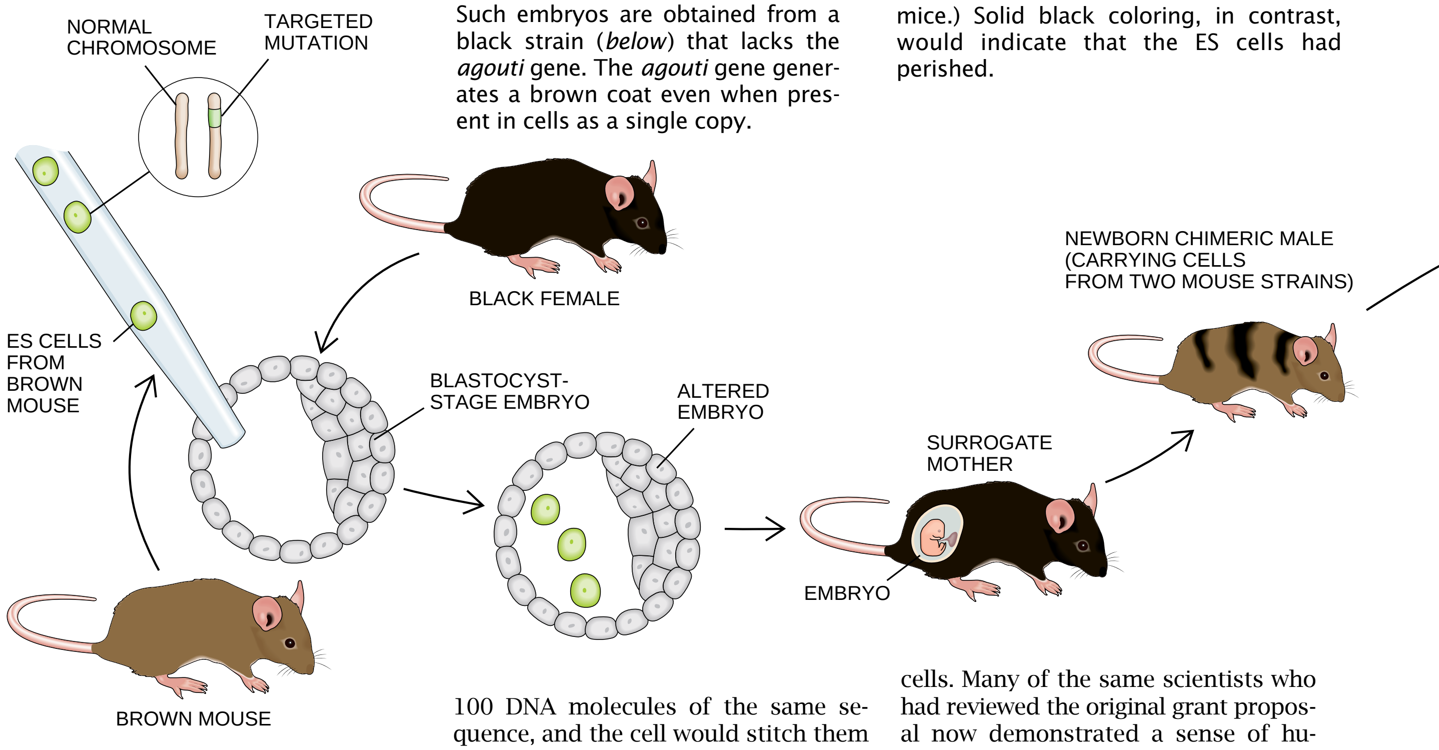
Development of gene-targeting technology has evolved over the past 15 years. In the late 1970s I was experimenting with using extremely small glass needles to inject DNA directly into the nuclei of mammalian cells. The needles were controlled by hydraulically driven micromanipulators and directed into nuclei with the aid of a high-powered microscope. The procedure turned out to be extremely efficient. One in three to five cells received the DNA in a functional form and went on to divide and stably pass that DNA on to its daughter cells.

When I followed the fate of these DNA molecules in cells, a surprising phenomenon captured my attention. Although the newly introduced DNA molecules were randomly inserted into one of the recipient cell's chromosomes, more than one molecule could be inserted at that site, and all of them were in the same orientation. Just as words in any language have an orienta-

How Targeted Gene Replacement Is Accomplished in Mice

1. Cells known as embryonic stem (ES) cells (*green at far left*) are isolated from a brown mouse strain and altered (by the process described in the illustration on pages 54 and 55) to carry a targeted mutation in one chromosome (*inset*). The ES cells are then inserted into young embryos, one of which is shown. Workers like to use the coat color of the future newborns as a guide to whether the ES cells have survived in the embryo. Hence, they typically put the ES cells into embryos that, in the absence of the ES cells, would acquire a totally black coat. Such embryos are obtained from a black strain (*below*) that lacks the *agouti* gene. The *agouti* gene generates a brown coat even when present in cells as a single copy.

2. The embryos containing the ES cells grow to term in surrogate mothers. Then workers examine the coats of the newborns. Brown shading intermixed with black indicates that the ES cells have survived and proliferated in an animal. (Such individuals are called chimeras because they contain cells derived from two different strains of mice.) Solid black coloring, in contrast, would indicate that the ES cells had perished.



tion (in English we read words from left to right), so, too, do DNA molecules. Apparently, before cells performed random insertion, some mechanism in the cell nucleus stitched virtually all the introduced DNA molecules together in the same orientation.

We went on to demonstrate that cells used a process called homologous recombination to achieve such linkages. Homologous recombination works only on DNA molecules with the same nucleotide sequence. Such molecules line up next to each other. Then both molecules are cut and are joined to each other at the cut ends. The joining is accomplished with such precision that the nucleotide sequences at the points of linkage are not altered.

This unexpected observation implied that all mouse cells, and presumably all mammalian cells, had the machinery to perform homologous recombination. At the time, there was no reason to suspect that somatic cells (those not involved in sexual reproduction) would have this machinery. Further, we knew the machinery was fairly efficient because we could microinject more than 100 DNA molecules of the same sequence, and the cell would stitch them all together in the same orientation. I realized immediately that if we could harness this machinery to carry out homologous recombination between a newly introduced DNA molecule of our choice and the same DNA sequence in a cell's chromosome, we would have the ability to rewrite the cell's instruction manual at will.

Excited by this prospect, in 1980 I requested funding from the government to test the feasibility of gene targeting. To my disappointment, the scientists who reviewed the grant proposal rejected it. In their view, the probability that the newly introduced DNA sequence would ever find its matching sequence within the 1,000 volumes of the genetic instruction manual seemed vanishingly small.

Despite the rejection, I decided to forge ahead using funds I was receiving for another project. It was a gamble. Had the experiments failed, I would have had little meaningful data to submit at grant renewal time. Fortunately, the experiments worked. By 1984, when we again asked for funds to pursue the research, we had ample evidence that gene targeting was in fact feasible in cells. Many of the same scientists who had reviewed the original grant proposal now demonstrated a sense of humor. The critique of the new proposal opened with the statement, "We are glad that you didn't follow our advice."

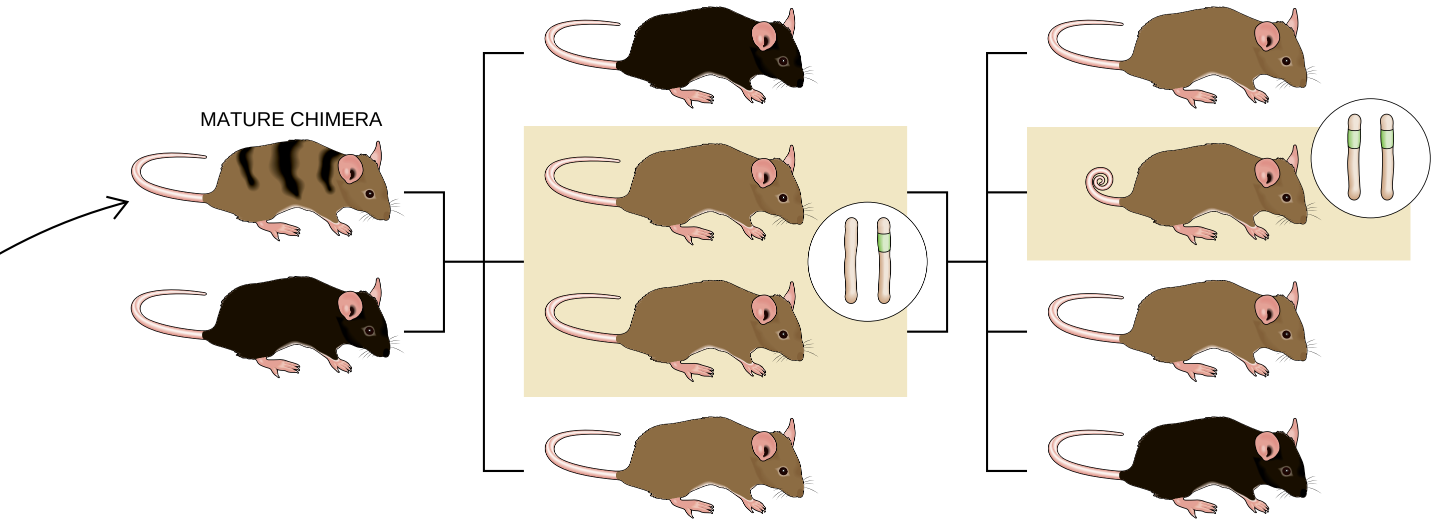
How is gene targeting in cells accomplished? The first step is to clone the gene of interest and propagate it in bacteria. This procedure provides a pure source of DNA containing the gene. Next, in a test tube, the nucleotide sequence of the gene is changed to meet the purpose of the experiment. The altered gene is referred to as the targeting vector.

The targeting vector is introduced into living cells by any of several means. Once within the cell nucleus, it forms a complex with proteins constituting the cell's machinery for homologous recombination. Aided by these proteins, it searches through all the sequences of the genome until it finds its counterpart (the target). If it indeed does find its target, it will line up next to that gene and replace it.

Regrettably, such targeted replacement occurs only in a small fraction of the treated cells. More often, the targeting vector inserts randomly at nonmatching sites or fails to integrate at

3. Chimeric males are mated to black (non-*agouti*) females. Then researchers screen the progeny for evidence of the targeted mutation (*green in inset*) in the gene of interest. They exclude black mice immediately; if the animals had been born of sperm made by ES cells—and so had a chance of harboring the chosen mutation—they would be brown. Direct examination of the genes in the brown mice reveals which of those animals (*boxed*) inherited the targeted mutation.

4. Males and females carrying the mutation are mated to each other to produce mice whose cells carry the chosen mutation in both copies of the target gene (*inset*) and thus lack a functional gene. Such animals (*boxed*) are identified definitively by direct analyses of their DNA. Then they are examined carefully for any physical or behavioral abnormalities.



all. We must therefore sort through the cells to identify those in which targeting has succeeded. Approximately one in a million treated cells has the desired targeted replacement.

To greatly simplify the search for that cell, we make use of two "selectable markers," which are introduced into the targeting vector from the start. Inclusion of a "positive" selectable marker promotes survival and growth of cells that have incorporated the targeting vector, either at the target site or at a random location within the genome. Inclusion of the "negative" selectable marker helps to eliminate most of the cells that have incorporated the targeting vector at a random location.

The positive marker, usually a *neomycin-resistance* (*neo*^{r}) gene, is positioned so that it will be flanked by DNA also present in the target gene. The negative marker, typically the *thymidine kinase* (*tk*) gene from a herpesvirus, is attached to one end of the targeting vector [*see illustration on pages 54 and 55*]. When homologous recombination occurs, the unchanged segments of the cloned gene, together with the *neo*^{r} gene sandwiched between them, replace the target sequence in the chromosome. But the *tk* gene, lying outside the zone of matching sequences, does not enter the chromosome and is degraded by the cell. In contrast, when cells randomly insert the targeting vector, they stitch the entire vector, complete with the *tk* gene, into the DNA. When no insertion occurs, the vector and both its markers are lost.

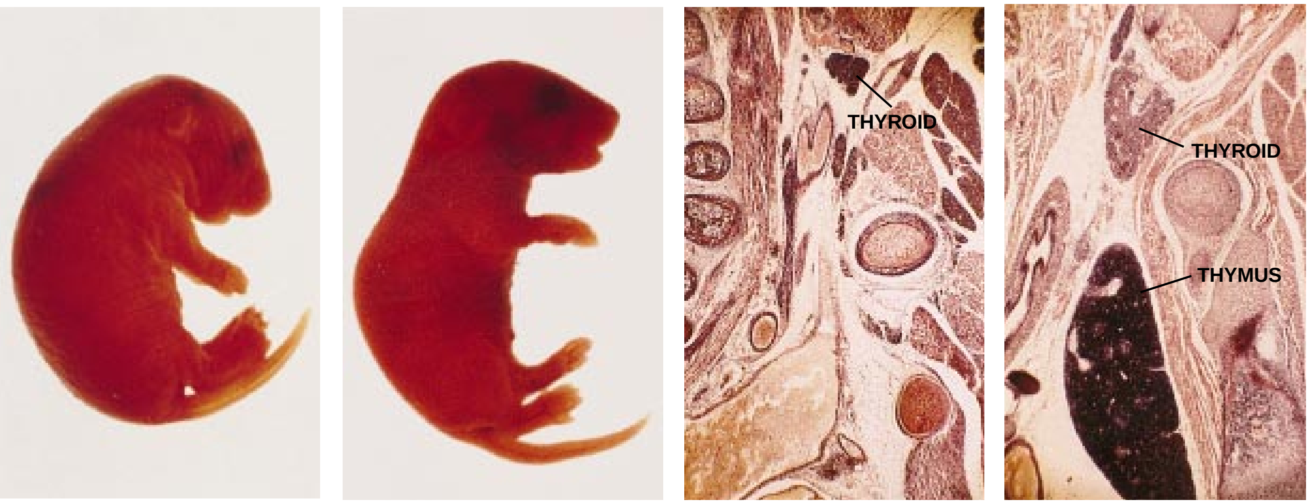
We do not have to examine the DNA directly to identify these different outcomes. Instead we grow the cells in a medium containing two drugs, an analogue of neomycin called G418 and the antiherpes drug ganciclovir. G418 kills cells that lack the protective *neo*^{r} gene in their chromosomes, namely, those that have failed to integrate vector DNA. But it allows cells that carry either random or targeted insertions to survive and grow. Concurrently the ganciclovir kills any cells that carry the herpes *tk* gene, namely, those that harbor a random insertion. In the end, virtually the only surviving cells are those bearing the targeted insertion (cells possessing the "positive selectable" *neo*^{r} gene and lacking the "negative selectable" *tk* gene).

By 1984 we had shown that it was possible to target specific genes in cultured mouse cells. We were then ready to extend the technology to alter the genome of living mice. To accomplish this aim, we used special cells developed in 1981 by Matthew H. Kaufman and Martin J. Evans of the University of Cambridge. These cells are embryo-derived stem (ES) cells. Such cells are obtained from an early mouse embryo. They can be cultured in petri dishes indefinitely, and they are pluripotent: capable of giving rise to all cell types.

In brief, by the procedure described earlier, we produce ES cells known to carry a targeted mutation in one copy of a chosen gene. Then we put the ES cells into early mouse embryos, which are allowed to develop to term. Some of the resulting mice, when mature, will produce sperm derived from the ES cells. By mating such mice to normal mice, we generate offspring that are heterozygous for the mutation—they carry the mutation in one of the two copies of the gene in every cell.

These heterozygotes will be healthy in most instances, because their second, undamaged copy of the gene will still be functioning properly. But mating of these heterozygotes to brothers or sisters bearing the same mutation yields homozygotes: animals carrying the targeted mutation in both copies of the gene. Such animals will display abnormalities that will reveal the normal functions of the target gene in all their tissues.

Of course, the procedure is more easily summarized than carried out. To actually do the work, we begin by injecting our modified ES cells into blastocyst-stage embryos, which have not yet become attached to the mother's uterus. Because we depend on coat color to indicate whether the procedure is going according to plan, we choose blas-



NEWBORN mouse (*above, left*) carries a targeted mutation in both copies of a gene called *HoxA-3.* Consequently, its body is more curved than that of a normal newborn (*second from left*). Tissue specimens from mutant (*center right*) and normal (*far right*) mice reveal that such mutants also lack a thymus and have an unusually small thyroid gland. These disorders and others indicate that the *HoxA-3* gene is needed for development of tissues and organs that originate in a narrow strip of cells (*colored band in drawing*) present in young embryos.

tocysts that would normally develop into pups bearing a different coat color than is found in pups produced by the mouse strain from which the ES cells are obtained.

The stem cells are isolated from a brown mouse carrying two copies of the *agouti* gene. This gene, even when present in a single copy, produces brown coloring by causing yellow pigment to be laid down next to black pigment in the hair shaft. (Production of the pigments themselves is under the control of other genes.) Hence, we typically select blastocysts that would normally develop into black mice. (Mice acquire black coats when the *agouti* gene inherited from both parents is defective.) Then we allow the embryo, containing the modified ES cells, to grow to term in a surrogate mother.

If all goes well, the altered ES cells reproduce repeatedly during this time, passing complete copies of all their genes to their daughter cells. These cells mix with those of the embryo and contribute to the formation of most mouse tissues. As a result, the newborns are chimeras: they are composed of cells derived both from the foreign ES cells and from the original embryo. We readily identify such chimeras by observing broad swatches of brown coloring in their otherwise black coats. If the animals bore no ES-derived cells, they would be completely black because of their lack of functional *agouti* genes.

By merely looking at the chimeras, though, we cannot determine whether the ES cells gave rise to germ cells, the vehicle through which the targeted mutation is passed to future generations. We find that out only when we move to the next stage: producing heterozygous mice harboring one copy of the mutation in all their cells. To generate such animals, we mate chimeric male mice to black female mice lacking the *agouti* gene. An offspring will be brown if the sperm that fertilized the egg was derived from ES cells (because all such sperm carry the *agouti* gene). An offspring will be black if the sperm derived from the original blastocyst cells (which lack functional *agouti* genes).

Consequently, when we see brown pups, we know that the genes carried by ES cells made their way to these offspring. We can then think about setting up matings between heterozygous siblings in order to produce mice with two defective copies of the target gene. First, though, we must discern which of the brown pups carry a copy of the mutated gene. This we do by examining their DNA directly for the targeted mutation. When matings are set up between heterozygous siblings, one in four of the offspring will have two defective copies of the gene. We pick out the homozygotes by again analyzing DNA directly, this time looking for the presence of two copies of the targeted mutation. These animals are then examined carefully for any anatomical, physiological or behavioral anomalies that can provide clues to the functions of the disrupted gene. The total procedure from cloning a gene to generating mice with a targeted mutation in that gene takes approximately one year.

Laboratories all around the world are now applying gene targeting in mice to study an array of biological problems. Since 1989, more than 250 strains carrying selected genetic defects have been produced. A few examples of the emerging findings should illustrate the kinds of insights these animals can provide.

In my own laboratory, we have been exploring the functions of homeotic, or *Hox,* genes. These genes serve as master switches ensuring that different parts of the body, such as the limbs, the organs, and parts of the head, form in the appropriate places and take on the correct shapes. Studies of homeotic genes in *Drosophila* have yielded important clues to their activities [see “The Molecular Architects of Body Design,” by William McGinnis and Michael Kuziora; SCIENTIFIC AMERICAN, February]. Yet many questions remain. For instance, *D. melanogaster* has only eight *Hox* genes, whereas mice and humans each have 38. Presumably, expansion of the *Hox* family played a critical part in the evolutionary progression from invertebrates to vertebrates, supplying extra machinery needed for a more complex body. Precisely what do all 38 genes do?

Before gene targeting became available, there was no way to answer these questions, because no one had found mice or humans with mutations in any of the 38 *Hox* genes. My colleagues and I are now embarking on a systematic effort to establish the function of the individual *Hox* genes. Later we will attempt to identify how these genes form an interactive network to direct the formation of our bodies.

As part of this program, we have discovered that targeted disruption of the *HoxA-3* gene leads to multiple defects. Mice carrying two mutated copies of the gene die at birth from cardiovascular dysfunction brought on by incomplete development of the heart and the major blood vessels issuing from it. These mice are also born with aberrations in many other tissues, including the thymus and parathyroid (which are missing), the thyroid gland, the bone and cartilage of the lower head, and the connective tissue, muscle and cartilage of the throat.

These abnormalities are diverse but share one striking commonality: the affected tissues all descend from cells that were originally clustered in a narrow zone in the upper part of the developing embryo. The rudiments of the heart, for instance, are located in this region before the heart takes up its more posterior location in the chest. It seems, then, that the assignment of the *HoxA-3* gene is to oversee construction of many of the tissues and organs that originate in this narrow region.

Unexpectedly, the disorder produced by knocking out the mouse *HoxA-3* gene mimics that found in an inherited human disease known as Di George syndrome. Chromosomal analysis of patients shows that the human *HoxA-3* gene is not the culprit; victims display genetic damage on a chromosome distinct from that housing *HoxA-3.* We now know, however, that the gene responsible for the syndrome acts by interfering either with activation of the *HoxA-3* gene or with the events set in motion by the *HoxA-3* gene. Also, a mouse model for the disease is now available and may eventually provide clues to treatment. This unanticipated benefit underscores once again the value of basic research: findings born of curiosity often lead to highly practical applications.

Like developmental biologists, immunologists have also benefited from gene targeting. They are now applying this technology to decipher the individual responsibilities of well over 50 genes that influence the development and operation of the body's two foremost classes of defensive cells—*B* and *T* lymphocytes.

Cancer researchers are excited by the technique as well. Often investigators know that mutations in a particular gene are common in one or more tumor types, but they do not know the normal role of the gene. Discovery of that role using our knockout technology can help to reveal how the mutant form of the gene contributes to malignancy.

The *p53* tumor suppressor gene offers a case in point. Tumor suppressor genes are ones whose inactivation contributes to the development or progression of cancer. Mutations in the *p53* gene are found in perhaps 80 percent of all human cancers, but until recently the precise responsibilities of the normal gene were obscure. The analysis of mice homozygous for a targeted disruption of *p53* indicated that *p53* probably acts as a watchdog that blocks healthy cells from dividing until they have repaired any damaged DNA that is present in the cell. Such damage often occurs in cells as a consequence of the frequent environmental insults to which we are subjected. The loss of functional *p53* genes eliminates this safeguard, allowing damaged DNA to be passed to daughter cells, where it participates in formation of cancers.

Many other diseases will be amenable to study by gene targeting. More than 5,000 human disorders have been attributed to genetic defects. As the genes and mutations for the disorders are identified, workers can create precisely the same mutations in mice. The mouse models, in turn, should make it possible to trace in detail the events leading from the malfunctioning of a gene to the manifestation of disease. A deeper understanding of the molecular pathology of the disease should permit the development of more effective therapies. Among the models now being constructed are mice with different mutations in the cystic fibrosis gene.

The study of atherosclerosis, a leading cause of strokes and heart attacks, is also beginning to involve gene targeting. In contrast to cystic fibrosis, atherosclerosis is not caused by mutations in a single gene. Defects in a number of genes combine with environmental factors to promote the buildup of plaque in arteries. Nevertheless, promising mouse models have been made by alterating genes known to be involved in the processing of triglycerides and cholesterol. I also anticipate that mouse models for hypertension, another culprit in heart disease and stroke, will soon be developed, now that genes thought to participate in its development are being identified.

As understanding of the genetic contribution to disease increases, so will the desire to correct the defects by gene therapy. At the moment, the techniques used for gene therapy rely on random insertion of healthy genes into chromosomes, to compensate for the damaged version. But the inserted genes often do not function as effectively as they would if they occupied their assigned places on the chromosome. In principle, gene targeting can provide a solution to this problem. Yet, before it can be used to correct the defective gene in a patient's tissue, investigators may need to establish cultures of cells able to participate in formation of that tissue in adults. Such cells, which like the ES cells in our studies are termed stem cells, are known to be present in bone marrow, liver, lungs, skin, intestines and other tissues. But research into ways to isolate and culture these cells is still in its infancy.

Before the technical hurdles to broad application of our methods in gene therapy are surmounted, gene targeting will find common usage in the study of mammalian neurobiology. Already mice have been prepared with targeted mutations that alter their ability to learn. As increasing numbers of neural-specific genes are identified, the pace of this research will surely intensify.

We can anticipate continued improvements in gene-targeting technology, but it has already created opportunities to manipulate the mammalian genome in ways that were unimaginable even a few years ago. To significantly aid in deciphering the mechanisms underlying such complex processes as development or learning in mammals, researchers will have to call on every bit of their available ingenuity, carefully deciding which genes to alter and modifying those genes in ways that will bring forth informative answers. Gene targeting opens a broad range of possibilities for genetic manipulations, the limitations of which will be set only by the creative limits of our collective imagination.

FURTHER READING

THE NEW MOUSE GENETICS: ALTERING THE GENOME BY GENE TARGETING. M. R. Capecchi in *Trends in Genetics,* Vol. 5, No. 3, pages 70–76; March 1989.

ALTERING THE GENOME BY HOMOLOGOUS RECOMBINATION. M. R. Capecchi in *Science,* Vol. 244, pages 1288–1292; June 16, 1989.

REGIONALLY RESTRICTED DEVELOPMENTAL DEFECTS RESULTING FROM TARGETED DISRUPTION OF THE MOUSE HOMEOBOX GENE *HOX*-1.5. O. Chisaka and M. R. Capecchi in *Nature,* Vol. 350, No. 6318, pages 473–479; April 11, 1991.

High-Speed Silicon-Germanium Electronics

The author has helped create electronic devices that outperform traditional silicon technology yet remain compatible with standard manufacturing methods

by Bernard S. Meyerson

Silicon-based logic chips are so much a part of modern life that the singing at a birthday party is as likely to come from the birthday card as from the party guests. Despite this pervasive success, some workers have been arguing for years that silicon technology is nearing its physical limits. The traditional technology, they fear, cannot attain higher speeds without being shrunk to a point at which the devices can no longer function. If so, continued progress in electronics would depend on finding an alternative to silicon, and the giant electronics industry would face a difficult and costly period of transition. Is silicon really headed into its twilight years, as so many writers and researchers have claimed?

To paraphrase Mark Twain, the reports of the demise of silicon technology have been greatly exaggerated. In collaboration with a team at IBM Research and Manufacturing, I have explored methods for extending the performance of silicon technology by modifying the composition of the chips. In the past, the stunning improvements in the speed and versatility of silicon electronics have mostly resulted from the miniaturization of circuitry. We have concentrated instead on using materials that sharply increase the velocity of electrons through the electronic devices, thereby offering an alternative route to faster performance. Our efforts have demonstrated that an alloy of silicon and germanium, two well-known semiconducting elements, can form the basis of exceptionally high speed transistors. Transistors are the simple switches that lie at the heart of modern electronics.

These newly developed devices attain switching speeds once thought to lie beyond the capabilities of silicon technology. Moreover, they can be produced on existing chip-fabrication lines, preserving the multibillion-dollar investment such facilities represent. For these reasons, I expect that silicon technology will continue to dominate electronics design—indeed, I predict that it will soon reclaim functions that today have been abandoned to other, more exotic materials.

In concert with a circuit design team from Analog Devices, IBM has just announced the first commercial products that incorporate high-performance silicon-germanium transistors. In the next few years such transistors and other devices based on silicon-germanium alloy will probably find a home in a wide range of products, including personal communications gear and electronic signal converters, which draw digital data from the network of fiber-optic cables rapidly proliferating across the country.

A fundamental factor controlling the capability of computers and other electronics is the operating speed of their component devices. Over the past few decades miniaturization has been the key to faster performance. The most basic device in modern electronic circuits is the transistor, which functions as a simple on-off switch. That switching action forms the basis of the digital computer. A closer look at the design of transistors clarifies how they operate and why shrinking them improves their function. Such an examination also elucidates why the size-reduction process cannot be extended indefinitely.

Silicon-based electronics incorporates two main kinds of transistors: bipolar transistors and field-effect transistors. In a field-effect transistor, electric current enters from a location known as the source and exits via the drain. The region through which the current traverses the device is called the channel. Part of the transistor, termed the gate, controls the flow of current through the channel. The gate does its job by creating an electric field that can either fill or empty the channel of charge, thereby starting or stopping the flow of current. When current can move through the channel, the device is on; when no current flows, it is off.

The field-effect transistor has a significant advantage: it consumes a modest amount of power. To operate such a transistor, one need only charge the gate to some critical voltage. Once the gate contains an adequate charge, no further current is necessary to maintain the transistor in its on or off state. Except during switching, therefore, field-effect transistors draw essentially no power. For this reason, these devices are ideally suited for applications

BERNARD S. MEYERSON has extensively explored the potential uses of nontraditional semiconducting materials. He received his Ph.D. in physics from the City University of New York in 1981. After graduation, he joined the staff of the IBM Thomas J. Watson Research Center, where he now manages the electronic materials group in the center's physical sciences and technology department. In 1992 he was appointed an IBM Fellow. Meyerson has explored techniques for growing crystal alloys at unusually low temperatures, which has opened up new possibilities in semiconductor design. This work led him to invent the ultrahigh vacuum/chemical vapor deposition technique, the method that permitted the successful fabrication of high-speed silicon-germanium transistors.

INNOVATIVE TRANSISTOR seen in this micrograph consists of a mix of silicon and silicon-germanium alloy. Four years ago this device showed that the new technology can greatly surpass the speed of traditional silicon electronics.

that require low power consumption: in portable computers, for example, not to mention in the singing birthday card.

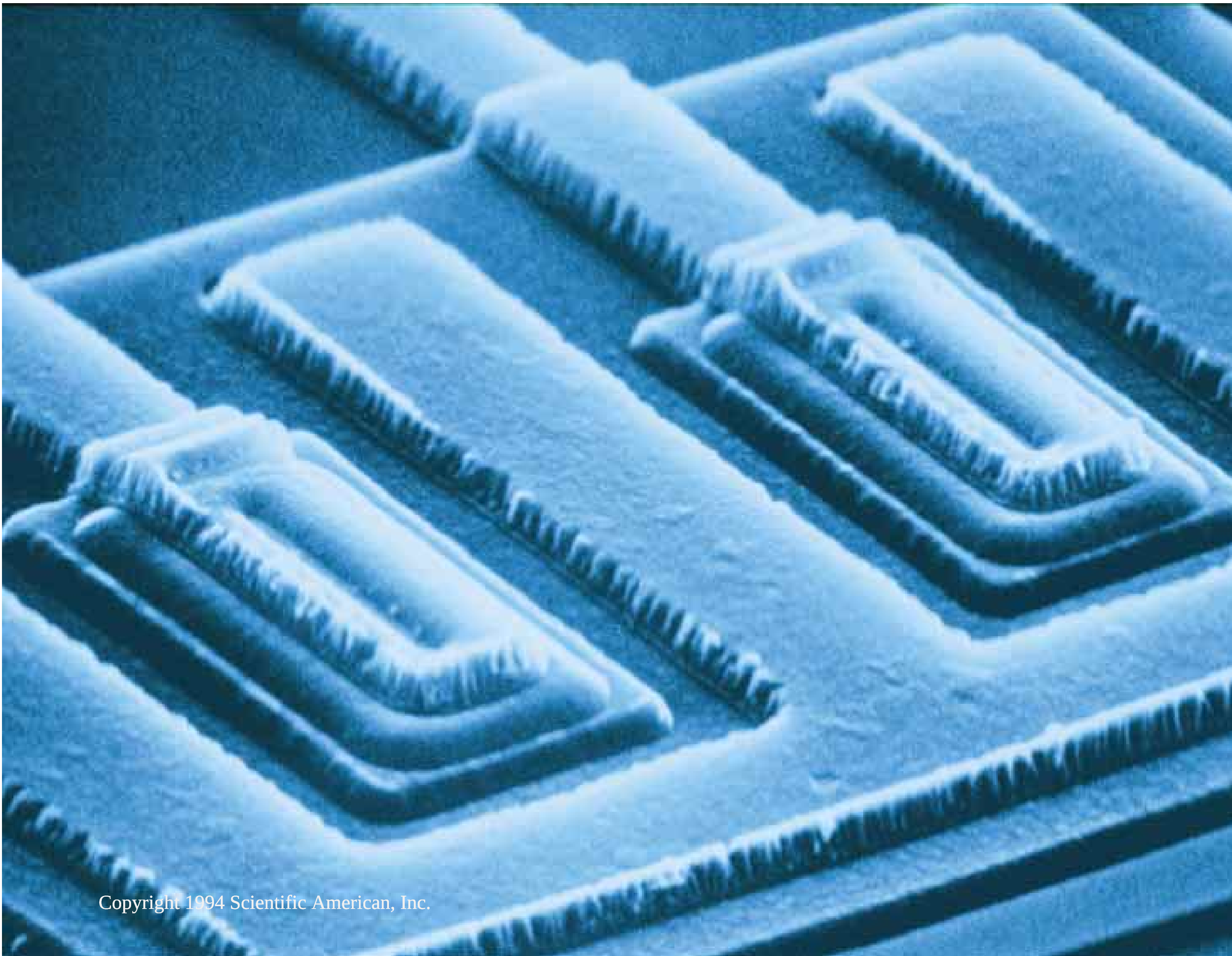
High-performance computers, in contrast, have relied primarily on bipolar transistors, which can operate significantly faster but consume a good deal more power. In the most common class of bipolar transistors, the NPN (negative-positive-negative) device, electrons flow from the emitter to the collector regions. When the transistor is activated, a small current is injected into the base region of the transistor. That current lowers a constant, built-in energy barrier that blocks the flow of electrons. As the barrier drops, current begins to pass through the transistor, and the device switches to an on state. The amount of current moving through the device is proportional to, but much larger than, the amount injected into the base.

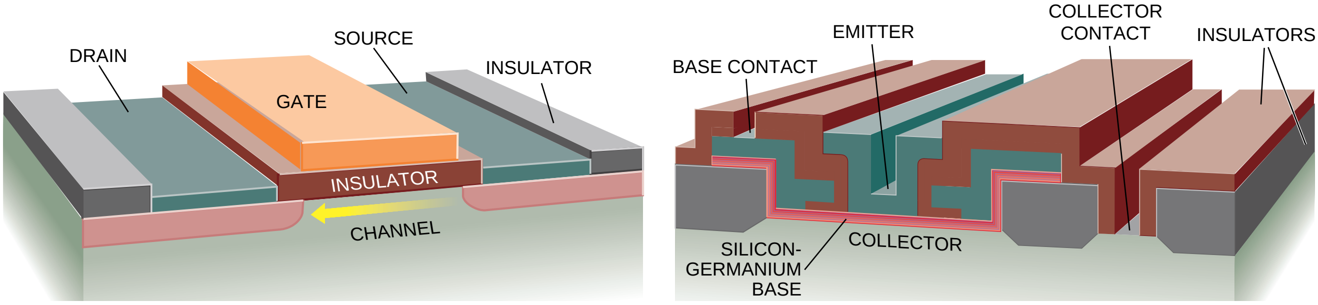
The base of a bipolar transistor must contain a constant electrical charge sufficient to keep the energy barrier up (so that the transistor stays off) when no current is applied to the base. Engineers create such built-in charges by introducing specific impurities, or dopant atoms, into the silicon when the transistor is fabricated. Depending on the nature of the dopant atoms, they add a net positive or negative charge to the silicon. Doped silicon is known as *n*-type silicon if it contains an excess of negative charges or *p*-type silicon if positive charges dominate.

The function of a bipolar transistor depends on the electrical properties at the interface between *n*-type and *p*-type silicon. An interface between two regions of semiconductor having the same basic composition—silicon, in most cases—but opposing types of doping is called a homojunction. The meeting of two dissimilar materials is known as a heterojunction. Homojunctions are far more easily fabricated, so they dominate in present circuit designs.

The size of a transistor, whether field effect or bipolar, profoundly influences its speed of operation. A major factor limiting the speed of a bipolar transistor, for example, is the time required for electrons to move across the base. A decrease in the thickness of the base reduces the distance that electrons must travel and so increases the rate at which the transistor can switch on and off. A thinner base, in turn, makes it possible to shrink the surface area of the entire transistor. The smaller area permits a tighter packing of the transistors, which improves the performance of the overall chip by diminishing the distances that electrical signals must traverse between transistors. Size reduction leads to similar improvements in the performance of field-effect transistors.

For these reasons, technologists have focused on creating ever smaller transistors. This strategy for enhancing the function of an electronic device by reducing its critical dimensions is commonly referred to as scaling. Although scaling has led to many years of im-





FIELD-EFFECT TRANSISTOR (*left*) operates by modulating the intensity of the electric field in the gate. Depending on the design of the transistor, that field can either enable or halt the flow of current through the channel, from the source to the drain. When current passes through the channel, the transistor is in its "on" state. A bipolar transistor (*right*) switches on when electrons move from the emitter to the collector. A built-in energy barrier in the base prevents that movement. When current is injected into the base, the barrier drops, and electrons flow through the transistor. The bipolar transistor shown here incorporates a silicon-germanium layer that accelerates the electrons and speeds the operation of the device.

provement in the speed and flexibility of silicon-based electronics, the trend cannot continue indefinitely. A consideration of the nature of bipolar transistors reveals the reason for this limit.

Scaling reduces the thickness and volume of the transistor base, so the dopant density must rise to keep the total base charge constant. In essence, one must squeeze the same total numbers of dopant atoms into an ever smaller volume base region. Unfortunately, a silicon homojunction having very high doping levels on both sides will necessarily leak current. If one continues to scale an NPN bipolar transistor, the density of dopant in the base eventually reaches the level at which current passes through the transistor even when it should be in its off state, rendering it useless.

Chip designs are beginning to push scaling to its natural conclusion, so researchers are actively pursuing alternative approaches to boost the speed of electronic devices. Indeed, the physical limit of scaling is a primary reason that some engineers have postulated the demise of silicon in favor of other, more exotic semiconducting materials, such as gallium arsenide. But industry has already invested tens of billions of dollars in tools and facilities for fabricating silicon-based devices. Clearly, it would be quite advantageous to find a path to faster performance that does not abandon silicon.

Much of my effort has centered on developing just such a revamped silicon technology. My present efforts actually build on an old idea. In the mid-1950s a number of researchers recognized that heterojunctions could, in principle, offer a way to quicken the switching rate of a transistor not by shrinking the device but by modifying its basic electronic properties. Naturally occurring electric fields in the two materials at a heterojunction can confine negative or positive charges on opposite sides of the interface. If the material at the junction changes gradually from one composition to the other, an extended electric field can be built into the graded region.

In the late 1950s Herbert Kroemer, now at the University of California at Santa Barbara, proposed using the field generated in a graded heterojunction to propel electrons rapidly across the base of a bipolar device. By inducing the electrons to travel more swiftly, heterojunctions could function much faster than homojunctions of a similar size. Kroemer envisioned several possible pairings of semiconductor materials that could speed the operation of transistors; the most promising pairing involved silicon on one side and germanium on the other. Although the idea appeared sound, the practical problem of building a workable silicon-germanium heterojunction proved to be monumental.

A promising development occurred during the 1960s, when researchers developed epitaxy, a technique that seemed well suited to such delicate fabrication tasks. In this process, layers of atoms are deposited onto an existing crystalline material. The underlying crystal, or substrate, serves as a template so that the newly accumulated layers follow the same atomic arrangement as the crystal itself.

Because silicon and germanium have the same crystal structure, a layer of one material can be deposited on the other, maintaining a consistent atomic order. But atoms in a germanium crystal have a natural spacing 4 percent greater than that of atoms in a silicon crystal. Germanium atoms would normally expand to their natural spacing, but when deposited on a much thicker silicon substrate, they are locked in place by the underlying silicon. A layer of germanium atoms deposited on top of a thick silicon substrate experiences tremendous strain, which mounts as additional layers accumulate.

Ultimately, defects form in the germanium to relieve the strain. When a defect occurs, entire rows of germanium atoms squeeze out of the lattice, thereby allowing the remaining germanium atoms in the layer to move apart from one another. Four of every 100 germanium atoms grown along a silicon-germanium juncture must work their way out of the lattice in order to bring the structure to a fully relaxed state. This exclusion of germanium atoms would result in several trillions of defects in an area the size of single chip, more than enough to prevent its proper function.

One way to reduce the strain in the crystal is to grow a silicon-germanium alloy, rather than layers of pure germanium, on the silicon substrate. Such an alloy has a characteristic atomic spacing intermediate between that of silicon and that of germanium. But great skill is still required to fabricate an alloy layer, because even a mixed silicon-germanium composition develops defects if the layer is too thick or if it is too rich in germanium.

The disparity between the atomic spacing in silicon and germanium crystals, known as lattice mismatch, proved exceedingly difficult to overcome. During the early 1980s, most of the efforts along those lines relied on a technique called molecular-beam epitaxy. In this approach, workers grow the crystal film in a steel chamber that has been evacuated to an internal pressure of less than one trillionth of an atmosphere. The silicon substrate is mounted in the chamber and is heated to a temperature of 1,100 degrees Celsius or more. Such searing temperatures evaporate contaminants from the silicon, leaving a clean surface on which a film can grow.

After the high-temperature cleaning, engineers allow the surface to cool somewhat and then deposit a buffer layer of pure silicon on the substrate to bury any residual contamination. Molten pools of silicon and germanium at the base of the apparatus provide a source of atoms; beams of those atoms are directed at the substrate to produce the desired film. The atoms strike the silicon substrate and accumulate in crystalline layers.

To minimize the strain that results from lattice mismatch, researchers concentrated on trying to build layers of silicon-germanium alloys containing less than 30 percent germanium. Molecular-beam epitaxy eventually allowed the fabrication of such moderately defect-free heterojunctions. But they were only good enough to serve as laboratory test beds.

Some researchers therefore gave up on molecular-beam epitaxy in favor of an alternative approach: chemical vapor deposition. Chemical vapor deposition utilizes gas molecules that incorporate the desired atoms—silicon and germanium, in this case. The flow of gas carries those atoms to the substrate surface, where they collect and form the new crystal layers. This technique, which has been well known for decades, is in many ways simpler than molecular-beam epitaxy.

The biggest drawback of chemical vapor deposition was that it required high temperatures: 1,100 degrees C during the initial cleaning and 1,000 degrees C while the films were growing. Under such intense heat, strained materials, such as silicon-germanium alloy, rapidly become defective. Furthermore, high temperatures make it impossible to place the dopant material accurately. At temperatures above 800 degrees C, dopant atoms in silicon or germanium quickly diffuse away from their initial position. Chemical vapor deposition could not produce usable heterojunctions between silicon and silicon-germanium so long as the process demanded such high temperatures.

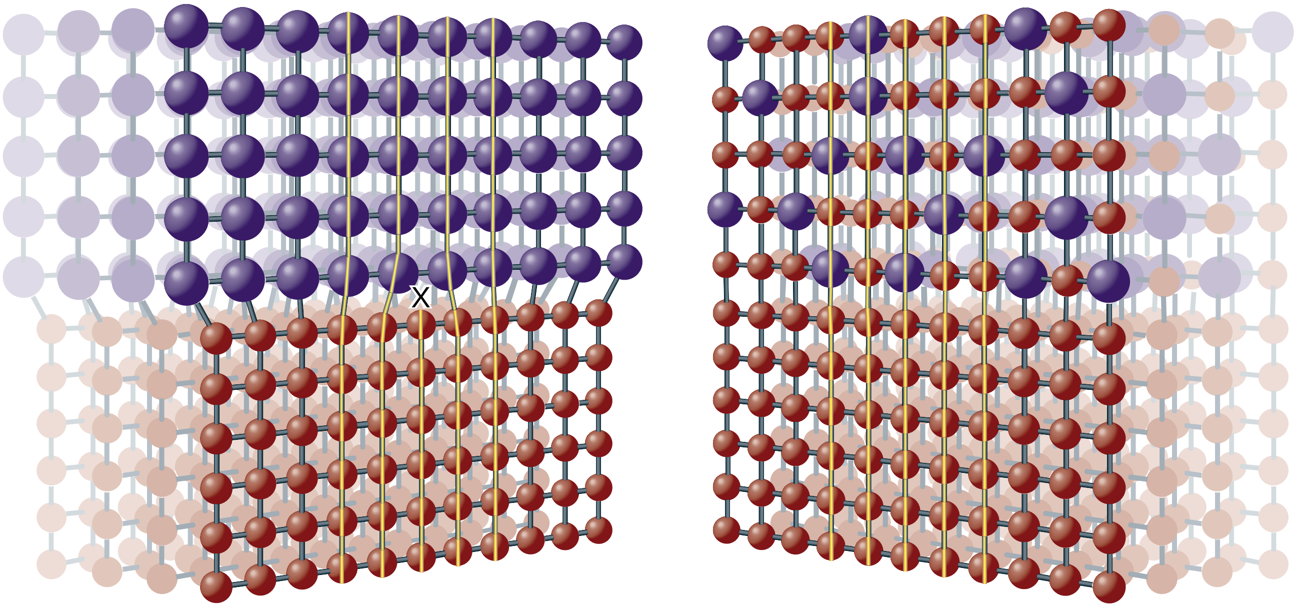
A number of researchers, including me, sought a way to perform chemical vapor deposition at lower temperatures. Our endeavor centered on the two essential steps in the process that seemed to demand high temperatures: cleaning the silicon surface before film growth and growing a defect-free film.

The primary purpose of the cleaning is to dislodge the silicon oxides that form when pure silicon comes in contact with air or moisture. Silicon oxides have no crystal structure, so they would interfere with epitaxy unless removed. Furthermore, such oxides absorb a silicon dopant, boron, out of the air, and trap it on the silicon surface at disastrously high concentrations. At room temperature, silicon oxide grows on the surface to a thickness of about 10 atomic layers. The oxide then acts as a barrier that protects the silicon below it from additional reaction with the air.

Scientists have long known that dipping a silicon wafer into hydrofluoric acid removes the coating of silicon oxide, but the conventional wisdom held that the oxides re-formed instantly when the silicon was again exposed to air. Consequently, all methods of epitaxy included a step in which the silicon was baked at high temperatures, even if the silicon oxide had already been chemically stripped using hydrofluoric acid. I suspected that the conventional wisdom was not correct.

As a graduate student, I had spent many hours handling silicon wafers and inevitably ended up dropping some of them. I noticed that when I rinsed off a wafer in water the wafer did not become wet; instead the water rolled off its surface. I knew that silicon oxide attracts water, so if the silicon wafer could not be made wet, then the oxide could not have been present. In many cases, however, several hours had passed since the wafers were cleaned in a bath of hydrofluoric acid. It seemed that the silicon oxide took quite some time to re-form.

After joining IBM, I examined the early literature in this field and found the source of the misconception that silicon oxide forms instantly. Years ago researchers using crude optical probes thought they had detected silicon oxide, when actually they had observed the thin coating of hydrogen that forms after silicon is cleaned in hydrofluoric acid. Modern, chemically selective probes have confirmed that silicon can



ATOMIC SPACING in a germanium crystal (*purple*) is slightly greater than it is in silicon (*red*). That lattice mismatch (*left*) has long frustrated engineers attempting to utilize the desirable electronic properties at junctions between silicon and germanium. Germanium atoms deposited onto silicon initially follow the underlying atomic arrangement. As the germanium atoms revert to their natural spacing, they give rise to defects (*marked by the "X"*) that create short circuits and ruin the junction. The author has succeeded in growing defect-free layers of silicon-germanium alloy on top of a silicon base (*right*). These hybrid crystals form the basis of a new class of high-speed electronic devices.

remain free of oxide for many hours after a hydrofluoric acid bath. The hydrogen layer that clings to the silicon protects the surface from the air, retarding oxide formation.

That protective layer obviates the need for a high-temperature preparatory step in epitaxy. In fact, hydrofluoric acid cleaning does not merely allow low temperatures, it demands them, because high temperatures would drive off the hydrogen layer.

The next challenge was to find a means by which to grow a high-quality film at relatively cool temperatures. Early work on silicon epitaxy indicated that the number of defects in the deposited films rises dramatically as the temperature drops. Impurities that are normally present during chemical vapor deposition—oxygen and water, in particular—incorporate themselves into the films far more readily at low temperatures than at high ones. These impurities can cluster together within a growing layer, causing defects in the material.

The obvious way to combat this problem is to minimize the concentration of foreign atoms in the chamber where vapor deposition takes place. Laboratory experiments indicated that growing uncontaminated films at temperatures below 700 degrees C would demand ultrahigh vacuum conditions, though not as severe as those needed for molecular-beam epitaxy.

How might possible sources of contamination from the interior of a crystal-growing apparatus be eliminated? At IBM, we managed to remove such contaminants with specialized vacuum pumps and tight seals. In this way, we can effect ultrahigh vacuum conditions in an inexpensive glass tube, which is surrounded by a furnace that supplies the heat required for epitaxy. Such an arrangement ensures that the apparatus does not contribute enough impurities to interfere with film growth. A special airlock makes it possible to load the main growth chamber without exposing it to the surrounding atmosphere. This feature is important because contaminants from the air cling tenaciously to the inside of the chamber; it takes considerable time to bake them off and pump them out.

The gases containing the silicon and germanium, which flow through the apparatus during chemical vapor deposition, can be another source of contaminants. We keep the pressure in the reactor quite low, typically one millionth of an atmosphere, to minimize the quantity of foreign material that enters in this way. We therefore call our approach the ultrahigh vacuum/chemical vapor deposition technique.

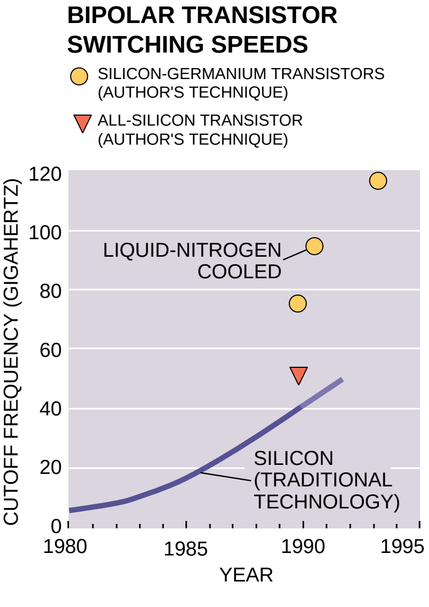
Because these procedures establish an extremely clean environment within the growth chamber, we can run the furnace at temperatures well below those used in conventional epitaxy. My group at IBM has found that temperatures of 400 to 500 degrees C suffice for preparing high-quality silicon and silicon-germanium alloy films.

The ability to fabricate a heterojunction at such moderate temperatures permits the creation of sophisticated and flexible chip designs. For instance, one can grow the silicon-germanium layer on top of a silicon wafer that already contains all the proper chemical regions for the electronic devices that the chip will eventually bear. Such imprinted chemical patterns make it possible to build a chip having an exceptionally high density of transistors or other devices. The temperatures used in traditional epitaxy would severely distort any such preexisting patterns.

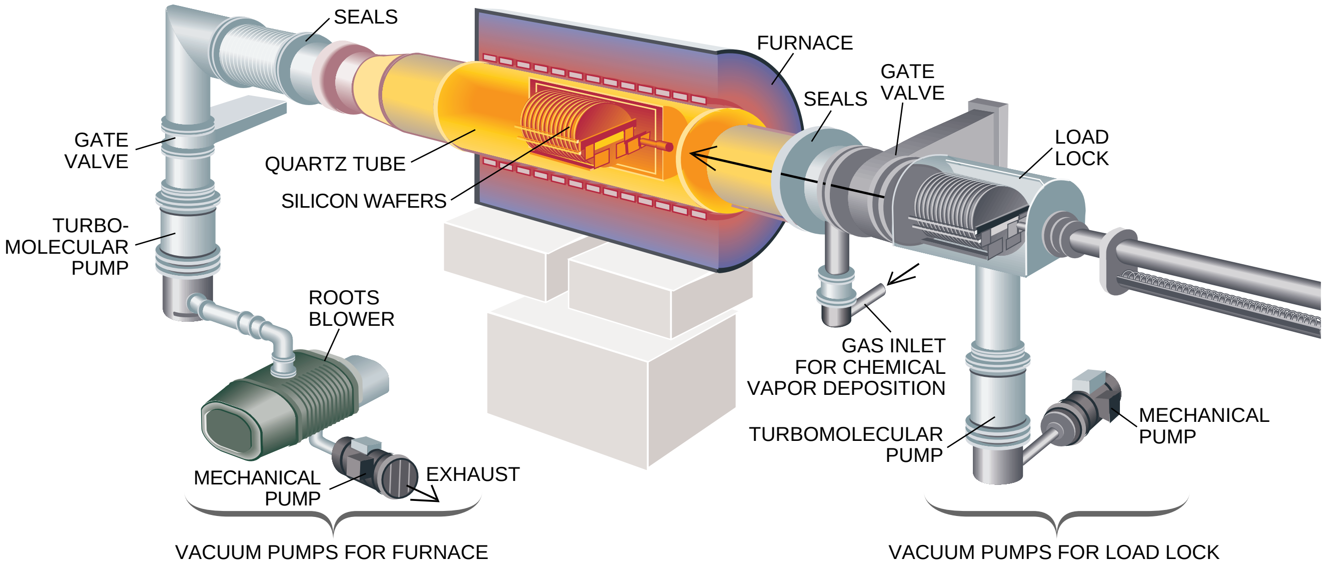
I had done much of the initial development of the ultrahigh vacuum/chemical vapor deposition process on my own. Then, beginning in 1988, I joined forces with a number of other specialists to tackle the much more difficult task of using the process to build workable, high-speed bipolar transistors. We went around beating on the doors of colleagues known to have a sense of adventure, not to mention a sense of humor—a requirement to survive the inevitable disasters that occur in such unproved work. In the end we pulled together a team that went on to break every existing performance record in silicon technology.

We started by using low-temperature epitaxy to build pure silicon homojunctions. These simple devices functioned perfectly, demonstrating the soundness of our technique. Starting in 1989, a group of IBM researchers fabricated the first in a series of NPN bipolar transistors that realized Kroemer's concept of a graded heterojunction between silicon and silicon-germanium alloy. Although the alloy contained less than 4 percent germanium, these transistors already exceeded the supposed capabilities of silicon technology. The built-in electric field (approximately 30,000 volts per centimeter across the base) in these devices accelerated electrons so strongly that they traversed the base of the device in half the time required in conventional, all-silicon transistors.

A standard measure of the performance of a bipolar transistor is the manner in which the gain of the transistor (the ratio of the current switched by the transistor to the current needed to turn the transistor on) depends on the frequency of switching. In a typical application in a computer, a conventional bipolar transistor might have a gain of about 100. At higher switching frequencies, the gain becomes progressively smaller. When the gain drops to one, the transistor becomes useless, because the current that must be put in to turn it on equals the current that comes out; in that case, the transistor func-



SWITCHING SPEEDS of silicon bipolar transistors have steadily improved over the years (*left*). Silicon-germanium technology accelerates that trend, yielding levels of performance previously considered impossible for silicon-based electronics. An innovative crystal-growing system (*right*) brought the new technology to fruition.



CRYSTAL-GROWING TECHNIQUE invented by the author avoids the destructively high temperatures utilized in earlier approaches. Vacuum pumps, special valves and seals prevent contaminants from entering the furnace, where a silicon-germanium alloy is deposited onto the silicon wafers. The actual deposition process occurs in a near vacuum, thereby minimizing the concentrations of foreign atoms that can interfere with the proper operation of electronics.

tions as no more than a simple wire.

Technologists gauge the speed of a transistor in terms of how rapidly it can switch on and off before its gain drops to one. The first graded heterojunctions that we built in 1989 switched at 75 gigahertz (billions of cycles per second), nearly twice the speed of the fastest comparable silicon devices. More recent work at IBM has pushed the heterojunctions to speeds in the range of 110 to 117 gigahertz, a level of performance previously considered impossible using silicon. In follow-up experiments my colleagues and I have incorporated these devices into complete circuits that operated at record-fast speed. That was a crucial test, because high-speed transistors often yield much lower practical performance speeds when wired into real circuits.

A collaboration between IBM and Analog Devices is now bringing such circuits to the marketplace. At last year's International Electron Devices Meeting in Washington, D.C., Analog Devices reported that it will soon begin selling silicon-germanium-based circuitry, including a digital-to-analog converter, a staple of home electronics. The silicon-germanium converter transforms numerical data into electronic output at a record-fast rate of one billion conversions per second. It matches the speed of the best such circuits built using gallium-arsenide junctions and yet operates on a fraction of the power they require.

The appearance of a commercially viable, silicon-germanium integrated circuit marks a milestone in the effort to find paths to higher performance other than scaling. Engineers at Analog Devices are already considering additional applications for the silicon-germanium devices, such as digital cordless telephones that can handle an unusually rapid data flow. Digital-to-analog converters are essential for translating the digital data from an optical fiber into analog signals for a telephone or television. Faster digital-to-analog circuits may therefore hasten the arrival of digital data networks into both home and business. Such converters will also lie at the heart of the portable communications devices that are likely to become increasingly prevalent.

At present, silicon-germanium technology is still in its infancy. Designers need to modify many existing circuit designs to exploit fully the speed of the new devices. To date, IBM is the only company that has demonstrated the ability to integrate a significant number of high-performance heterojunction bipolar transistors into circuits. My group has shown that silicon-germanium materials can boost the performance of field-effect transistors as well, but we have yet to integrate such devices into larger circuits. Eventually the technology should make it possible to combine multiple functions (transmitter, signal converter, receiver) onto a single chip. In this way, all kinds of exotic gear, such as "Dick Tracy"–style two-way wrist televisions, could pass from fantasy to reality.

Leybold-A.G. recently began manufacturing a commercial version of our ultrahigh vacuum/chemical vapor deposition apparatus. Now that they have standardized equipment to work with, engineers can focus on developing increasingly complex circuits and finding ways to broaden the variety of silicon-germanium heterojunction devices that may be combined onto a single chip. It is only a matter of time—and probably not much time—before many novel electronic devices come to fruition.

Earlier, facing the limits of scaling, many researchers had mistakenly concluded that silicon was reaching the end of its life as an electronic material. In the heyday of this view, I saw a poster hanging on the wall of a colleague's office. It depicted an automobile's taillights disappearing into a cloud of tire smoke; the caption read, "Gallium arsenide leaves silicon in the dust." Quite the contrary. I would say our team has shown that silicon is still very much in the race, and wagering against its long-term success would be a bad bet.

FURTHER READING

PHYSICS OF SEMICONDUCTOR DEVICES. S. M. Sze. Wiley Interscience, 1981.

EVOLUTION OF THE MOS TRANSISTOR: FROM CONCEPTION TO VLSI. Chih-Tang Sah in *Proceedings of the IEEE,* Vol. 76, No. 10, pages 1280–1326; October 1988.

UHV/CVD GROWTH OF SILICON AND SILICON–GERMANIUM ALLOYS: CHEMISTRY, PHYSICS, AND DEVICE APPLICATIONS. Bernard S. Meyerson in *Proceedings of the IEEE,* Vol. 80, No. 10, pages 1592–1608; October 1992.

The Quantum Physics of Time Travel

Common sense may rule out such excursions—but the laws of physics do not

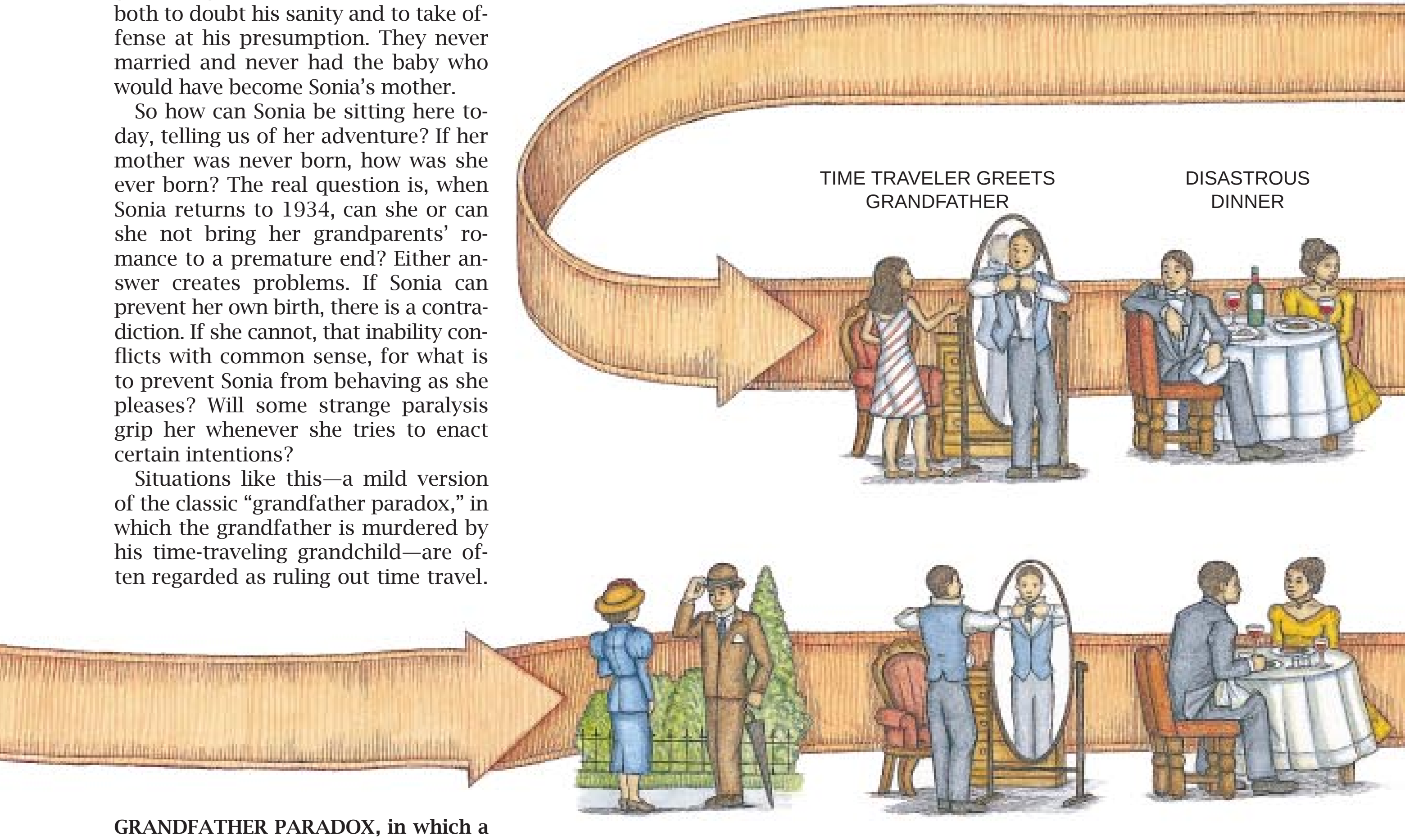
by David Deutsch and Michael Lockwood

Imagine, if you will, that our friend Sonia keeps a time machine in her garage. Last night she used it to visit her grandfather in 1934, when he was still courting her grandmother. Sonia convinced him of her identity by referring to family secrets that he had not yet revealed to anyone. This left him stunned, but worse was to follow. When he told his sweetheart over dinner that he had just met their future granddaughter, the lady's response was both to doubt his sanity and to take offense at his presumption. They never married and never had the baby who would have become Sonia's mother.

So how can Sonia be sitting here today, telling us of her adventure? If her mother was never born, how was she ever born? The real question is, when Sonia returns to 1934, can she or can she not bring her grandparents' romance to a premature end? Either answer creates problems. If Sonia can prevent her own birth, there is a contradiction. If she cannot, that inability conflicts with common sense, for what is to prevent Sonia from behaving as she pleases? Will some strange paralysis grip her whenever she tries to enact certain intentions?

Situations like this—a mild version of the classic "grandfather paradox," in which the grandfather is murdered by his time-traveling grandchild—are often regarded as ruling out time travel. Yet, surprisingly, the laws of physics do not forbid such adventures.

Another paradox, which often appears in science fiction, has been discussed by the Oxford philosopher Michael Dummett. An art critic from the future visits a 20th-century painter, who is regarded in the critic's own century as a great artist. Seeing the painter's current work, the critic finds it mediocre and concludes that the artist has yet to produce those inspired paintings that so impressed future generations. The critic shows the painter a book of reproductions of these later works. The painter contrives to hide this book, forcing the critic to leave without it, and then sets about meticulously copying the reproductions onto canvas. Thus, the reproductions exist because


TIME TRAVELER GREETS GRANDFATHER
DISASTROUS DINNER
GRANDFATHER MEETS CHARMING LADY
GRANDFATHER DRESSES FOR DINNER
DELIGHTFUL DINNER

GRANDFATHER PARADOX, in which a time traveler prevents his or her own birth, is a stock objection to time travel.

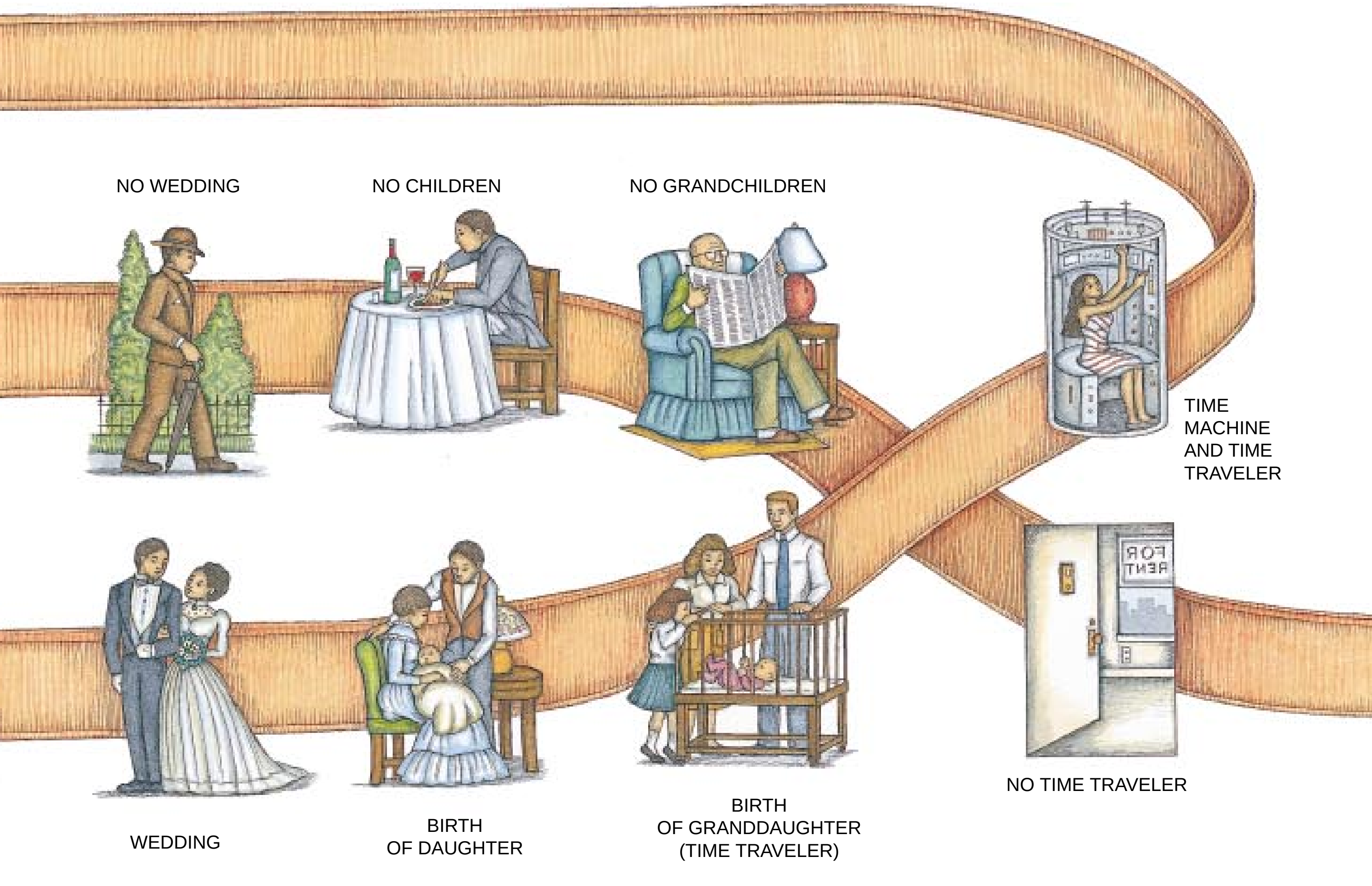
they are copied from the paintings, and the paintings exist because they are copied from the reproductions. Although this story threatens no contradiction, there is something very wrong with it. It purports to give us the paintings without anyone's having to expend artistic effort in creating them—a kind of artistic "free lunch."

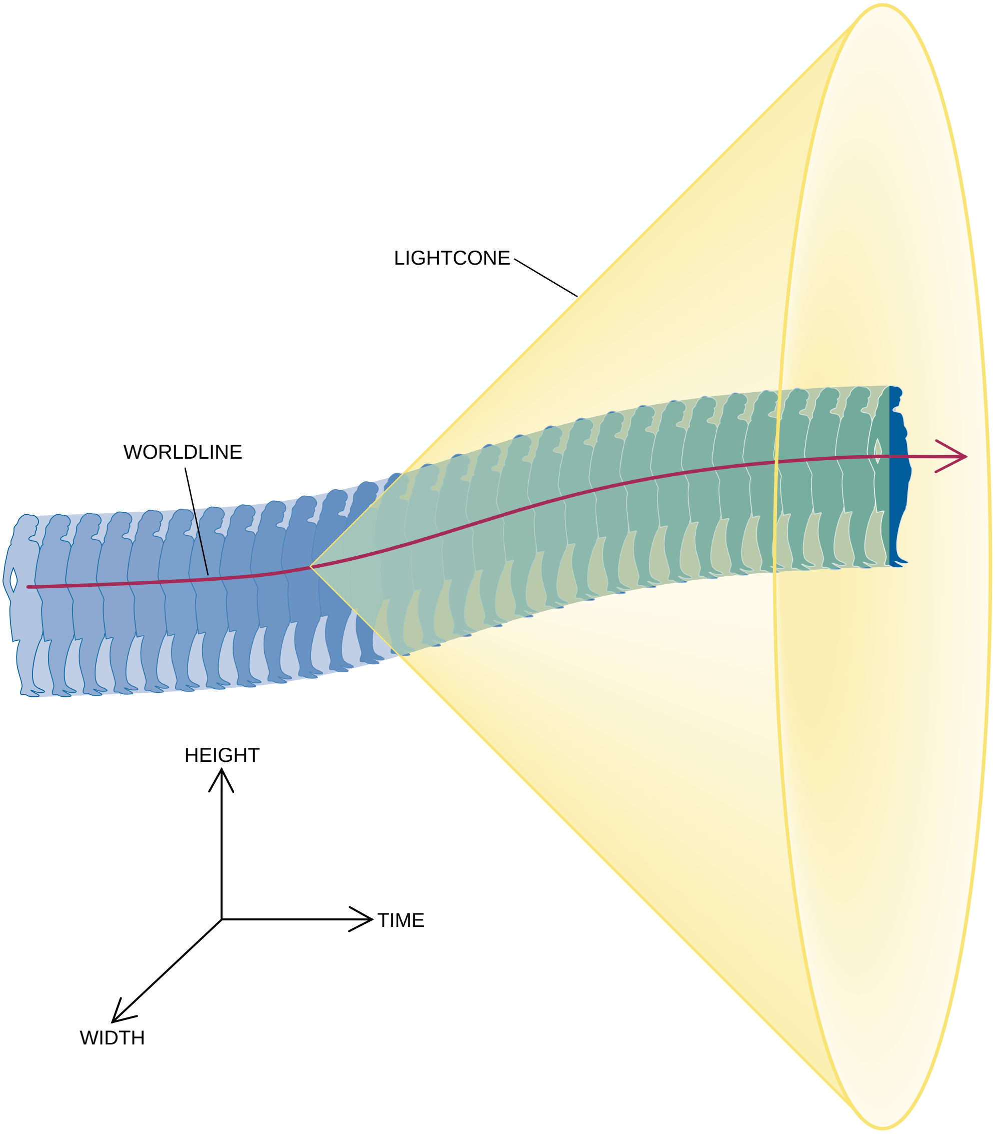
Persuaded by such objections, physicists have traditionally invoked a chronology principle that, by fiat, rules out travel into the past. One-way travel into the future raises no such problems. Einstein's special theory of relativity predicts that, with sufficient acceleration, astronauts could go on a journey and return to the earth decades into the future, while physically aging only a year or two. It is important to distinguish between predictions such as this, which are merely surprising, and processes that would violate physical laws or independently justifiable philosophical principles.

We shall shortly explain why traveling into the past would not violate any such principle. To do so, we must first explore the concept of time itself, as physicists understand it. In Einstein's special and general theories of relativity, three-dimensional space is combined with time to form four-dimensional space-time. Whereas space consists of spatial points, space-time consists of spatiotemporal points, or events, each of which represents a particular place at a particular time. Your life forms a kind of four-dimensional "worm" in space-time: the tip of the worm's tail corresponds to the event of your birth, and the front of its head to the event of your death. An object, seen at any one instant, is a three-dimensional cross section of this long, thin, intricately curved worm. The line along which the worm lies (ignoring its thickness) is called that object's worldline.

At any point on your worldline, the angle it makes with the time axis is a measure of your speed. The worldline of a ray of light is typically drawn as making an angle of 45 degrees; a flash of light spreading out in all directions forms a cone in space-time, called a lightcone [*see illustration on next page*]. An important difference between space and space-time is that a worldline—unlike, say, a line drawn on paper—cannot be any arbitrary squiggle. Because nothing can travel faster than light, the worldline of a physical object can never stray outside the lightcone emanating from any event in its past. Worldlines that meet this criterion are called

DAVID DEUTSCH and MICHAEL LOCKWOOD, both of the University of Oxford, share an interest in the philosophical foundations of physics. Deutsch is a research fellow at Wolfson College; he earned his doctorate in physics at Oxford under Dennis Sciama, did postdoctoral research under John A. Wheeler, Bryce DeWitt and Roger Penrose and is now working on the quantum theory of computation. Deutsch is writing a book on physics and philosophy entitled *The Fabric of Reality*. Lockwood is a fellow of Green College and lecturer at the department for continuing education; he earned his doctorate in philosophy, also at Oxford, under the late Sir Alfred Ayer. His book, *Mind, Brain, and the Quantum*, was published in 1989, and he is currently writing another on the nature of time. The authors believe that the real universe is far stranger than anything conceived of in science fiction but is also ultimately more intelligible.





SPACE AND TIME are combined into one four-dimensional entity, space-time. Here we show two space dimensions and time. A worldline connects all events in our life in space-time; since we have some size, a person's worldline is more like a worm extending from birth to death than a line. The worldlines of light rays emanating in all space directions from an event trace out a cone in space-time, called a lightcone. The worldline of any object, such as the navel of this figure, cannot stray outside a lightcone emanating from any point in its past.

timelike. Time, as measured by a watch, increases in one direction along a worldline.

Einstein's special theory of relativity requires worldlines of physical objects to be timelike; the field equations of his general theory of relativity predict that massive bodies such as stars and black holes distort space-time and bend worldlines. This is the origin of gravity: the earth's worldline spirals around the sun's, which spirals around that of the center of our galaxy.

Suppose space-time becomes so distorted that some worldlines form closed loops [*see illustration on opposite page*]. Such worldlines would be timelike all the way around. Locally they would conform to all the familiar properties of space and time, yet they would be corridors to the past. If we tried to follow such a closed timelike curve (or CTC) exactly, all the way around, we would bump into our former selves and get pushed aside. But by following part of a CTC, we could return to the past and participate in events there. We could shake hands with our younger selves or, if the loop were large enough, visit our ancestors.

To do this, we should either have to harness naturally occurring CTCs or create CTCs by distorting and tearing the fabric of space-time. So a time machine, rather than being a special kind of vehicle, would provide a route to the past, along which an ordinary vehicle, such as a spacecraft, could travel. But unlike a spatial route, a CTC (or rather, the surrounding closed timelike tube) gets used up if repeatedly traversed; just so many worldline worms can fit into it, and no more. If one travels on it to a particular event, one will meet everyone who has ever traveled, or will ever travel, to that event.

Does our universe now, or will it ever, contain CTCs? We do not know, but there are various theoretical conjectures about how they might be formed. The mathematician Kurt Gödel found a solution to Einstein's equations that describes CTCs. In that solution, the whole universe rotates (according to current evidence, the actual universe does not). CTCs also appear in solutions of Einstein's equations describing rotating black holes. But these solutions neglect infalling matter, and how far they apply to realistic black holes is a matter of controversy. Also, a time traveler would be trapped inside the black hole after reaching the past, unless its rotation rate exceeded a critical threshold. Astrophysicists think it unlikely that any naturally occurring black holes are spinning that fast. Perhaps a civilization far more advanced than ours could shoot matter into them, increasing their rotation rate until safe CTCs appeared, but many physicists doubt that this would be possible.

A kind of shortcut through space-time, called a wormhole, has been mooted by Princeton University physicist John A. Wheeler. Kip S. Thorne of the California Institute of Technology and others have shown how two ends of a wormhole could be moved so as to form a CTC. According to a recent calculation by J. Richard Gott of Princeton, a cosmic string (another theoretical construct that may or may not exist in nature) passing rapidly by another would generate CTCs.

We are at present a very long way from finding any of these CTCs. They may, however, become accessible to future civilizations, which might well attempt to enact time-travel paradoxes. Let us therefore take a closer look at the paradoxes to see what principles, if any, time travel would violate, according to classical and quantum physics.

Classical physics says, unequivocally, that on arriving in the past Sonia must do the things that history records her doing. Some philosophers find this an unacceptable restriction of her “free will.” But as an argument against time travel within classical physics, that objection is unpersuasive. For classical physics in the absence of CTCs is deterministic: what happens at any instant is wholly determined by what happens at any earlier (or later) instant. Accordingly, everything we ever do is an inevitable consequence of what happened before we were even conceived. This determinism alone is often held to be incompatible with free will. So time travel poses no more of a threat to free will than does classical physics itself.

The real core of the grandfather paradox is not the violation of free will but of a fundamental principle that underlies both science and everyday reasoning; we call this the autonomy principle. According to this principle, it is possible to create in our immediate environment any configuration of matter that the laws of physics permit locally, without reference to what the rest of the universe may be doing. When we strike a match, we do not have to worry that we might be thwarted because the configuration of the planets, say, might be inconsistent with the match being lit. Autonomy is a logical property that is highly desirable for the laws of physics to possess. For it underpins all experimental science: we typically take for granted that we can set up our apparatus in any configuration allowed by physical law and that the rest of the universe will take care of itself.

In the absence of CTCs, both classical and quantum physics conform to the autonomy principle. But in their presence, classical physics does not, because of what John L. Friedman of the University of Wisconsin and others call the consistency principle. This states that the only configurations of matter that can occur locally are those that are self-consistent globally. Under this principle, the world outside the laboratory can physically constrain our actions inside, even if everything we do is consistent, locally, with the laws of physics. Ordinarily we are unaware of this constraint, because the autonomy and consistency principles never come into conflict. But classically, in the presence of CTCs, they do.

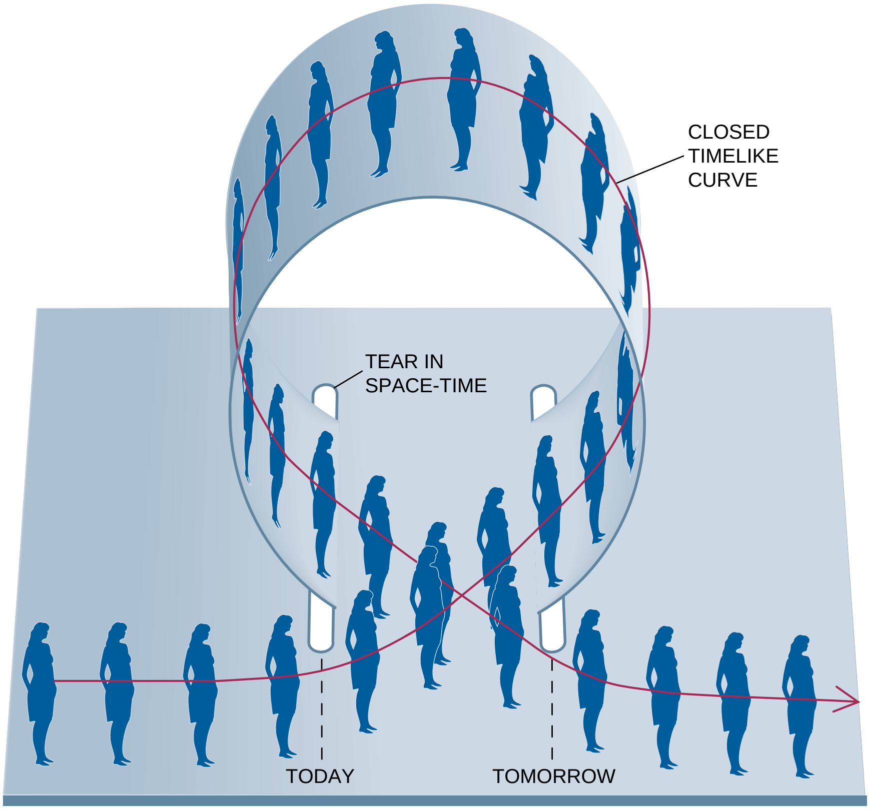
Classical physics says there is only one history, so try as she might to do other than what history dictates, consistency requires Sonia to act out her part in it. She may visit her grandfather. But perhaps when he tells Sonia’s grandmother-to-be what happened, she becomes worried about his state of health. He is very touched and proposes to her; she accepts. Not only could this happen—under classical physics something like it must happen. Sonia, far from altering the past, becomes part of it.

What if Sonia is determined to rebel against history? Suppose she travels back to meet her earlier self. At this meeting, her younger self records what her older self says and, in due course, having become that older self, deliberately tries to say something different. Must we suppose, absurdly, that she is gripped by an irresistible compulsion to utter the original words, contrary to her prior intentions to do otherwise? Sonia could even program a robot to speak for her: Would it somehow be forced to disobey its program?

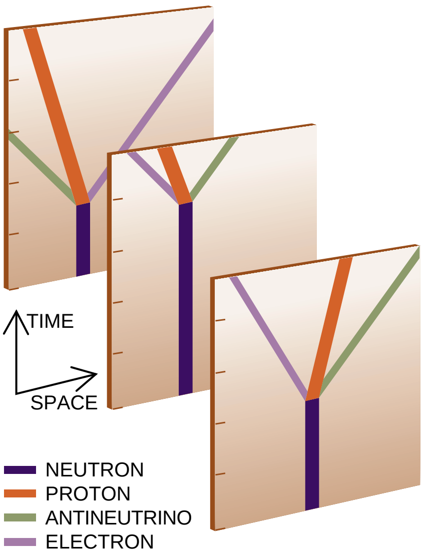
Within classical physics, the answer is yes. Something must prevent Sonia or the robot from deviating from what has already happened. It need not be anything dramatic, however. Any commonplace hitch will suffice. Sonia’s vehicle breaks down, or the robot’s program turns out to contain a bug. But one way or another, according to classical physics, consistency requires the autonomy principle to fail.

Now let us return to the story of the time-traveling art critic. We call this violation of common sense a knowledge paradox (the grandfather paradox is an inconsistency paradox). We use the term “knowledge” here in an extended sense, according to which a painting, a scientific article, a piece of machinery and a living organism all embody knowledge. Knowledge paradoxes violate the principle that knowledge can come into existence only as a result of problem-solving processes, such as biological evolution or human thought. Time travel appears to allow knowledge to flow from the future to the past and back, in a self-consistent loop, without anyone or anything ever having to grapple with the corresponding problems. What is philosophically objectionable here is not that knowledge-bearing artifacts are carried into the past—it is the “free lunch” element. The knowledge required to invent the artifacts must not be supplied by the artifacts themselves.

In an inconsistency paradox, physical events seem to be more tightly constrained than we are used to. In a knowledge paradox, they are less tightly constrained. For instance, the state of the universe before the art critic arrives does not determine who, if anyone, will arrive from the future or what he or she will bring along: the generally deterministic laws of classical physics allow the critic to bring back great pictures, poor pictures or no pictures at all. This indeterminacy is not what we



CLOSED TIMELIKE CURVE can be formed if space-time loops around. Entering such a curve tomorrow and moving forward in time, we can end up at today.



NEUTRON DECAY can occur at any time, though some times are more likely than others. For each instant in which the neutron might decay, there is a universe in which it decays at that instant, according to Everett's multiverse interpretation of quantum mechanics.

usually expect from classical physics, but it constitutes no fundamental impediment to time travel. Indeed, the indeterminacy would allow the classical laws to be supplemented with an additional principle, stating that knowledge can arise only as a result of problem-solving processes.

Yet that principle would land us in the same problem regarding autonomy as we encountered in the grandfather paradox. For what is to prevent Sonia from carrying new inventions into the past and showing them to their supposed originators? So although classical physics can, after all, accommodate the kind of time travel that is usually considered paradoxical, it does this at the cost of violating the autonomy principle. Hence, no classical analysis can wholly eliminate the paradox.

All this, however, is in our view academic. For classical physics is false. There are many situations in which it is an excellent approximation to the truth. But when closed timelike curves are involved, it does not even come close.

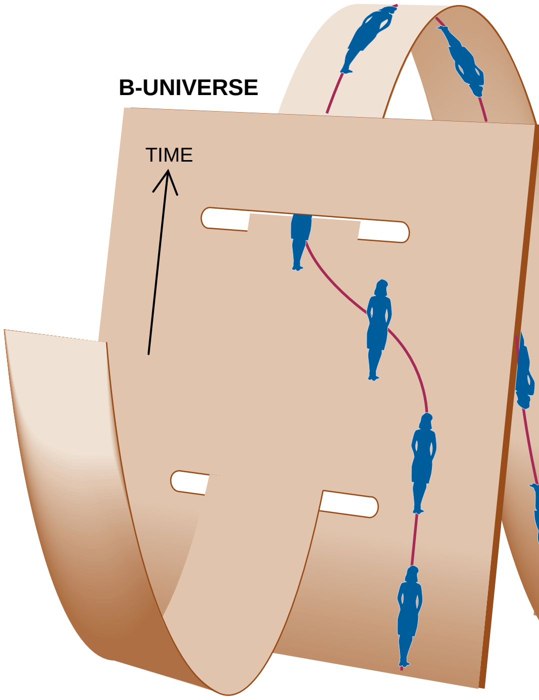
One thing we already know about CTCs is that if they exist, we need quantum mechanics to understand them. Indeed, Stephen W. Hawking of the University of Cambridge has argued that quantum-mechanical effects would either prevent CTCs from forming or would destroy any would-be time traveler approaching one. According to Hawking's calculations, which use an approximation that ignores the gravitational effects of quantum fields, fluctuations in such fields would approach infinity near the CTC. Approximations are inevitable until we discover how to apply quantum theory fully to gravity; but space-times containing CTCs push current techniques beyond the limits where they can be confidently applied. We believe that Hawking's calculations reveal only the shortcomings of those techniques. The quantum-mechanical effects that we shall be describing, far from preventing time travel, would actually facilitate it.

Quantum mechanics may necessitate the presence of closed timelike curves. CTCs, while hard to find on large scales, may well be plentiful at submicroscopic scales, where the effects of quantum mechanics predominate. There is as yet no fully satisfactory theory of quantum gravity. But according to many versions that have been proposed, space-time, though it appears smooth at large scales, has a foamlike submicroscopic structure containing many wormholes as well as CTCs reaching about 10^{-42} second into the past. For all we know, time travel by subatomic particles may be going on all around us.

More important, quantum mechanics can resolve the paradoxes of time travel. It is our most basic physical theory and constitutes a radical departure from the classical worldview. Rather than predicting with certainty what we shall observe, it predicts all possible outcomes of an observation and the probability of each. If we wait for a neutron to decay (into a proton, an electron and an antineutrino), we are most likely to observe this in about 20 minutes. But we might observe it immediately or be kept waiting indefinitely. How are we to understand this randomness? Is there something about the internal state of neutrons, currently unknown, that differs from one neutron to another and explains why each neutron breaks up when it does? This superficially attractive idea turns out to conflict with predictions of quantum mechanics that have been experimentally corroborated.

Other attempts have been made to preserve our classical intuitions by modifying quantum mechanics. None are generally deemed to have succeeded. So we prefer to take quantum mechanics at face value and to adopt a conception of reality that straightforwardly mirrors the structure of the theory itself. When we refer to quantum mechanics, we mean its so-called many-universes interpretation, first proposed by Hugh Everett III in 1957. According to Everett, if something physically can happen, it does—in some universe. Physical reality consists of a collection of universes, sometimes called a multiverse. Each universe in the multiverse contains its own copy of the neutron whose decay we wish to observe. For each instant at which the neutron might decay, there is a universe in which it decays at that instant. Since we observe it decaying at a specific instant, we too must exist in many copies, one for each universe. In one universe we see the neutron break up at 10:30, in another at 10:31 and so on. As applied to the multiverse, quantum theory is deterministic—it predicts the subjective probability of each outcome by prescribing the proportion of universes in which that outcome occurs.

Everett's interpretation of quantum mechanics is still controversial among physicists. Quantum mechanics is commonly used as a calculational tool that, given an input—information about a physical process—yields the probability of each possible output. Most of the time we do not need to interpret the mathematics describing that process. But there are two branches of physics—quantum cosmology and the quantum theory of computation—in which this is not good enough. These branches have as their entire subject matter the inner workings of the physical systems under study. Among researchers in these two fields, Everett's interpretation prevails.



MULTIVERSE PICTURE OF REALITY unravels the time travel paradoxes. Sonia plans to enter the time machine tomor-

What, then, does quantum mechanics, by Everett's interpretation, say about time travel paradoxes? Well, the grandfather paradox, for one, simply does not arise. Suppose that Sonia embarks on a "paradoxical" project that, if completed, would prevent her own conception. What happens? If the classical space-time contains CTCs, then, according to quantum mechanics, the universes in the multiverse must be linked up in an unusual way. Instead of having many disjoint, parallel universes, each containing CTCs, we have in effect a single, convoluted space-time consisting of many connected universes. The links force Sonia to travel to a universe that is identical, up to the instant of her arrival, with the one she left, but that is thereafter different because of her presence.

So does Sonia prevent her own birth or not? That depends on which universe one is referring to. In the universe she leaves, the one she was born in, her grandfather did marry her grandmother because, in that universe, he received no visit from Sonia. In the other universe, the one whose past Sonia travels to, her grandfather does not marry that particular woman, and Sonia is never born.

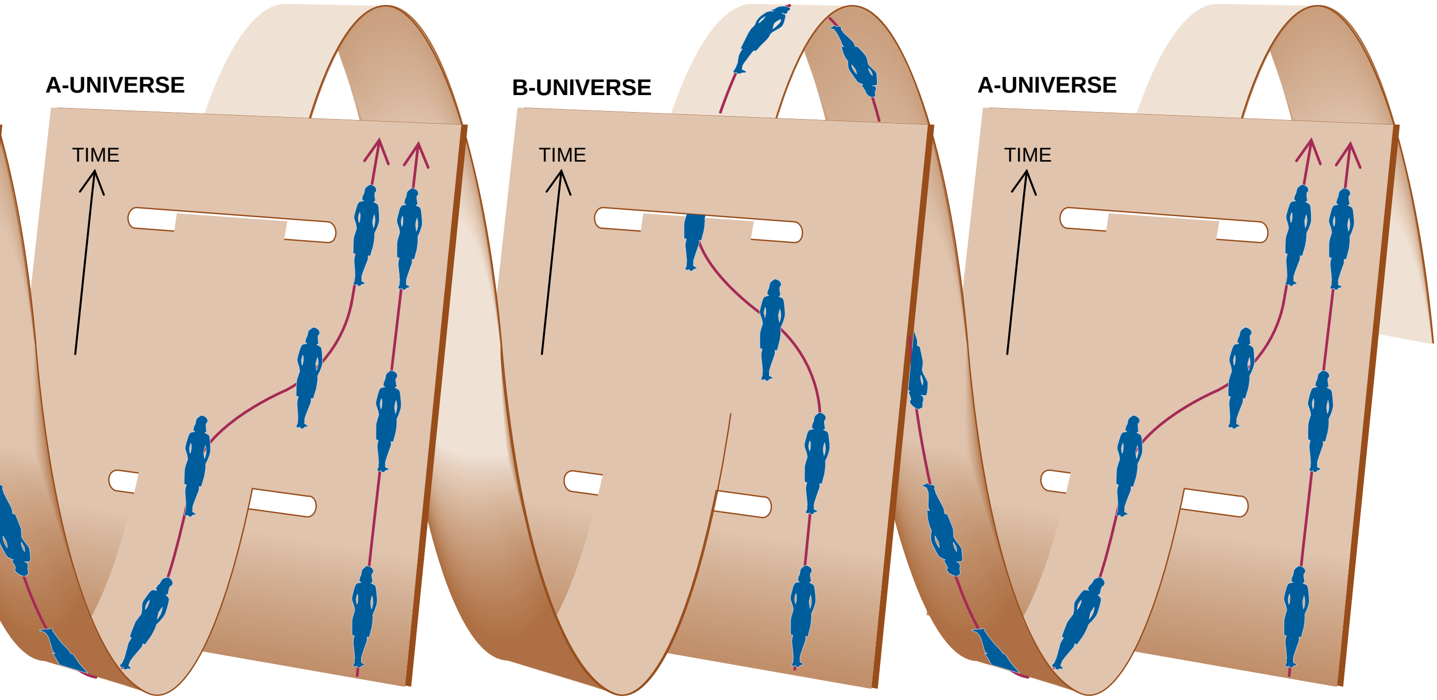
Thus, the fact that Sonia is traveling in time does not constrain her actions. And it turns out, according to quantum mechanics, that it never would. Quantum mechanics, even in the presence of CTCs, conforms to the autonomy principle.

Suppose Sonia tries her best to enact a paradox. She resolves that tomorrow she will enter the time machine and emerge today, unless a version of her first emerges today, having set out from tomorrow; and that if a version of her does emerge today, she will not enter the time machine tomorrow. Within classical physics, that resolution is self-contradictory. But not under quantum physics. In half the universes—call them A—an older Sonia steps out of the time machine. Consequently, just as she resolved, Sonia does not enter the time machine tomorrow, and each A-universe thereafter contains two Sonias of slightly different ages. In the other (B) universes, no one emerges from the time machine. So Sonia sets out and arrives in an A-universe where she meets a younger version of herself. Once again, she can behave as she likes in the past, doing things that depart from her (accurate) recollections.

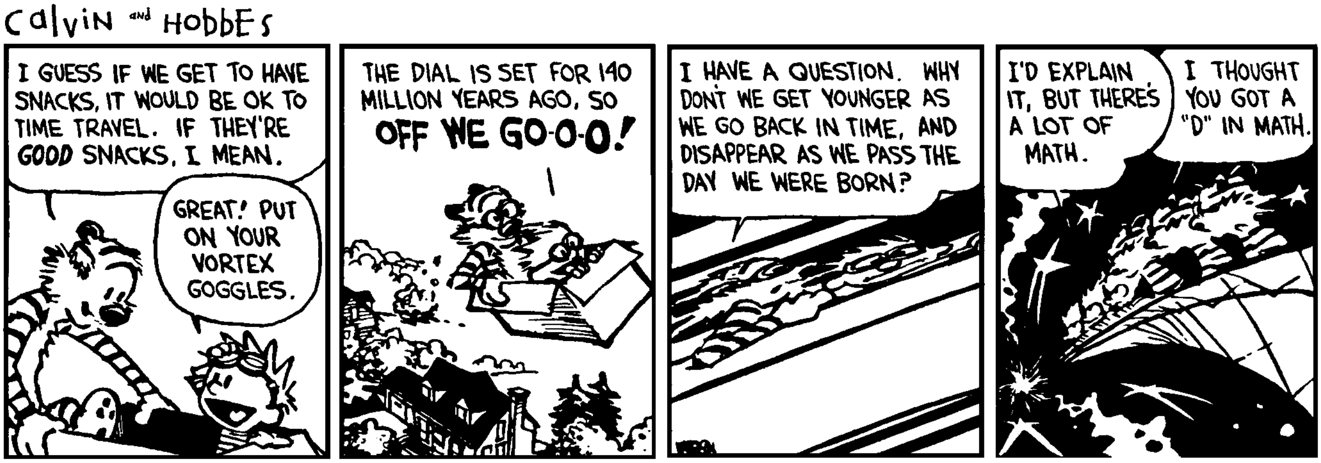
So in half the universes there is a meeting between two Sonias, and in half there is not. In the A-universes an older Sonia appears "from nowhere," and in the B-universes she disappears "into nowhere." Each A-universe then contains two Sonias, the older one having started life in a B-universe. Sonia has gone missing from each B-universe, having emigrated to an A-universe.

However convoluted Sonia's plans might be, quantum mechanics says the universes link up in such a way that she can carry them out consistently. Suppose Sonia tries to cause a paradox by traveling around the link twice. She wants to reappear in the universe she started from and join her previous self for a spaghetti dinner instead of the stir-fry she remembers having. She can behave as she likes, and in particular eat whatever she likes, in company with her younger self; however, the multiverse, by being linked up in a way different from that of the previous paradox, prevents her from doing so in her original universe. Sonia can succeed in sharing spaghetti with a version of herself only in another universe, while in the original universe she is still alone, eating stir-fry.

Time travel would make possible another curious phenomenon, which we call asymmetric separation. Suppose that Sonia's boyfriend, Stephen, stays behind while she uses her time machine in one of the ways we have described. In half the universes, she enters it and never returns. Thus, from Stephen's point of view, there is a possibility that he will be separated from her. Half the versions of him will see Sonia departing, never to return. (The other half will be joined by a second Sonia.) But from Sonia's point of view, there is no possibility of her being separated from Stephen, because every version of her



row and travel back to today but resolves that if she emerges from the time machine today, she will not enter tomorrow. She is able to carry out this plan, without paradox. In a B-universe she does not emerge today and so enters the time machine tomorrow. She then emerges today, but in an A-universe, and meets her copy—who does not enter the time machine.


Calvin and Hobbes
I GUESS IF WE GET TO HAVE SNACKS, IT WOULD BE OK TO TIME TRAVEL. IF THEY'RE GOOD SNACKS, I MEAN.
GREAT! PUT ON YOUR VORTEX GOGGLES.
THE DIAL IS SET FOR 140 MILLION YEARS AGO, SO OFF WE GO-O-O!
I HAVE A QUESTION. WHY DON'T WE GET YOUNGER AS WE GO BACK IN TIME, AND DISAPPEAR AS WE PASS THE DAY WE WERE BORN?
I'D EXPLAIN IT, BUT THERE'S A LOT OF MATH.
I THOUGHT YOU GOT A "D" IN MATH.

© 1991 Watterson (distributed by Universal Press Syndicate)

will end up in a universe containing a version of him—whom she will have to share with another version of herself.

If Stephen and Sonia follow a similar plan—entering the time machine if and only if the other does not first emerge—they can separate completely, ending up in different universes. If they carry out more complex intentions, each of them could end up in the company of any number of versions of the other. If time travel were achievable on a grand scale, competing galactic civilizations could use these asymmetric separation effects to have the whole galaxy to themselves. Also, an entire civilization could "clone" itself into any number of copies, just as Sonia did. The more often it did this, the likelier it would be that an observer would see it disappear from his universe, just as Stephen sees Sonia disappear from the A-universe when her "clone" appears in the B-universe. (Perhaps this explains why we have not yet encountered any extraterrestrials!)

In the art critic story, quantum mechanics allows events, from the participants' perspective, to occur much as Dummett describes. The universe that the critic comes from must have been one in which the artist did, eventually, learn to paint well. In that universe, the pictures were produced by creative effort, and reproductions were later taken to the past of another universe. There the paintings were indeed plagiarized—if one can be said to plagiarize the work of another version of oneself—and the painter did get "something for nothing." But there is no paradox, because now the existence of the pictures was caused by genuine creative effort, albeit in another universe.

The idea that time travel paradoxes could be resolved by "parallel universes" has been anticipated in science fiction and by some philosophers. What we have presented here is not so much a new resolution as a new way of arriving at it, by deducing it from existing physical theory. All the claims we have made about time travel are consequences of using standard quantum mechanics to calculate the behavior of logic circuits—just like those that are used in computers, except for the additional supposition that information can travel into the past along CTCs. The time travelers in this computer model are packets of information. Similar results have been obtained using other models.

These calculations definitively dispose of the inconsistency paradoxes, which turn out to be merely artifacts of an obsolete, classical worldview. We have argued that the knowledge paradoxes would likewise present no obstacle to time travel. But one cannot make that argument airtight until concepts like knowledge and creativity have been successfully translated into the language of physics. Only then can one tell if the "no-free-lunch" principle we require—that it takes problem-solving processes to create knowledge—is consistent, in the presence of CTCs, with quantum mechanics and the rest of physics.

There is a final argument that is often raised against time travel. As Hawking puts it, "The best evidence that time travel never will be possible is that we have not been invaded by hordes of tourists from the future." But this is a mistake. For a CTC reaches only as far back as the moment it was created. If the earth's first navigable CTC is constructed in 2054, subsequent time travelers could use it to travel to 2054 or later, but no earlier. Navigable CTCs might already exist elsewhere in the galaxy. But even then we should not expect "hordes of tourists from the future." Given the limited capacity of CTCs and that our stock of them at any given time cannot be replenished in this universe, a CTC is a nonrenewable resource. Extraterrestrial civilizations or our descendants will have their own priorities for its use, and there is no reason to assume that visiting the earth in the 20th century would be high on their list. Even if it were, they would arrive only in some universes, of which this, presumably, is not one.

We conclude that if time travel is impossible, then the reason has yet to be discovered. We may or may not one day locate or create navigable CTCs. But if anything like the many-universes picture is true—and in quantum cosmology and the quantum theory of computation no viable alternative is known—then all the standard objections to time travel depend on false models of physical reality. So it is incumbent on anyone who still wants to reject the idea of time travel to come up with some new scientific or philosophical argument.

FURTHER READING

CAUSAL LOOPS. Michael Dummett in *The Nature of Time.* Edited by R. Flood and M. Lockwood. Basil Blackwell, 1986.

DO THE LAWS OF PHYSICS PERMIT CLOSED TIME-LIKE CURVES? Kip S. Thorne in *Annals of the New York Academy of Sciences,* Vol. 631, pages 182–193; August 1991.

QUANTUM MECHANICS NEAR CLOSED TIMELIKE LINES. David Deutsch in *Physical Review D,* Vol. 44, No. 10, pages 3197–3217; November 15, 1991.

THE PARADOXES OF TIME TRAVEL. David Lewis in *American Philosophical Quarterly,* Vol. 13, No. 2, pages 145–152; April 1976. Reprinted in *The Philosophy of Time.* Edited by Robin Le Poidevin and Murray MacBeath. Oxford University Press, 1993.

MUST TIME MACHINE CONSTRUCTION VIOLATE THE WEAK ENERGY CONDITION? Amos Ori in *Physical Review Letters,* Vol. 71, No. 16, pages 2517–2520; October 18, 1993.

The Dynamics of Social Dilemmas

Individuals in groups must often choose between acting selfishly or cooperating for the common good. Social models explain how group cooperation arises—and why that behavior can suddenly change

by Natalie S. Glance and Bernardo A. Huberman

Imagine that you and a group of friends are dining at a fine restaurant with an unspoken agreement to divide the check evenly. What do you order? Do you choose the modest chicken entrée or the pricey lamb chops? The house wine or the Cabernet Sauvignon 1983? If you are extravagant, you could enjoy a superlative dinner at a bargain price. But if everyone in the party reasons as you do, the group will end up with a hefty bill to pay. And why should others settle for pasta primavera when someone is having grilled pheasant at their expense?

This lighthearted situation, which we call the Unscrupulous Diner's Dilemma, typifies a class of serious, difficult problems that pervade society. Sociologists, economists and political scientists find that this class of social dilemma is central to a wide range of issues, such as protecting the environment, conserving natural resources, eliciting donations to charity, slowing military arms races and containing the population explosion. All these issues involve goals that demand collective effort and cooperation. The challenge is to induce individuals to contribute to common causes when selfish actions would be more immediately and personally beneficial. Studies of these problems cast light on the nature of interactions among individuals and the emergence of social compacts. Moreover, they explain how personal choices give rise to social phenomena.

Social dilemmas have often been studied using groups of people who are given choices that present a conflict between the general good and the costs to an individual. Such experiments confirmed the hypothesis, first made by the economist Mancur L. Olson in the 1950s, that small groups are more likely to secure voluntary cooperation than are larger ones. They also revealed that repeated iterations of a situation tend to promote cooperative attitudes. The amount of cooperation further increases when communication among the participants is permitted.

More recently, powerful computers have been drafted for simulations of the social behavior of groups. The computer experiments gloss over the complexities of human nature, but we believe they can help elucidate some of the principles that govern interactions involving many participants. For the past three years, we have investigated social cooperation using both analytical techniques and computer simulations. We have tried to look not just at the outcomes of the dilemmas but also at the dynamics of the interactions and the ways in which those outcomes evolve in various groups.

Our mathematical theory of social dilemmas indicates that overall cooperation cannot generally be sustained in groups that exceed a critical size. That size depends on how long individuals expect to remain part of the group

NATALIE S. GLANCE and BERNARDO A. HUBERMAN explore their joint interest in the dynamics of social systems at the Xerox Palo Alto Research Center. For several years, Glance has studied the role of expectations and beliefs in systems of intentional agents. She received her Ph.D. in physics from Stanford University last June. Huberman is a Xerox Research Fellow and has been a visiting professor at the University of Paris and the University of Copenhagen. He received his physics degree from the University of Pennsylvania and has worked in condensed matter physics, statistical mechanics and chaotic dynamics. He is a co-recipient of the 1990 Prize of the Conference on Economics and Artificial Intelligence.

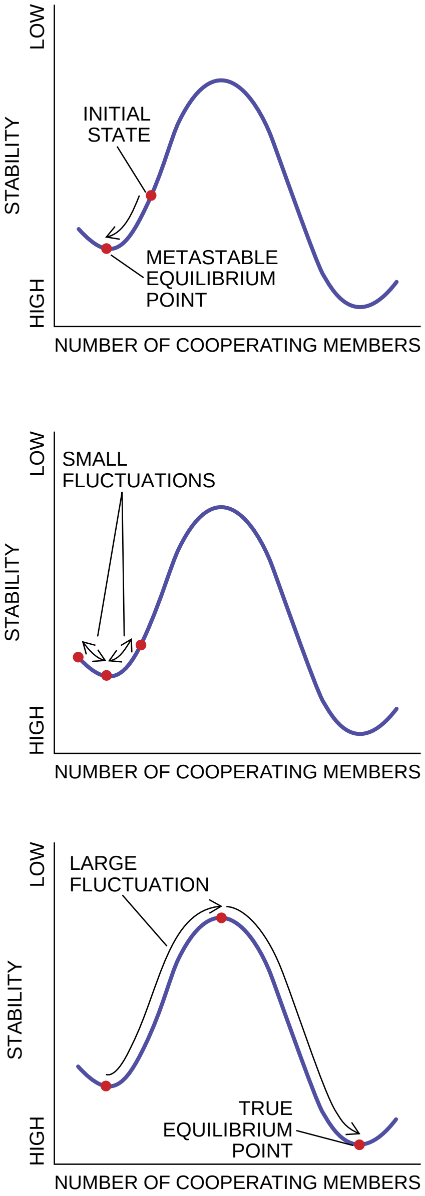
as well as on the amount of information available to them. Moreover, both general cooperation and defection can appear suddenly and unexpectedly. These results can serve as aids for interpreting historical trends and as guidelines for constructively reorganizing corporations, trade unions, governments and other group enterprises.

Mathematical theories of social dilemmas have traditionally been formulated within the framework of game theory. The mathematician John von Neumann and the economist Oskar Morgenstern developed that discipline in the mid-1940s to model the behavior of individuals in economic and adversarial situations. An individual's choices are ranked according to some payoff function, which assigns a numerical worth—in dollars or apples or some other commodity—to the consequences of each choice. Within game theory, individuals behave rationally: they choose the action that yields the highest payoff. (Real people may not be consistently rational, but they do behave that way when presented with simple choices and straightforward situations.)

Social dilemmas can readily be mapped into game settings. In general terms, a social dilemma involves a group of people attempting to provide themselves with a common good in the absence of central authority. In the Unscrupulous Diner scenario, for instance,

WHAT SHOULD I ORDER? That is the question for individuals in groups that have agreed to split the bill equally. An individual can get a modest meal and lower everyone's bill or get a sumptuous meal and eat at the others' expense—but thereby increase the chance that others, too, will follow that strategy. The Diner's Dilemma is typical of a class of social problems in which individuals must choose between cooperating with the group or defecting for personal gain.

the common good is achieved by minimizing the amount of the check. The individuals are said to cooperate if they choose a less expensive meal; they defect if they spare no expense (for the group, that is!). Of course, the game is only an idealized mathematical model—how well can one quantify intangibles such as the enjoyment of the meal or guilt over saddling friends with a large bill? Nevertheless, the dynamics of the game are still instructive.



STABILITY FUNCTION explains the dynamics of groups confronting social dilemmas. No matter what a group's initial state may be, it quickly shifts into a state of relative equilibrium, in which either many or few people are cooperating (*top*). Small fluctuations around this equilibrium point are routine (*middle*). Large fluctuations, however, which are rare, can carry the group over a stability barrier. The group will then very rapidly advance to a lower true equilibrium state (*bottom*). In the long run, a group will always settle into the lowest equilibrium state.

Each individual can choose either to contribute to the common good or to shirk and "free ride" on the sacrifices of others. All individuals share equally in the common good, regardless of their actions. Each person who cooperates therefore increases the common good by a fixed amount but receives back only some fraction of that added value. (The return is diminished by free riders who benefit without contributing.)

When an individual realizes that the costs of cooperating exceed her share of the added benefit, she will rationally choose to defect and become a free rider. Because every individual faces the same choice, all the members of a group will defect. Thus, the individually rational strategy of weighing costs against benefits has an inferior outcome: no common good is produced, and all the members of the group are less well off than they could be.

The situation changes, however, if the players know they will repeat the game with the same group. Each individual must consider the repercussions of a decision to cooperate or defect. The issue of expectations then comes to the fore. Individuals do not simply react to their perceptions of the world; they choose among alternatives based on their plans, goals and beliefs.

Of what do these expectations and beliefs consist? First, an individual has a sense of how long a particular social interaction will last, and that estimate affects her decision. A diner who goes out with a group once is more likely to splurge at the expense of others than is one who goes out with the same friends frequently. We call the expected duration of a game the horizon length. A short horizon reflects a player's belief that the game will end soon, whereas a long one means the player believes the game will repeat far into the future.

Second, each player has beliefs about how her actions will influence the rest of the group's future behavior. A diner may reject the option of an expensive meal out of fear that it would prompt others to order lavishly at the next gathering. The size of the group bears directly on this thinking. In a large crowd, a player can reasonably expect that the effect of her action, cooperative or not, will be diluted. (Ten dollars more or less on the group's bill matters less when it is divided among 30 diners rather than five.) The player will reason that her actions become less influential as the size of the group increases.

For groups beyond some size, overall cooperation becomes unsustainable. The likelihood of bad consequences from an individual's defection becomes so small, whereas the potential gain stays so large, that the disincentive to defect vanishes. As our experiments have determined, this critical size depends on the horizon length: the longer that players expect the game to continue, the more likely they are to cooperate. That conclusion reinforces the commonsense notion that cooperation is most likely in small groups with lengthy interactions.

The smallest possible social group, consisting of only two players, raises the special limiting case widely known as the Prisoner's Dilemma. It is so named because of one common way in which it is framed: a prisoner is given the choice of betraying a fellow prisoner (defecting) and going free or keeping silent (cooperating) and thereby risking a harsh punishment if the other prisoner betrays him. Because the psychology of the interactions is unique, certain strategies that work well for individuals in the Prisoner's Dilemma fail in larger groups. The highly successful one known as tit-for-tat depends on retaliation and forgiveness. A player initially cooperates and thereafter does whatever the other player last did. Tit-for-tat works because it allows each player to recognize that the other's actions are in direct response to her own. In groups of more than two, however, it is impossible for one player to punish or reward another specifically because any modification of her own actions affects the entire group.

In larger groups, an individual caught in a social dilemma forms a strategy for conditional cooperation from a calculation of the expected payoffs: she will cooperate if at least some critical fraction of the group is also cooperating. When enough of the others are cooperating, she expects that her future gains will compensate for present losses. If the number of cooperating individuals falls below that threshold, then her expected losses rule out cooperation, and she will defect. The strategies, expectations and thresholds of the individuals determine whether cooperation within a group is sustainable.

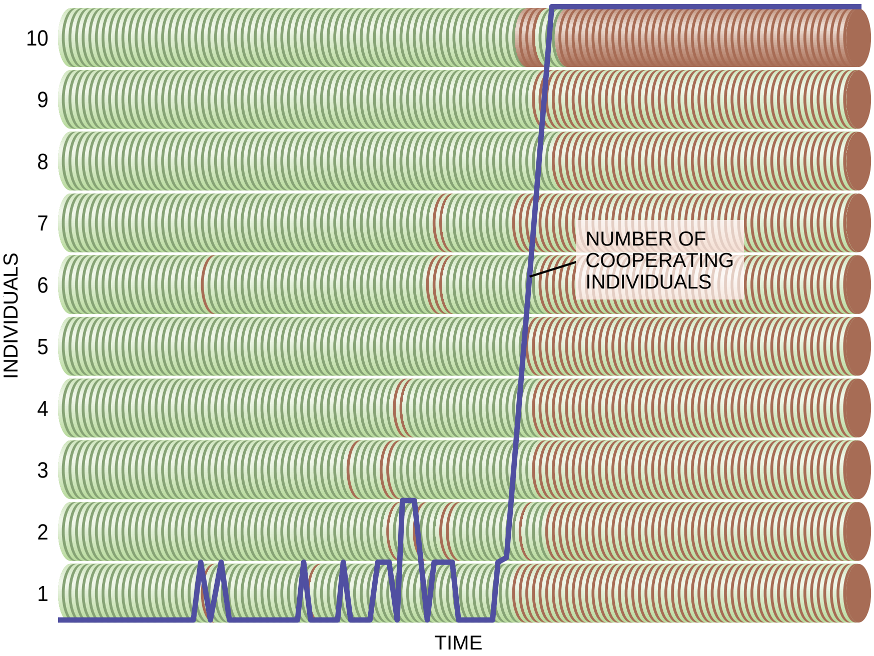
Quite aside from the question of whether a group can achieve cooperation is the equally important matter of how cooperation or defection emerges in a social setting. Imagine that the hypothetical diners, after many consecutive budget-busting meals, decide to split into smaller groups, hoping that the limited size of the resulting tables will aid cooperation. How long does it take for the small groups of defectors to switch? Is the process smoothly evolutionary or sudden?

To study the evolution of social cooperation, we borrowed methods from statistical thermodynamics. This branch of physics attempts to derive the macroscopic properties of matter from the interactions of its constituent molecules. We adapted the approach to study the aggregate behavior of individuals confronted with social choices.

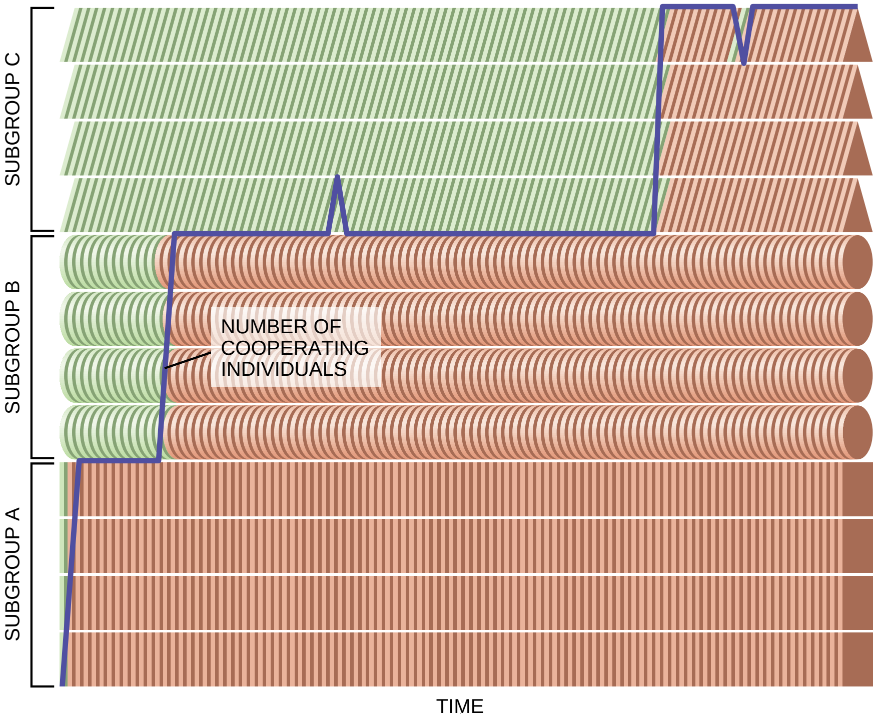
Our method relies on the mathematical construction of a curve called a stability function. This curve describes the relative stability of a group's behavior in terms of the amount of cooperation present. The values of the curve derive from a knowledge of the costs, benefits and individual expectations associated with a given social dilemma. The stability function generally has two minima, or troughs, which represent the most stable states of the group: widespread defection and widespread cooperation. They are separated by a high barrier, which is the least stable state. The relative heights of these features depend on the size of the group and the amount of information available to its members. From this function, one can predict the possible outcomes of the dilemma and how long the group will stay in a particular state.

Like a ball rolling downhill, the group's behavior will always gravitate from its initial state toward the closest trough. Once in a trough, however, the system does not become static. Instead it jiggles back and forth randomly, just as a small ball would be moved by vibrations. These random perturbations are caused by the uncertainty that individuals have about the behavior of others. If an individual misperceives the level of cooperation in the group, she may erroneously defect and thereby briefly move the system away from equilibrium. The more uncertainty there is in the system, the more likely there will be fluctuations around an equilibrium state.

These perturbations are usually small, so in the short run the system stays near one minimum. Over the long run, however, large fluctuations become important. Such fluctuations, caused by many individuals switching from defection to cooperation, or vice versa, can push the group over the barrier between the minima. Consequently, given sufficient time, a group will always end up in the more stable of the two equilibrium states, even if it initially moves into the other, metastable one.



OUTBREAKS OF COOPERATION can be simulated using computer agents that act like individuals. In a homogeneous group of agents that are all initially defectors (*green*), the shift to widespread cooperation (*orange*) is sudden and rapid.



HETEROGENEOUS GROUPS evolve stepwise toward overall cooperation, with each subgroup experiencing a distinct transition on its own.

Huge random fluctuations are extremely rare—on average, they occur over periods proportional to the exponential of the size of the group. Once the transition from the local minimum to the maximum of the function takes place, however, the system slides down to the global minimum very quickly—in a period proportional to the logarithm of the group size. Thus, the theory predicts that although the general behavior of a group in a dilemma stays the same for long periods, when it does change, it does so very fast.

Computer experiments demonstrate those predictions. A society of computational agents, or programs acting like individuals, can be presented with a social dilemma. The agents intermittently and asynchronously reevaluate their options and decide whether to cooperate or to defect. They base their decisions on information, which may be imperfect and delayed, about how many of the others are cooperating. The sum of all the agents' actions reveals the degree of cooperation or defection in the group. The experimenter can compile statistics on the level of cooperation over time.

One typical experiment features a group of 10 agents, all of which are initially defecting. If one agent misjudges how many others are cooperating and switches its behavior, that change might lead the rest of the group to make a similar shift. The group therefore stays at or near its initial metastable state of mutual defection for a long time, until a sudden and abrupt transition carries the group to mutual cooperation.

That abrupt appearance of cooperation in a computer simulation well describes certain real social phenomena, such as the recent upsurge in environmental awareness and activism. In many parts of the U.S. and Europe, voluntary recycling has become a normal part of daily life. A decade ago that was not the case. Recycling poses a social dilemma for the consumer: the environmental benefits are great if most of the population recycles but marginal if only a few do, and the individual's invested effort in bringing bottles and newspapers to the recycling center is the same in either case. Our theory may help explain why the population, after a long period of relative apathy, has so quickly embraced recycling, emissions controls and other environmental protection measures.

REGIONAL RECYCLING PROGRAMS are spreading in accordance with the rules of cooperation in hierarchies. The enjoyment of benefits from recycling in one community spurs neighboring communities to join the effort.

In the hypothetical social dilemmas we have described so far, all the individuals evaluate their payoffs the same way and share the same expectations about the outcomes of their actions. In any real group of humans, however, individuals have largely disparate beliefs. We have therefore looked at how diversity affects the dynamics of social dilemmas.

A heterogeneous group can display two different types of diversity: variation around a common average or segregation into factions. The first involves a simple spread in opinion or concern among individuals who are fundamentally the same. For example, some unscrupulous diners may anticipate and value more future meals than others. If the typical diner looks about 10 meals into the future, then individuals will have horizons that vary but cluster around that average.

Although models of social dilemmas that include this type of diversity are more complicated than ones for homogeneous groups, their dynamics still follow a clear pattern. Basically the diversity acts as an additional form of uncertainty, instigating fluctuations in the state of the group. If most individuals are defecting, the first to decide to cooperate will probably be the one who has the longest horizon. That decision might then convince others who have longer-than-average horizons to cooperate, too. Those transitions can trigger a cascade of further cooperation, until the whole group is cooperating.

The events that led to the mass protests in Leipzig and Berlin and to the subsequent downfall of the East German government in November 1989 vividly illustrate the impact of such diversity on the resolution of social dilemmas. Earlier that year Mikhail S. Gorbachev, then president of the Soviet Union, stopped backing the Eastern European governments with the force of the Soviet military. His new policy reopened the issue of whether the Eastern European population would still subscribe to the existing social compact. The citizens of Leipzig who desired a change of government faced a dilemma. They could stay home in safety or demonstrate against the government and risk arrest—knowing that as the number of demonstrators rose, the risk declined and the potential for overthrowing the regime increased.

A conservative person would demonstrate against the government only if thousands were already committed; a revolutionary might join at the slightest sign of unrest. That variation in threshold is one form of diversity. People also differed in their estimates of the duration of a demonstration as well as in the amount of risk they were willing to take. Bernhard Prosch and Martin Abraham, two sociologists from Erlangen University who studied the Leipzig demonstrations, claim that the diversity in thresholds was important in triggering the mass demonstrations. They also documented that over just six weeks the number of demonstrators grew from a handful of individuals to more than 500,000.

A second type of diversity within a social group describes differences that do not range around an average value. It is found in groups composed of several distinct factions, each characterized by a distinct set of beliefs. Among the diners, for example, might be a mix of students and professionals. Students on a tight budget have concerns different from those of well-off professionals. On the whole, the variation among the students' preferences would be small as compared with the average differences between the two subgroups.

When a large group containing several factions changes from overall defection to cooperation, it does so through progressive transitions. The subgroup with the greatest tendency to cooperate (for example, the one with the longest horizon in its average expectations or the one with the lowest average costs for cooperation) will usually be the first

PUBLIC DEMONSTRATIONS signaled the end of the old social compact in East Germany. People dissatisfied with the government saw that the risk of arrest declined as more people joined the protests, which fueled the explosive growth of the crowds.

to cross over. The other groups will then follow in turn, probably in the order of their willingness to cooperate.

Relationships among subgroups may powerfully influence the evolution of cooperation, a fact that is notably important in large hierarchical organizations. The weight that an individual in one division gives to the actions of others depends on those persons' placement in the hierarchy. Hierarchies are therefore very different from level groups.

Functional hierarchies often hide in informal settings. Air pollution is a problem that the whole world faces and must solve collectively. Yet each person is usually bothered more by a neighbor burning a compost pile than by someone across town doing the same. The dilution of environmental impact with distance can be represented as a hierarchy of layered interactions between neighborhoods, towns, counties, states, countries and continents. The effect of someone else's actions on your own choices will depend on how many layers distant she is from you.

The effective size of the hierarchy is therefore much smaller than the number of its constituents. Suppose that in its effect on your decisions, the action of your nearby neighbor counts as much as the summed actions of an entire distant neighborhood. Then the effective number of people influencing your decision is much smaller than the total population of your town. We can say that the hierarchy has been rescaled, because the whole is smaller than the sum of its parts.

Computer experiments show how cooperation can spread in large hierarchical organizations. Transitions from defection to cooperation (or the other way around) tend to originate within the smallest units, which usually occupy the lowest level of the hierarchy. Cooperation can then progressively spread to higher levels. The switching trend can even terminate if the cooperative influence of distant units is too attenuated to be felt. In such a case, the organization may contain some branches that cooperate and others that defect for long periods.

These results suggest practical ways to restructure organizations to secure cooperation among members faced with a social dilemma. Corporations benefit, for example, when managers share their knowledge with one another. Yet managers may withhold information they fear their colleagues can use for their own advancement. To volunteer information, a person needs to feel secure that others will, too. Setting up a network of smaller groups of managers could overcome the dilemma by promoting that sense of security. Moreover, restructuring a large corporation into smaller units may encourage the appearance of pockets of collaboration that might spread rapidly.

Conversely, when organizations grow without a major reorganization, the tendency to ride for free grows and lowers efficiency. The act of reorganizing does not guarantee instant improvement: the switch to collective cooperation may still take a long time. That time can be shortened by increasing the benefits for individuals who cooperate and by dispersing the most cooperative managers among small core groups throughout the organization.

The study of social dilemmas provides insight into a central issue of behavior: how global cooperation among individuals confronted with conflicting choices can be secured. These recent advances show that cooperative behavior can indeed arise spontaneously in social settings, provided that the groups are small and diverse in composition and that their constituents have long outlooks. Even more significantly, when cooperation does appear, it does so suddenly and unpredictably after a long period of stasis.

The world still echoes with the thunderous political and social events marking the past few years. The fall of the Berlin Wall, leading to a unified Germany, and the breakdown of the centralized Soviet Union into many autonomous republics are examples of abrupt global defections from prevailing social compacts. The member countries of the European Union currently face their own social dilemma as they try to secure supranational cooperation. The pressing issue is whether or not those countries can build a beneficial cooperative superstructure while each one remains autonomous. If our predictions are accurate, these restructurings will not proceed smoothly. Rather they will always be punctuated by unexpected outbreaks of cooperation.

FURTHER READING

THE LOGIC OF COLLECTIVE ACTION: PUBLIC GOODS AND THE THEORY OF GROUP. Mancur Olson, Jr. Harvard University Press, 1965.

THE TRAGEDY OF THE COMMONS. Garrett Hardin in *Science,* Vol. 162, pages 1243–1248; December 13, 1968.

COLLECTIVE ACTION. Russell Hardin. Johns Hopkins University Press, 1982.

INSTITUTIONAL STRUCTURE AND THE LOGIC OF ONGOING COLLECTIVE ACTION. Jonathan Bendor and Dilip Mookherjee in *American Political Science Review,* Vol. 81, No. 1, pages 129–154; March 1987.

THE OUTBREAK OF COOPERATION. N. S. Glance and B. A. Huberman in *Journal of Mathematical Sociology,* Vol. 17, Issue 4, pages 281–302; April 1993.

SOCIAL DILEMMAS AND FLUID ORGANIZATIONS. N. S. Glance and B. A. Huberman in *Computational Organization Theory.* Edited by K. M. Carley and M. J. Prietula. Lawrence Erlbaum Associates (in press).

Frogs and Toads in Deserts

Amphibians seem unlikely desert denizens. But those living in dry climes reveal a diverse and unusual array of adaptations to life at the extremes

by Lon L. McClanahan, Rodolfo Ruibal and Vaughan H. Shoemaker

With their moist skin and aquatic tendencies, frogs and toads seem best suited to life in or near bodies of water. Yet these creatures are found in arid regions throughout the world—from the Colorado Desert in California to African savannas. To survive in such climates, they have developed behavioral and physiological mechanisms that allow them to conserve water and remain cool.

The range of adaptations that we and other researchers have observed challenges some of the classical views of anuran, or frog and toad, physiology. Most of our previous knowledge was based on temperate species. The study of desert-dwelling amphibians has offered insights into the remarkable diversity of these animals.

Some 300 million years ago amphibians were the first vertebrates to invade the land, and they maintain a strong connection to fresh water. Modern amphibians include salamanders, caecilians—legless amphibians that resemble worms—and anurans. Contrary to most people's perception, no biological distinction separates frogs and toads: terrestrial anurans with warty skins are usually called toads; aquatic forms with smooth skins are called frogs. Most amphibians lay their eggs in water and have larvae that lead a fishlike existence until they undergo metamorphosis. Once equipped with legs and lungs, they spend at least some time on land.

Despite the ability to survive on terra firma, most amphibians inhabit sites near fresh water or areas of elevated humidity and rainfall. Relatively few live in arid regions. This geographic distribution reflects the physiology of amphibians: they are generally ill suited to face the rigors of the desert.

Other terrestrial vertebrates, including reptiles, birds and mammals, have an integument, or skin, that can protect them against desiccation. The outer part, known as the stratum corneum, is composed of multiple layers of flattened, dead epidermal cells. This skin deters water loss. Unlike these other vertebrates, amphibians typically have a stratum corneum consisting only of a single-cell layer. This very permeable skin offers some benefits. Amphibians do not drink but absorb water across the skin from moist surfaces, such as wet rocks or leaves, and soil as well as from pools. Oxygen and carbon dioxide pass readily through the skin; for example, lungless salamanders rely on the integument for gas exchange.

At the same time that it is ideally suited to absorb water, the thin amphibian skin is a perfect conduit for evaporation. Under moderate conditions of temperature and humidity, most amphibians cannot survive for more than a day in circulating air because they quickly dehydrate. Even anurans, which can tolerate much larger water losses than can other vertebrates, are jeopardized in this situation.

Amphibians also differ from other vertebrates in the way they excrete wastes. The kidneys of desert animals are challenged to conserve water while eliminating the nitrogen waste produced during the metabolism of protein and other nitrogen-containing compounds. Birds and reptiles have solved this problem by synthesizing uric acid, a poorly soluble, nitrogen-rich compound. This waste can be excreted as a solid precipitate, and little water is lost. Mammals incorporate their

LON L. McCLANAHAN, RODOLFO RUIBAL and VAUGHAN H. SHOEMAKER have spent more than 20 years studying desert frogs and toads. Their association began in 1965, when McClanahan was a graduate student working with Ruibal at the University of California, Riverside, and Shoemaker arrived there as an assistant professor. The three of them formed a lasting friendship that led to research ventures in North and South American deserts. In particular, they discovered bizarre amphibian adaptations in the unique habitat of the Gran Chaco in Argentina and Paraguay. Ruibal and Shoemaker are professors of biology at Riverside; McClanahan is professor of biology at the California State University at Fullerton and director of the university's Ocean Studies Institute.

TREE FROG *Phyllomedusa sauvagei* is one of several species of frog adapted to arid regions. This South American anuran can survive in hot, dry environments because it tolerates high body temperatures. It also conserves water by losing little liquid through its skin or from its excretory system.

waste nitrogen into urea, a water-soluble molecule. They conserve precious water by creating urine that has a high concentration of dissolved materials, primarily urea and salts. This ability is particularly well developed in desert rodents. The kangaroo rat, for instance, produces urine that is 14 times as concentrated as blood; Australian desert mice achieve ratios of 20 or more.

Like mammals, adult anurans usually produce urea, but in contrast to mammals, they cannot make urine that is more concentrated than blood. They therefore require a great deal of water to eliminate their waste. Most amphibians are also less tolerant of high body temperatures than are other terrestrial vertebrates. Birds, mammals and desert lizards can maintain body temperatures of about 40 degrees Celsius. But many frogs and toads die at temperatures of 35 degrees C or less.

Terrestrial frogs and toads share some characteristics that help to offset the handicaps imposed by their permeable skin and inefficient kidneys. When water is unavailable, these animals stop producing urine and allow wastes to accumulate in the body fluids. Dehydration also results in beneficial changes in the skin and urinary bladder. Increased permeability of the bladder permits the animals to restore water lost by evaporation by reclaiming water stored as dilute urine. In addition, dehydrated frogs and toads absorb water through the skin much more readily than when they are hydrated. These responses are mediated by a posterior pituitary hormone called arginine vasotocin.

However useful in the short term, these protective measures do not permit anurans to survive without water for a long time in hot, dry environments. Thus, frogs and toads in these regions have evolved additional strategies. Perhaps the most straightforward adaptation to life in dry lands is to stay near what little water is available, and many species of amphibians do just that. Springs, seeps, canyons that drain higher ground as well as man-made impoundments are all permanent water sources where frogs can be found.

The frog *Hyla cadaverina,* or the California tree frog, found in the Colorado Desert, is one such oasis dweller. These frogs live near seeps and water holes. Although air temperatures often soar above 40 degrees C, the frogs' bodies do not exceed 30 degrees C, because they undergo evaporative cooling. *H. cadaverina* can store up to 25 percent of its body weight in the form of dilute urine. As it loses water through evaporation, water from urine is recycled back into the body. When these reserves are exhausted, the frogs return to the seep and take in more water.

The seeps also provide a place for mating and egg development. On summer days the frogs remain in clumped masses near water holes. The reasons for clustering in this fashion are not clear. We think huddling may decrease the amount of water lost by an individual. As evening approaches, the frogs slowly disperse and start to feed at the water's edge in preparation for night, when they chorus, mate and lay eggs in the pond. If rain forms other pools, the frogs will spread out and lay eggs in these waters as well.

The red-spotted toad, *Bufo punctatus,* uses the canyon waters of the Colorado Desert for spawning. But unlike *H. cadaverina,* these toads do not permanently reside in a moist environment. Using transmitters, we tracked toads that moved up to 100 meters

away from the water. Even in the middle of a summer day, we discovered toads burrowed in the coarse soil of rock crevices far from water. These crevices provide a shelter that is shared with other members of the same species: at one site we found eight individuals in the same cranny.

Because *B. punctatus* is unable to tolerate temperatures higher than 35 degrees C, we presumed that the toads seek microhabitats that remain relatively cool during the day. Studies using implanted location and temperature sensors confirmed this theory. We tracked one toad from September, late in the active season, until December, one of the colder months. At night dur-

Anuran Adaptations in Arid Regions

COLORADO DESERT

Hyla cadaverina
(top left)
This frog lives near permanent water supplies in the desert. It can store up to 25 percent of its body weight as dilute urine. If it loses water through its skin, the frog can reabsorb water from the bladder.

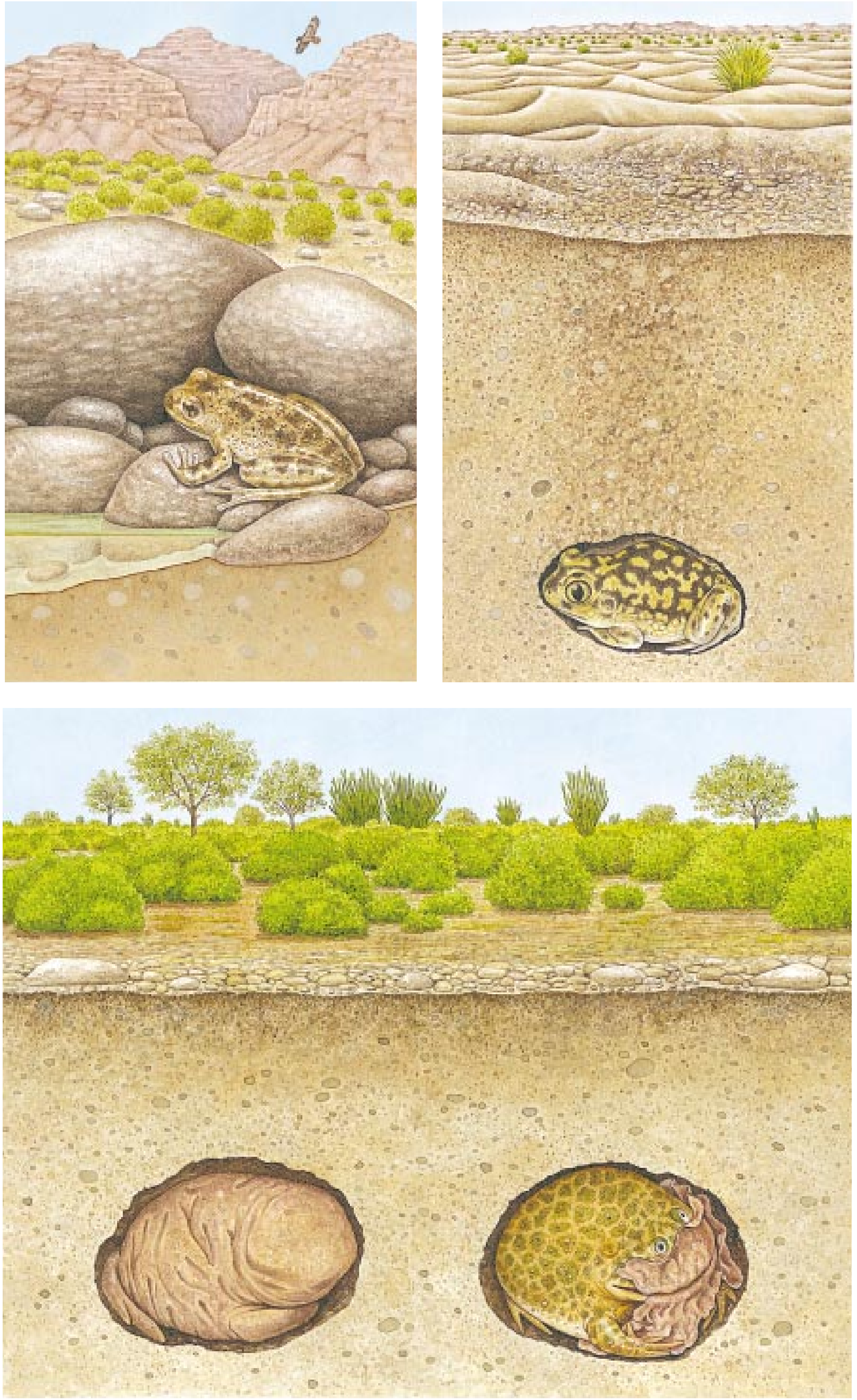
Scaphiopus couchi
(top center)
This species of spadefoot toad inhabits areas of the desert where there are no year-round water sources. To survive the dry season, the toad buries itself as deep as one meter underground. Some spadefoots have survived two-year droughts in this manner.

Bufo punctatus
(top right)
This toad can travel 100 meters from water sources, looking for food. To stay cool, it finds rock crevices and, like *H. cadaverina*, recycles water from urine. The toad can tolerate losing 40 percent of its body water (in contrast, camels can survive a 20 percent depletion).

GRAN CHACO REGION

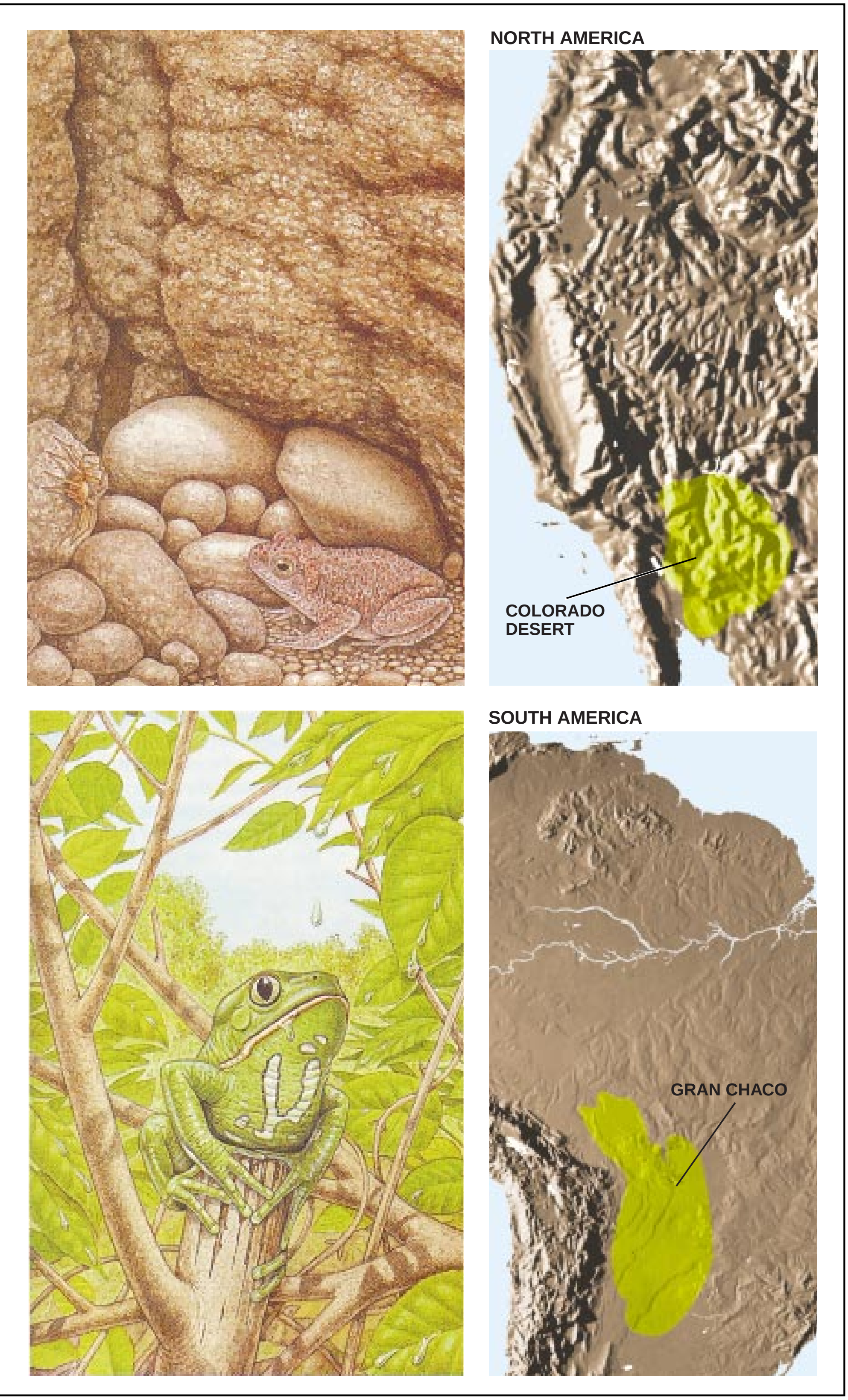
Lepidobatrachus laevis
(bottom left)
This toad survives dry periods by burrowing into mud and becoming entombed. It then constructs a multilayered cocoon (*left side of panel*) to resist water loss. At the beginning of the rainy season, the toad pulls the cocoon over its head and eats it (*right side of panel*). Afterward *L. laevis* emerges.

Phyllomedusa sauvagei
(bottom right)
This tree frog coats itself with a waxy substance to prevent water loss. It also excretes uric acid to conserve water. *P. sauvagei* is the only anuran known to drink water—most absorb it through their skin. It does so by letting drops roll into its nearly closed mouth.



ing the warm season, it traveled 85 meters from its burrow to a small stream; early in the morning the toad returned to the same burrow.

The strategy is clearly successful. During the toad's active months, daytime air temperatures routinely exceeded the critical limits for the toad, but its body temperature never rose above 31 degrees C. Reduced temperatures in the burrow as well as evaporation controlled the toad's body temperature.



Habitat selection also proved vital when the toad became inactive in winter, protecting it from cold nightly temperatures. In December, when air temperatures ranged from 12 degrees C at sunset to four degrees C at sunrise, the toad's body temperature inside the burrow remained constant at 25 degrees C.

Not only does *B. punctatus* choose protective habitats, it possesses a characteristic that allows it to abandon its water source and forage for insects during hot summer nights. The toads can store 40 percent of their body weight as dilute urine. If all the bladder reserves are used, the toads have another safety feature: they can tolerate the loss of up to 40 percent of their body water. Humans can survive only a 10 percent loss of their body water; camels, 20 percent.

Most areas of the desert lack permanent water sources. Amphibians that live in such regions do not retreat to rocky crevices; they burrow underground. The soil protects anurans from the extreme surface heat during the dry season. They obtain water through the soil, and as the earth dries, they conserve water because little is lost to evaporation.

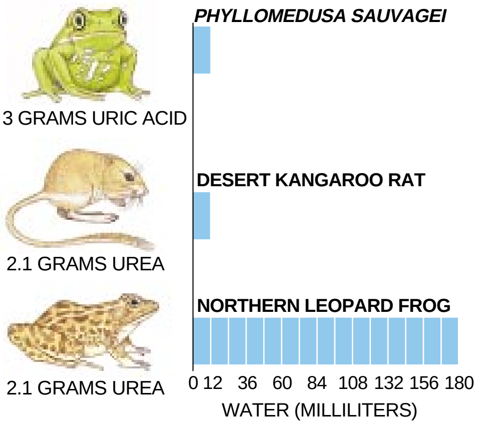
We first studied the ecology of burrowing toads in southeastern Arizona. Three species of *Scaphiopus,* or spadefoot toads, are abundant in this area and can easily be found after the summer rains bring them out of the burrow. Although some toads emerge during light rains, the majority come out during the evening of the first heavy downpour, when temporary ponds form. We wondered what cues might trigger this exodus. We determined that gently and silently moistening the soil by pouring water on it did not elicit a response, whereas sprinkling the soil to imitate rain caused the animals to surface. The toads came out even when the ground was kept dry with plastic. Sound alone is a sufficient cue.

After they have left the burrow, adults make their way to the ponds to breed. Toads captured on the way often have a stomach full of termites that have simultaneously emerged from their underground nests. Spadefoots have a prodigious appetite and can consume 55 percent of their body weight in a single night. One meal of lipid-rich termites can provide enough energy to maintain a toad for more than a year; it may even be sufficient to allow a female to produce eggs.

Once in the water, the toads usually stay about 24 hours, just long enough to mate and lay eggs. They then leave and sometimes travel miles to take up

Comparison of Nitrogen Excretion*

Most amphibians, such as the northern leopard frog, excrete nitrogen as urea in solution, using ample water. Desert-dwelling mammals use less water to excrete the same amount of urea. Desert frogs, such as *Phyllomedusa sauvagei*, resemble mammals in that they use little water; they differ in that they excrete nitrogen waste as precipitated uric acid.



*per gram of nitrogen

residence in shallow burrows. On windless nights they forage on the still damp desert floor. The toads return to the burrow after feeding, reburying themselves in the friable soil. By late summer they disappear and are not seen until they unearth themselves the next summer. In parts of the Colorado Desert, one of the spadefoot toads, *Scaphiopus couchi,* was found to survive for nearly two years without rain. Fat reserves are sufficient because the toad drastically lowers its metabolic rate during its retreat underground.

We considered whether spadefoot toads could regulate the composition of their body fluids while below ground. So before the rain came, we tried to locate toads by digging near the dried-out ponds where they breed; we even used a bulldozer to excavate a large area. We did not find any toads. Fortunately, one of the local ranchers informed us that he occasionally uncovered toads when digging holes for fence posts, some at the depth of nearly one meter. Moreover, the burrowed creatures were surprisingly far, at 100 meters, from the vanished pond.

Armed with this information, we were able to find some sites where buried toads could be excavated and studied. At a location in Arizona, we dug up toads at various times during the year. We took samples of plasma and urine and analyzed them for electrolytes, urea and total solute concentration. In addition, we measured the volume of urine in the bladder and the moisture content of the soil at the burrow site.

We discovered that from the time the toads burrowed in September until we unearthed them in March the solute concentration in plasma and urine was typical of fully hydrated anurans. Furthermore, in March most individuals had retained a lot of dilute urine in their bladders—an amount equivalent to between 25 and 50 percent of their body weight. The soil adjacent to the toads was also fairly moist, and the forces binding water to soil were sufficiently weak, so that water could move osmotically into the animal.

In late June, just before the first rains, the toads were still in excellent condition but contained high concentrations of urea in both plasma and urine. By then the soil had dried so much that water could not pass from soil to animal—unless there was enough buildup in the total solute concentration in the toads to overcome the increased forces binding water to the soil.

The accumulation of urea in the spadefoots' body fluids tilts this exquisite osmotic balance. In laboratory studies we have found that the toads can produce and store urea on demand. If they are placed in relatively dry soil, they make more urea than when they are in wetter soils. This same strategy of storing urea is used by frogs that can adapt to brackish water (there are no truly marine frogs). In this case, the accumulation of urea makes the body fluids more concentrated than those fluids surrounding the animal, causing water to enter the creature by osmosis. For the same reason, some marine fishes such as sharks and coelacanths also retain elevated concentrations of urea in the body fluids.

A bizarre burrowing toad from the family Ceratophryidae has an even more elaborate survival tactic than the spadefoot toad. Unlike spadefoots, these animals do not store high concentrations of urea in their body fluids; they depend on a cocoon to prevent water loss. These toads inhabit the Gran Chaco, a semiarid region that extends from north-central Argentina into Paraguay, where they live in temporary ponds that fill up during the summer rains. One of the creatures, *Lepidobatrachus laevis,* is voracious and pugnacious. It screams loudly and bites when threatened. The animal is known as kururú-chiní, or "the toad that shrieks," in Guaraní, the language of Paraguay.

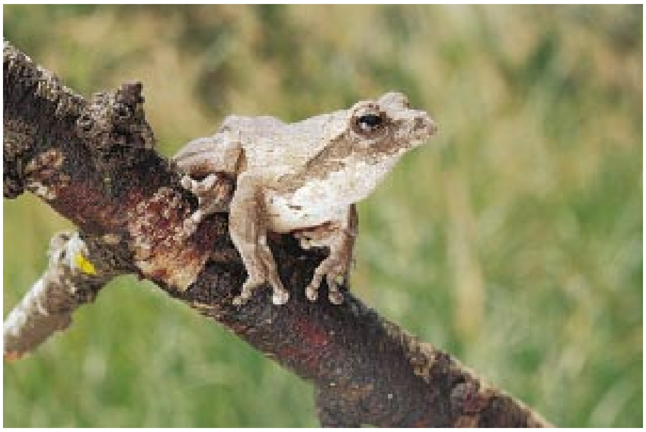
Unlike spadefoot toads, the kururú-chiní remain in ponds during times of drought. As water evaporates, they burrow to a shallow depth and become entombed as the mud dries. The toads then start to produce a multilayered cocoon. The kururú-chiní is not alone in this practice. Various other burrowing species in Australia, Africa and Mexico are known to form cocoons—indeed, it seems cocoon formation evolved independently several times.

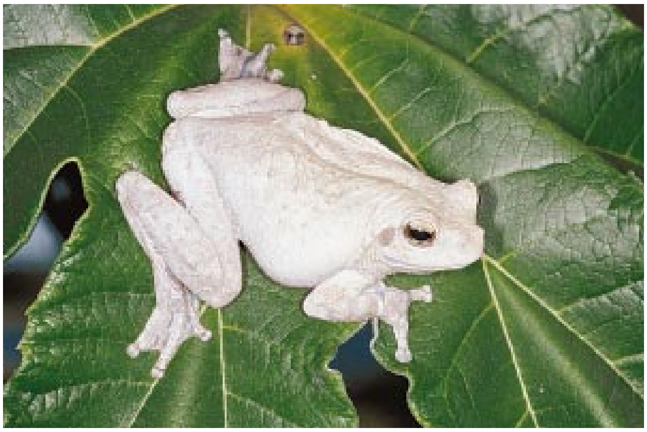
All anurans periodically shed the outermost cell layer of their skin after a replacement layer has been formed. During cocoon formation, however, the outermost layer of skin is not shed and stays in place as additional layers grow beneath. Indeed, the kururú-chiní forms a new layer every 24 hours until it is enclosed by a multilayered cocoon of flat cells with dried mucus between each layer. When plotted against time, evaporative water loss during cocoon formation shows a hyperbolic decline as each layer forms.

In the laboratory the kururú-chiní will construct a cocoon if it is deprived of water and placed in a quiet, dark place. Time-lapse filming shows that the toad moves only slightly as the cocoon thickens; after a few weeks, it remains motionless for days on end. But even when ensconced in its fully formed cocoon, the kururú-chiní retains its unnerving ability to shriek when disturbed. If the toad is gently moistened with water, it will awaken and start to shed the cocoon in one piece. The animal uses its legs to roll the cocoon up from the posterior part of the body, over its head. The kururú-chiní then promptly eats the entire wet casing.

Perhaps the most striking anuran adaptation to life in the desert was discovered serendipitously. In 1970 we received a surprising reprint from John P. Loveridge of the University of Zimbabwe (then Rhodesia). It described experiments showing that the gray tree frog, *Chiromantis xerampelina,* could survive for long periods in open, dry containers. The frog lost weight at a fraction of the rate of other frogs—rates similar to those of a lizard kept under identical conditions. Loveridge also noted that most of the dry mass of the frog's urine was uric acid.

His findings were heretical. All frogs and toads were thought to have water-permeable skins—with the exception of cocoon-forming burrowers—and to excrete nitrogen as urea. Loveridge's description suggested a frog with a





COLOR CHANGE permits *Chiromantis xerampelina* to endure direct sunlight. By abandoning its dark, protective coloring (*left*) and adopting white (*right*), the tree frog reflects the sun's rays. The frog also survives the heat by storing large volumes of water in its bladder and then using the reserve for evaporative cooling.

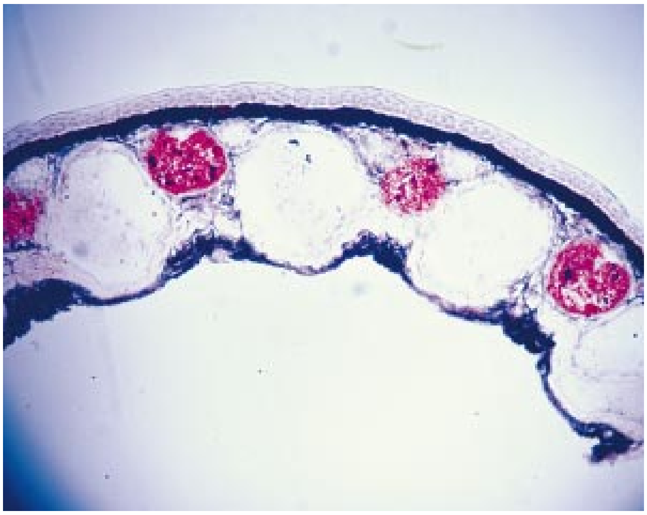
reptile's impermeable skin and a reptile's capacity to excrete uric acid.

At the time the paper appeared, we were beginning studies of amphibians in the Gran Chaco. We were impressed with the diversity of the amphibian fauna living in this region, which included a green arboreal frog, *Phyllomedusa sauvagei,* known locally as *rana verde.* Whereas most frogs living in arid regions must remain underground or near water, except during the rainy season, *Phyllomedusa* and *Chiromantis* can remain perched in trees, where they feed. *P. sauvagei* is active before the onset of summer rains in Paraguay, and *Chiromantis* can be found during the dry season in Zimbabwe. Because of Loveridge's work, we made crude measurements of water loss in *P. sauvagei* in the laboratory, but we observed nothing unusual.

A few weeks later one frog voided a large blob of semisolid urine while being handled. A quick trip to the spectrophotometer revealed that the major component was uric acid. Studies of nitrogen balance in *Phyllomedusa* and *Chiromantis* have amply confirmed the benefits of uric acid excretion. In both species, about 80 percent of the waste nitrogen is emitted as uric acid or urate salts. In addition, sodium and potassium precipitate along with uric acid, which further increases the excretory capacity of the kidneys. Thus, these frogs can feed while deprived of water for long periods. Across species, there appears to be a wide range in the ability to synthesize uric acid: from 230 milligrams per kilogram per day in *P. sauvagei* to only 40 milligrams per kilogram per day in *P. bicolor,* which inhabits tropical regions of Brazil.

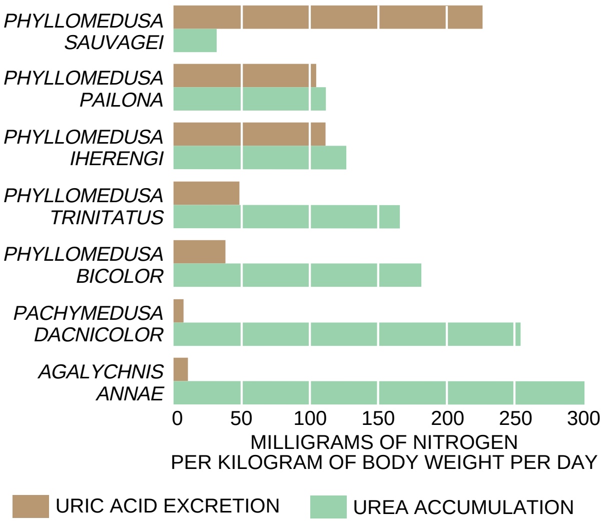
The ability to make and excrete uric acid turned out to be only one aspect of the tree frog's adaptations. We remeasured the evaporative water loss in *Phyllomedusa*—with care and patience this time—and found that *Phyllomedusa,* like *Chiromantis,* could reduce its water loss to very low levels when it was allowed to perch and behave normally.

We had observed that the frog uses each foot in turn to wipe its entire body. After this ritual, *rana verde* looks as if it consisted of plastic. Water dripped onto the creature's skin beads up, as it does on a waxed surface. Histological studies revealed the presence in the skin of a novel form of gland, which was interspersed between the mucous glands and poison glands that are typically found in frogs. The glands are tiny and numerous, about 30 per square millimeter of skin; they stain intensely when treated with lipid-soluble dyes. It is now clear that the waterproofing process involves the synchronous dis-



WIPING ITS SKIN with each of its four legs, in turn, allows *Phyllomedusa sauvagei* to completely cover itself (*left*) with the waxy substance that it secretes to prevent water loss. Once coated, the frog appears to be made of plastic. The protective lipid is produced in tiny skin glands that are shown magnified 150 times and dyed red in this micrograph (*right*).

charge of these glands, immediately followed by wiping. The coating, a rather heterogeneous mixture of lipids, is primarily wax ester. Like *Phyllomedusa,* insects and plants use a variety of waxes to retard water loss. Curiously, *Chiromantis* does not have lipid glands. The mechanism by which it prevents evaporation remains obscure.



URIC ACID PRODUCTION varies greatly. Frogs in arid regions, such as *P. sauvagei,* convert 80 percent of their nitrogen waste into semisolid uric acid, reducing the urea in the blood and saving water. Tropical frogs, such as *P. bicolor,* convert 20 percent. (*Pachymedusa dacnicolor* and *Agalychnis annae* are closely related to *Phyllomedusa.*)

Phyllomedusa and *Chiromantis* are able to survive in very hot environments. They accomplish this remarkable feat by controlling their body temperature. Before the summer rains, the air may exceed 40 degrees C. Body temperatures of *P. sauvagei* were found to track air temperature, except during the hottest parts of the hottest days. At these times, the frogs remained about 40 degrees C—two to four degrees cooler than the air and three to five degrees cooler than a thermometer constructed to match the size, shape and absorptive characteristics of the frog.

The frogs achieve such thermoregulation by controlling evaporation rates. Laboratory work has shown that the frogs can match evaporative heat loss to increases in ambient heat over a wide range of temperatures, wind speeds and relative humidities. The mechanism appears to be analogous to sweating. Microscopic observation of the skin shows periodic discharge from many of the gland ducts that dot the skin. We presume but have not conclusively demonstrated that mucous glands are responsible. Pharmacological studies of *Chiromantis* indicate the glands are controlled by sympathetic nerves that stimulate beta-adrenergic receptors.

The regulation of body temperature requires the most water during the dry period, just before the summer rains. The frogs obviate this problem because they can tolerate high body temperatures. On most days, there is no need to evaporate water for thermoregulation; on very warm days, thermoregulation is required for a few hours.

Like other frogs, *Phyllomedusa* and *Chiromantis* can store large volumes of water in the bladder and use it to offset loss through evaporation. When bladder reserves are exhausted, the frogs allow body temperatures to reach even higher levels, thereby reducing the need for evaporative cooling. *Phyllomedusa* appears to remain shaded while perched in trees, at least for most of the day. *Chiromantis,* however, can be observed sitting in the full sun. *Chiromantis* minimizes the effects of solar radiation by undergoing a dramatic color change. It forsakes its gray or brown protective coloration and instead becomes white to reflect sunlight.

We wondered if *Phyllomedusa* could take advantage of the light rain that frequently precedes a heavy downpour. Because its skin is waterproofed, the animal cannot absorb moisture that way. Therefore, we performed experiments in which water was dripped on its head. Astonishingly, we observed *Phyllomedusa* lift its head to gulp drops of water. Experiments using water with dyes incapable of diffusing through the skin showed coloration in the esophagus, stomach and small intestine—the water had clearly been ingested, not absorbed. *Phyllomedusa* is the only anuran known to drink.

The study of amphibians in deserts and semiarid regions has revealed extensive and diverse specializations for terrestrial existence. The emerging picture of anuran physiology defies the stereotype based on earlier studies of temperate species. Similarly, amphibians in other habitats have other capabilities—such as tolerance to freezing—that were unknown until recently [see "Frozen and Alive," by Kenneth B. Storey and Janet M. Storey; SCIENTIFIC AMERICAN, December 1990]. Despite this diversity, or perhaps because of it, there is evidence of a worldwide reduction in some amphibian populations and the extinction of others. The decline is not restricted to any particular habitat.

Some instances are clearly the result of human intervention. Various places that were once home to spadefoot toads in southern California are now housing tracts. We have seen widespread destruction of the habitat of *Phyllomedusa* in the Gran Chaco, where trees are cut for fuel. Air and water contamination, the introduction of predatory fishes and even consumption of frog legs contribute as well. In some cases, humans encourage survival. The great abundance of spadefoot toads in southwestern Arizona is probably the result of cattle tanks constructed by ranchers to catch runoff from thunderstorms. Such facilities serve as breeding sites.

Yet populations of anurans have declined or disappeared in relatively undisturbed or protected places. Although a portion of such events may represent natural fluctuation, concern grows that amphibian decline may indicate subtle environmental deterioration on a global scale. The complex life cycles and reproductive specializations of amphibians may make them doubly susceptible. Desert-adapted species appear no better able to withstand the effects of human activity than are their counterparts in wet conditions. Even creatures that spend most of the year underground must find abundant food and suitable aquatic breeding sites when they emerge. To prevent further extinctions, it is imperative that we understand the panoply of and limits to the devices that frogs and toads use to thrive in all habitats.

FURTHER READING

BEHAVIOR AND THERMAL RELATIONS OF THE ARBOREAL FROG *PHYLLOMEDUSA SAUVAGEI.* Lon L. McClanahan and Vaughan H. Shoemaker in *National Geographic Research,* Vol. 3, No. 1, pages 11–21; Winter 1987.

PHYSIOLOGICAL ECOLOGY OF AMPHIBIANS IN ARID ENVIRONMENTS. Vaughan H. Shoemaker in *Journal of Arid Environments,* No. 14, Vol. 2, pages 145–153; March 1988.

EXCHANGE OF WATER, IONS, AND RESPIRATORY GASES IN TERRESTRIAL AMPHIBIANS. Vaughan H. Shoemaker, with Stanley S. Hillman, Stanley D. Hillyard, Donald C. Jackson, Lon L. McClanahan, Philip C. Withers and Mark L. Wygoda in *Environmental Physiology of the Amphibians.* Edited by Martin E. Feder and Warren W. Burggren. University of Chicago Press, 1992.

Wire Pirates

by Paul Wallich, *staff writer*

Someday the Internet may become an information superhighway, but right now it is more like a 19th-century railroad that passes through the badlands of the Old West. As waves of new settlers flock to cyberspace in search of free information or commercial opportunity, they make easy marks for sharpers who play a keyboard as deftly as Billy the Kid ever drew a six-gun. Old hands on the electronic frontier lament both the rising crime rate and the waning of long-established norms of open collaboration.

It is difficult even for those who ply it every day to appreciate how much the Internet depends on collegial trust and mutual forbearance. The 30,000 interconnected networks and 2.5 million or more attached computers that make up the system swap gigabytes of information based on nothing more than a digital handshake with a stranger. (Even estimates of the Internet's size, compiled by SRI International, rely on the cooperation of system administrators around the globe.) Most people know, for example, that electronic-mail messages can be read by many people other than their intended recipients, but they are less aware that e-mail and other communications can be almost trace-

42
46
52
11
1
6
40
55
1
47
51
20
21
14
2
27
24
28
28
32
15
53
28
29
1
38
10
30
25
36

1. USERS
2. CRACKER
3. WIZARD
4. GURU
5. GOPHER
6. ARCHIE
7. VERONICA
8. WORLD-WIDE WEB
9. MOSAIC
10. FETCH
11. CRON DAEMON
12. FINGER DAEMON
13. FTP (FILE TRANSFER PROTOCOL) DAEMON
14. TELNET DAEMON
15. MAILER-DAEMON
16. ROOT
17. NOBODY
18. ZOMBIE PROGRAMS
19. SATAN
20. CRACK
21. WORM
22. COPS
23. TRIPWIRE
24. TROJAN HORSE
25. PACKET SNIFFER
26. DOT FILES
27. DEVICE FILES
28. SHELLS
29. NETWORK FILE SYSTEM (NFS)

INHABITANTS OF CYBERSPACE may be villains, victims or bystanders (some of them are shown at the right). Self-defense can be difficult in an environment ruled by trust.

Consumers and entrepreneurs crowd onto the information highway, where electronic bandits and other hazards await them

18
8
12
45
4
9
54
35
17
16
44
7
22
27
5
13
33
48
50
34
49
39
3
23
26
37
43
19
41
31

30. DOMAIN NAME SERVER
31. IDENT DAEMON
32. INTERNET-DAEMON
33. LOGIN DAEMON
34. NFS-DAEMON
35. PAGE DAEMON
36. PASSWORD-AUTHORIZATION DAEMON
37. QUOTA DAEMON
38. REMOTE-LOGIN DAEMON
39. STAT-DAEMON
40. SWAPPER
41. SYSLOG DAEMON
42. UPDATE DAEMON
43. USENET NEWS DAEMON
44. FINGER HACKER
45. SHOULDER SURFER
46. CELLULAR-PHONE CLONES
47. REMOTE PROCEDURE CALL
48. MAGIC COOKIES
49. SHELL SCRIPTS
50. FILTER
51. MIME
52. FIRE WALL
53. ROUTER
54. CHALLENGE-RESPONSE SYSTEM
55. KERBEROS

lessly forged—virtually no one receiving a message over the net can be sure it came from the ostensible sender.

Electronic impersonators can commit slander or solicit criminal acts in someone else's name; they can even masquerade as a trusted colleague to convince someone to reveal sensitive personal or business information. Of those few who know enough to worry about electronic forgeries, even fewer understand how an insidiously coded e-mail message can cause some computers to give the sender almost unlimited access to all the recipient's files. And mail-transfer programs are only one of the wide range of ways that an attacker can gain access to a networked computer. "It's like the Wild West," says Donn B. Parker of SRI: "No laws, rapid growth and enterprise—it's shoot first or be killed."

To understand how the Internet, on which so many base their hopes for education, profit and international competitiveness, came to this pass, it can be instructive to look at the security record of other parts of the international communications infrastructure. A computer cracker may become a "phone phreak" to avoid paying for the long-distance habit that computer intrusion sometimes requires, or he may take up phone phreaking as a related hobby (much as a poacher might both hunt and fish). Not only are some of the players the same, so are many of the basic design issues. Furthermore, engineers building each new generation of technology appear to make the same mistakes as their predecessors.

The first, biggest error that designers seem to repeat is adoption of the "security through obscurity" strategy. Time and again, attempts to keep a system safe by keeping its vulnerabilities secret have failed. Consider, for example, the running war between AT&T and the phone phreaks. When hostilities began in the 1960s, phreaks could manipulate with relative ease the long-distance network in order to make unpaid telephone calls by playing certain tones into the receiver. (One phreak, John Draper, was known as "Captain Crunch" for his discovery that a modified cereal-box whistle could make the 2,600-hertz tone required to unlock a trunk line.) The precise frequencies were "hidden" in technical manuals and obscure journal articles, but college students and others soon ferreted them out. Phreaks built so-called black, blue and red boxes to produce the required signals, and a small cottage industry flourished until the telephone company adopted methods that were less vulnerable to spoofing through the telephone mouthpiece.

Telephone credit cards underwent an evolutionary process that continues today. When the cards were first introduced, recalls Henry M. Kluepfel of Bell Communications Research, credit-card numbers consisted of a sequence of digits (usually area code, number and billing office code) followed by a "check digit" that depended on the other digits. Operators could easily perform the math to determine whether a particular credit-card number was valid. And phreaks could easily figure out how to generate the proper check digit for any given telephone number. The telephone company had to rely on detecting fraudulent calls as they occurred, tracking phreaks down and prosecuting them, a strategy that "never worked in the long term," according to Kluepfel.

In 1982 AT&T finally put in place a more robust method. The corporation assigned each card four check digits (the "PIN," or personal identification number) that could not be computed easily from the other 10. A nationwide on-line database made the numbers available to operators so that they could determine whether a card was valid.

Since then, theft of telephone credit-card numbers has become a matter of observation and guile rather than mathematics. "Shoulder surfers" haunt train stations, hotel lobbies, airline terminals and other likely venues. When they see a victim punching in a credit-card number, they transmit it to confederates for widespread use. Kluepfel noted ruefully that his own card was compromised one day in 1993 and used to originate more than 600 international calls in the two minutes before network-security specialists detected and canceled it. "I made a call from a coin phone and shielded the number from the scruffy-looking guy on my left, but I was unaware of the guy in the business suit on my right," he confesses.

Kluepfel cites estimates that stolen calling cards cost long-distance carriers and their customers on the order of half a billion dollars a year. The U.S. Secret Service has placed the total volume of telecommunications fraud at $2.5 billion; industry numbers range from $1 billion to $9 billion.

Somebody Else's Problem

Over the course of a generation, AT&T developed monitoring tools to thwart callers trying to evade toll charges. The corporation also used them to foil individuals who dialed into telephone switching systems to manipulate the facilities directly. Such access let phreaks forward other people's telephones to new locations, route calls around the world or even cut off one another's telephone service. After the Bell system breakup in 1984, however, AT&T was no longer the global policeman of the telecommunications world. In particular, tens of thousands of large and small companies that purchased PBXs (so-called private branch exchanges) to automate their internal telephone networks found themselves the targets of "finger hackers," but they had nothing like AT&T's expertise in self-defense.

"It's only insecure if it's configured wrong."
—Mark Abene, "Phiber Optik," hacker

The simplest way to commit finger hacking, says Kevin Hanley of AT&T, is "dial 1-800 and seven digits." At the other end of many toll-free lines sits a PBX remote-access unit, a subsystem that permits company employees to call their home office and then dial out from there to any number in the world. Most such units require a security code for outgoing calls, Hanley notes, but sometimes "the chairman doesn't want to remember a password." No one knows precisely what such oversight and self-indulgence cost. Industry estimates for PBX fraud range from a few hundred million to more than a billion dollars a year. "It's a bonehead crime," sneers Mark Abene, a hacker now serving a one-year sentence in federal prison for computer intrusion. (During an interview, Abene sketches out the architecture of digital interoffice signaling systems while complaining about what he considers unwarranted slurs on his character in internal telephone-company memos.)

Nevertheless, relentless trial-and-error dialing is highly profitable. Finger hackers can make tens or hundreds of thousands of dollars in calls before being detected. Most such crimes, Hanley notes, are committed by organized groups of criminals, who may even set up storefronts where they sell long-distance calls at cut price. Customers walk in, pay their money and tell an attendant the number they want to call. Less image-conscious thieves may sell calls out of a corner telephone booth.

Almost all these attacks can be circumvented by configuring a system to block calls to locations where a company does no business and by logging incoming calls to detect attempts at intrusion. Yet many PBX owners are unaware of the danger they face, Hanley says. Even those who do know their peril may not have managers on duty at night or during weekends, when most frauds occur. Of the 40 or 50 attendees at the security seminars that Hanley conducts, only "two or three know how toll fraud works," he asserts: "Maybe 30 or 40 percent know what it is, and that it's a bad thing."

Designers of the next big innovation in telecommunications, the cellular telephone, apparently ignored the lessons that their wire-bound predecessors so painfully learned. Leaving aside the fact that anyone with the right radio receiver can listen in on calls, the units are uniquely vulnerable to toll fraud. Every cellular telephone call begins with a broadcast of the telephone's serial number and billing number. Cellular switches check the pair against a database of working telephones to decide whether a call should go through. Unfortunately, these numbers are also the only information a thief needs to impersonate a legitimate caller.

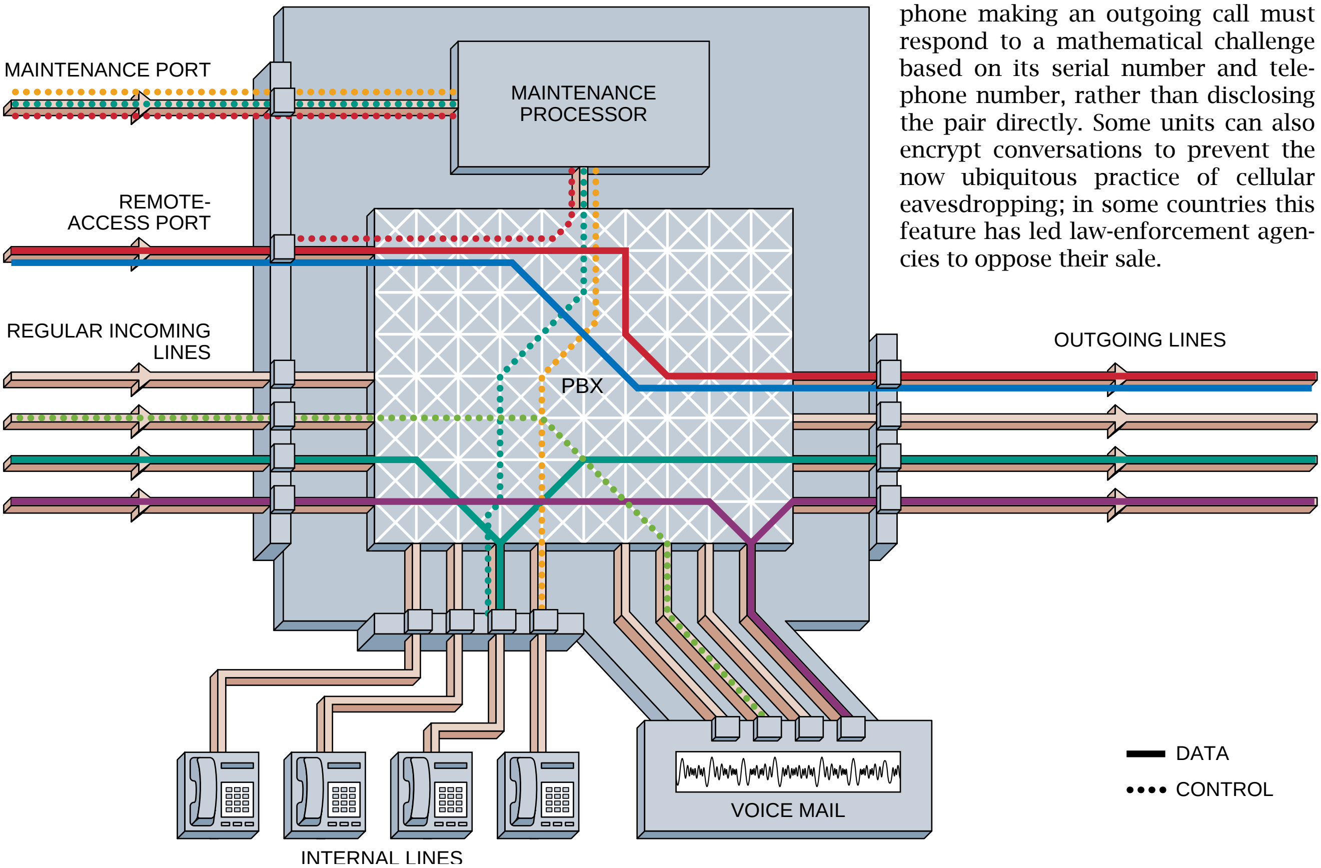
As early as 1984, communications expert Geoffrey S. Goodfellow (who got his start in the field when he broke into a computer at SRI and ended up with a job offer) wrote an article laying out a road map for cellular-phone fraud; only during the past year or two have thieves used the more sophisticated schemes he outlined. The fundamental problem, Goodfellow asserted, was that cellular-phone engineers underestimated both the technical expertise and the persistence of those who might want to subvert their equipment. He called for immediate replacement of current cellular-phone standards with more secure alternatives, but to little avail.

The simplest attack is known as cloning: reprogramming the serial number and telephone number of a pirate unit to that of a telephone currently in use by a legitimate customer. Although standards call for telephones to be built so that it will be impossible to change a serial number without irreparable damage, some early cellular-phone manufacturers supplied the numbers in a memory chip that could be popped out with a screwdriver. Others now place the information in an electrically programmable chip that can be accessed simply by applying the appropriate voltage to the telephone.

More sophisticated frauds adapt the same circuitry that allows a cellular phone to listen for incoming calls to decode the numbers as they are broadcast. A thief can then replay them to make calls that will be billed to others. The companies that carry cellular-phone traffic—and thus are financially responsible for fraudulent charges—have adopted a number of monitoring techniques to detect illicit calls; in addition, some have set their switching equipment to prevent long-distance cellular traffic to areas of the world where fraudsters often call (the countries that topped one carrier's list early in 1993, for example, were the Dominican Republic, Egypt, Pakistan, India, the former Soviet Union, El Salvador, China, Colombia, Mexico and Ghana).

Instead of just monitoring calling patterns, telecommunications engineers have been working on hardware fixes that could block the vast bulk of fraudulent cellular calls. In the fall of 1993 TRW announced a technique for analyzing the analog transmission "signature" of each telephone and storing it along with the serial number and telephone number. If a unit's serial number and telephone number do not match the signature, it must be fraudulent. Details of the characteristics that make up the signature are, of course, supposed to remain secure to prevent evildoers from devising countermeasures.

Such measures are a stopgap. The major cellular carriers are already preparing to replace the current analog cellular-phone system with a digital one. Most proposed digital-cellular standards incorporate protocols in which a telephone making an outgoing call must respond to a mathematical challenge based on its serial number and telephone number, rather than disclosing the pair directly. Some units can also encrypt conversations to prevent the now ubiquitous practice of cellular eavesdropping; in some countries this feature has led law-enforcement agencies to oppose their sale.



TELEPHONE SYSTEMS are vulnerable to a wide range of attacks. "Finger hackers" can dial into a toll-free remote-access port of a PBX and use a security code to make fraudulent calls (*blue*), or they can dupe a voice-mail system into transferring them to an outside line (*purple*). The voice-mail system can be suborned into divulging its contents, transmitting unauthorized messages or destroying stored data. Attackers can also dial into a PBX's maintenance port and disable security on the remote-access port (*red*), forward calls to outside numbers (*green*) or otherwise reconfigure the system.

During the same years that telephone companies were fighting the phone phreaks and cellular-phone architects were designing an estimated billion-dollar annual fraud bill for users and providers of service, computer scientists were laying the foundations of the Internet. Although initial funding came from the Department of Defense's Advanced Research Projects Agency, security was not really a concern, recalls ARPAnet veteran David J. Farber, now at the University of Pennsylvania. In the early days, only researchers had access to the net, and they shared a common set of goals and ethics, points out Eugene H. Spafford of Purdue University.

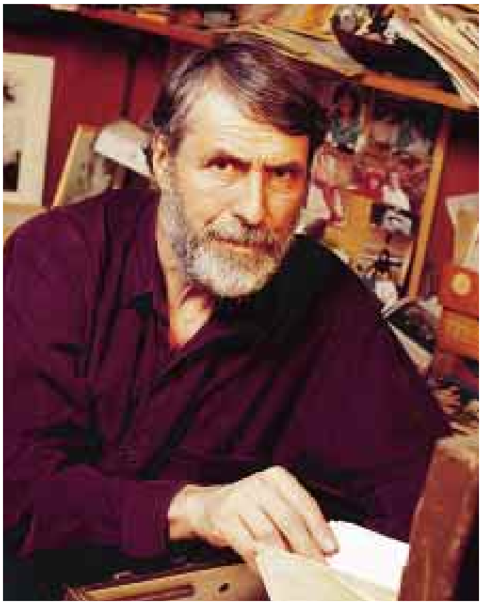
The very nature of Internet transmissions embodies this collegial attitude: data packets are forwarded along network links from one computer to another until they reach their destination. A packet may take a dozen hops or more, and any of the intermediary machines can read its contents. Indeed, many Internet packets start their journeys on a local-area network (LAN), where privacy is even less protected. On a typical LAN, computers broadcast each message to all the other computers attached to the network. Only a gentleman's agreement assures the sender that the recipient and no one else will read the message.

The Cyber-Neighborhood Goes Downhill

A lack of security on the ARPAnet did not bother anyone, because that was part of the package, according to Dorothy E. Denning, a professor of computer science at Georgetown University: "The concerns that are arising now wouldn't have been legitimate in the beginning." As the Internet grew, however, the character of its population began changing, and many of the newcomers had little idea of the complex social contract—and the temperamental software—guiding the use of their marvelous new tool.

By 1988, when a rogue program unleashed by Robert T. Morris, Jr., a Cornell graduate student, brought most Internet traffic to a halt for several days, a clear split had developed between the "knows" and the "know-nots." Willis Ware of Rand, one of the deans of computer security, recalls that "there were two classes of people writing messages. The first understood the jargon, what had happened and how, and the second was saying things like, 'What does that word mean?' or 'I don't have the source code for that program, what do I do?' "

"Crackers are ahead of legitimate administrators; they're way ahead of everybody else."
—Peter G. Neumann, SRI



Since then, the Internet's vulnerabilities have only gotten worse. Peter G. Neumann of SRI, a security researcher who edits the RISKS Forum, an on-line discussion of computer vulnerabilities, characterizes the situation as "unbelievably bad." Anyone who can scrounge up a computer, a modem and $20 a month in connection fees can have a direct link to the Internet and be subject to break-ins—or launch attacks on others. A few years ago the roster counted established names such as "mit.edu," "stanford.edu" or "ibm.com"; today you are as likely to find "mtv.com" or even "pell.chi.il.us," the nom du net of a battered PC-compatible with a hardware-store brass handle screwed into its case for portability.

Moreover, as the Internet becomes a global entity, U.S. laws become mere local ordinances. In European countries such as the Netherlands, for instance, computer intrusion is not necessarily a crime. Spafford complains—in vain, as he freely admits—of computer science professors who assign their students sites on the Internet to break into and files to bring back as proof that they understand the protocols involved.

Ironically, the more thoroughly computerized and networked an organization is, the more risk it may face when making its initial connections to the outside world. The internal network of a high-technology company may look much like the ARPAnet of old—dozens or even hundreds of users, all sharing information freely, making use of data stored on a few central file servers, not even caring which workstation they use to access their files.

As long as such an idyllic little pocket of cyberspace remains isolated, carefree security practices may be defensible. System administrators can safely configure each workstation on their network to allow connections with any other workstation. They can even set up their network file system to export widely used file directories to "world"—allowing everyone to read them—because, after all, the world ends at their corporate boundaries. Indeed, computer companies have made a practice of shipping their wares preconfigured so that each machine automatically shares resources with all its peers.

It does not take much imagination to see what can happen when such a trusting environment opens its digital doors to the Internet. Suddenly, "world" really means the entire globe, and "any computer on the network" means every computer on *any* network. Files meant to be accessible to colleagues down the hall or in another department can now be reached from Finland or Fiji. What was once a private lane is now a highway open to as much traffic as it can bear.

Dan Farmer of Sun Microsystems and Wietse Venema of the Eindhoven University of Technology report that even some of the most respected domains on the Internet contain computers that are effectively wide open to all comers—the equivalent of a car left unattended with the engine running. A recent security alert from CERT (the Computer Emergency Response Team), a clearinghouse for security-related problems based at Carnegie Mellon University, illustrates the point. The alert disclosed that all Sun workstations with built-in microphones had been preset to give their audio input "world-readable" status. Anyone who could gain network access to such a workstation could listen to conversations nearby.

Another alert warned system administrators that the memory buffers that store images displayed on workstation screens might also be preset to world-readable status, as might those that store characters typed on the keyboard. Patient attackers have watched until someone logged in to a privileged account and then simply read the password out of the computer's memory. Once intruders have gained such "root" access, they can masquerade as any legitimate user and read, alter or delete any files. They can also install programs to help them invade other computers and even modify system logs to erase signs of their intrusion.

Even if newcomers to the net try to secure their systems, they do not always have an easy time finding the information they need. Computer hardware and software vendors are often loath to talk about security problems, Neumann says. And CERT generally issues net-wide advisories only after manufacturers have developed a definitive fix—weeks or months later, or sometimes never. Most advisories do not explain the security flaw in question; instead they name the software and hardware involved and specify the modifications that should be made to reduce the chances of intrusion.

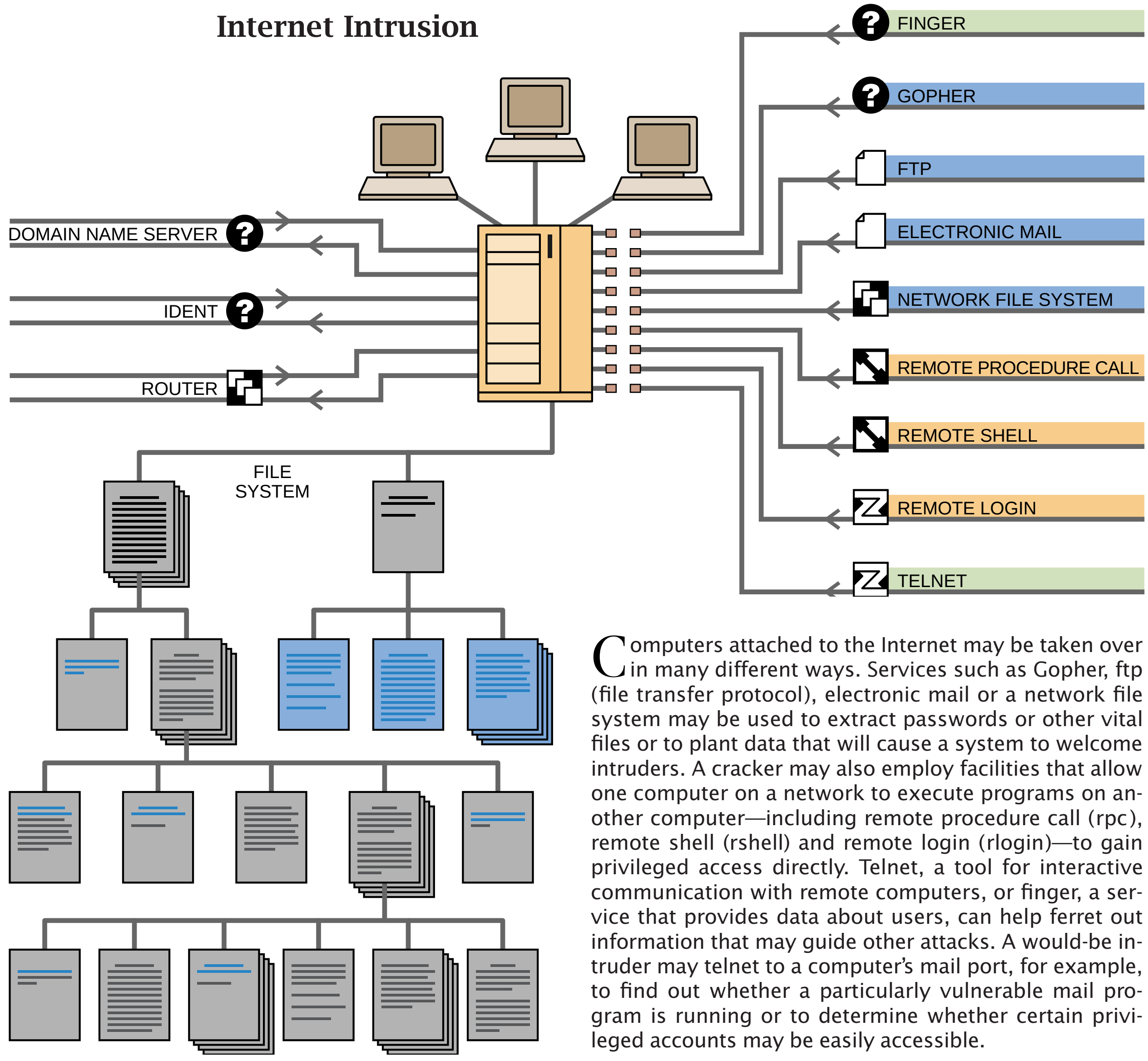
This policy keeps potentially useful information away from those crackers who are not well connected within the illicit community. But it also keeps many novice system administrators in the dark, Neumann complains: "You find out by having a buddy who's a system administrator somewhere else." Neumann estimates that between half and three quarters of the security holes currently known to hackers have yet to be openly acknowledged. "People don't know the risks," Spafford comments. "They know the benefits because people talk about those."

Many of those benefits come from programs such as Gopher, World-Wide Web or Mosaic, which help people navigate through cyberspace in search of information. A single menu selection or click of a mouse may take a researcher from a computer in Minnesota to another in Melbourne or Zurich. Files containing U.S. census data, pictures of the aft plumbing of the space shuttle and lists of British pubs or of artificial-intelligence software are available free for the finding. Such tools, many of them just a gleam in researchers' eyes a few years ago, are drawing tens of thousands of people to the Internet.

Yet that rapid evolution may have leapfrogged steps that could have contributed to security, Spafford notes. Cybersurfers are relying on "the first or second version" of programs originally written to test ideas rather than to provide industrial-strength services.

The popular Gopher program, according to CERT advisories and other sources, has security flaws that make it possible to access not only public files

Internet Intrusion



Computers attached to the Internet may be taken over in many different ways. Services such as Gopher, ftp (file transfer protocol), electronic mail or a network file system may be used to extract passwords or other vital files or to plant data that will cause a system to welcome intruders. A cracker may also employ facilities that allow one computer on a network to execute programs on another computer—including remote procedure call (rpc), remote shell (rshell) and remote login (rlogin)—to gain privileged access directly. Telnet, a tool for interactive communication with remote computers, or finger, a service that provides data about users, can help ferret out information that may guide other attacks. A would-be intruder may telnet to a computer's mail port, for example, to find out whether a particularly vulnerable mail program is running or to determine whether certain privileged accounts may be easily accessible.

but private ones as well. ("It's only insecure if you configure it wrong," insists Abene, who will return to his job as system administrator for an on-line service in New York City when he is released from prison.) Running Gopher initiates a dialogue between a client program on the user's machine and a Gopher server somewhere on the Internet. The server presents the client with a menu of choices for information, along with a set of "magic cookies"—shorthand specifications for the location of additional information.

If the user wants to delve deeper into a particular subject, the client program sends the server the cookie corresponding to that piece of information. Yet it is relatively easy to modify the cookie so that it specifies the location of information on the server's machine that is supposed to be kept private. An unsuspecting Gopher server will deliver those private files without checking whether the cookie it receives matches one of the items in the menu it presented. Although Gopher servers can readily be confined so that they have access only to public information, by default they have free rein.

By Bits Deceived

Failure to check the propriety of commands is a common oversight—as Othello with Iago, computers on the Internet trust not wisely but too well. E-mail, one of the net's most basic services, sets the tone: an electronic letter consists simply of a text file containing a header specifying the sender, addressee, subject, date and routing information, followed by a blank line and the body of the message. Although mail programs generally fill in the header lines accurately, there is little to prevent a whimsical or malicious person from inserting whatever information they please. A message from "president @whitehouse.gov" could as easily originate from a workstation in Amsterdam as from the Executive Office Building in Washington, D.C. Forging e-mail is "trivial," Farber asserts; what makes such forgeries a problem as the Internet grows is that the incentive for successful forgeries does so, too, as do the dangers of being taken in. Companies and individuals have already begun doing business via e-mail; real money and goods change hands on the basis of electronic promises.

Computer scientists have developed protocols for verifying the source of e-mail messages, but spoofers are also improving their techniques. Correspondence on Usenet discussion groups such as "comp.security.misc" illustrates this coevolution: some security-minded system administrators have advocated the use of "IDENT daemons." If a spoofer connects to a mail server and offers a false identification (the first step in sending a forged message), the mail server can query the IDENT daemon on the spoofer's machine.

"People are attacking individual systems, not the net per se."
—Dorothy E. Denning, Georgetown University

Others disparage IDENT; they point out that the name returned by the daemon is only as trustworthy as the computer it runs on. Once hackers have gained control of a machine—either by breaking in or because they own it—they can configure the IDENT daemon to respond to queries with whatever name they please.

Some system administrators are meeting the threat of such deceptions by barring connections to their computers from untrustworthy parts of the Internet. Each range of numbered addresses on the Internet corresponds to a particular organization, or domain, and so it is simple to refuse connections from computers in a domain believed to serve as a vehicle for hackers.

Even this step has a countermeasure. Spafford points out that most machines rely on "domain name servers" to translate back and forth between numbered network addresses and domains such as "xerox.com" or "umich. edu." But the name servers are just ordinary computers. They are vulnerable to deception or intrusion, and so the road maps they provide can be rewritten to serve deceitful ends. A cracker can modify the name server's database so that it tells any computer querying it that the address belonging to, say, "evil.vicious.hackers.org" is instead that of "harvard.edu." A computer that accepts connections from Harvard University will then allow the hackers in as well. Indeed, Spafford laments, it is almost impossible for any program to know for sure where the data packets reaching it over the Internet really come from or where the packets it sends out are going.

Another class of security problems comes not from misplaced trust in domain name servers or IDENT daemons but rather from the same versatility that makes networked computers so useful. Perhaps the best example of this is the "sendmail bug," a disastrous loophole that has reappeared time after time in the history of the net.

The bug arises because most mail programs make it possible to route messages not only to users but also directly to particular files or programs. People forward mail, for instance, to a program called vacation, which sends a reply telling correspondents that the intended recipient is out of town. Many people also route mail through filter programs that can forward it to any of several locations depending on sender, subject matter or content.

But this same mechanism can be subverted to send electronic mail to a program that is designed to execute "shell scripts," which consist of a series of commands to perform system functions, such as extracting information from files or deleting all files older than a certain date. This program will then interpret the body of the message as a script and will execute any commands it contains. Those commands could cause a copy of the receiving computer's password file to be sent to an intruder for analysis, fashion a subtle back door for later entry or simply wreak havoc on the recipient's stored data. Mail sent to certain files can have similar effects.

Some fixes for the latest incarnation of the sendmail bug have been published on the Internet and presumably have been implemented by most system administrators who saw them, but many systems remain vulnerable. Furthermore, other programs that process electronic mail contain analogous holes.

Even more ominous is the fact that e-mail is by no means the sole way to plant uncontrolled data in a victim's computer. Steven M. Bellovin, a researcher at Bell Labs, points out that Gopher and other information-retrieval programs also transfer large, potentially ill-identified files. A hacker would have to go to some trouble to set up a corrupt Gopher server and would even have to stock it with useful information to entice people into making connections to it. “I won’t be surprised when it starts happening,” Bellovin says.

Walls in Cyberspace

If the Internet, storehouse of wonders, is also a no-computer’s-land of invisible perils, how should newcomers to cyberspace protect themselves? Security experts agree that the first layer of defense is educating users and system administrators to avoid the particularly stupid mistakes. People still tape passwords to their keyboards or use no password at all for privileged computer accounts. One graduate student, pressed into service as administrator for a cluster of workstations at the University of Michigan, found that a simple password-guessing program could compromise a quarter of her users’ accounts. Five of 80 users had chosen their names as passwords. Some administrators have installed programs that reject passwords based on dictionary words or obvious personal identifiers, but their use is far from widespread.

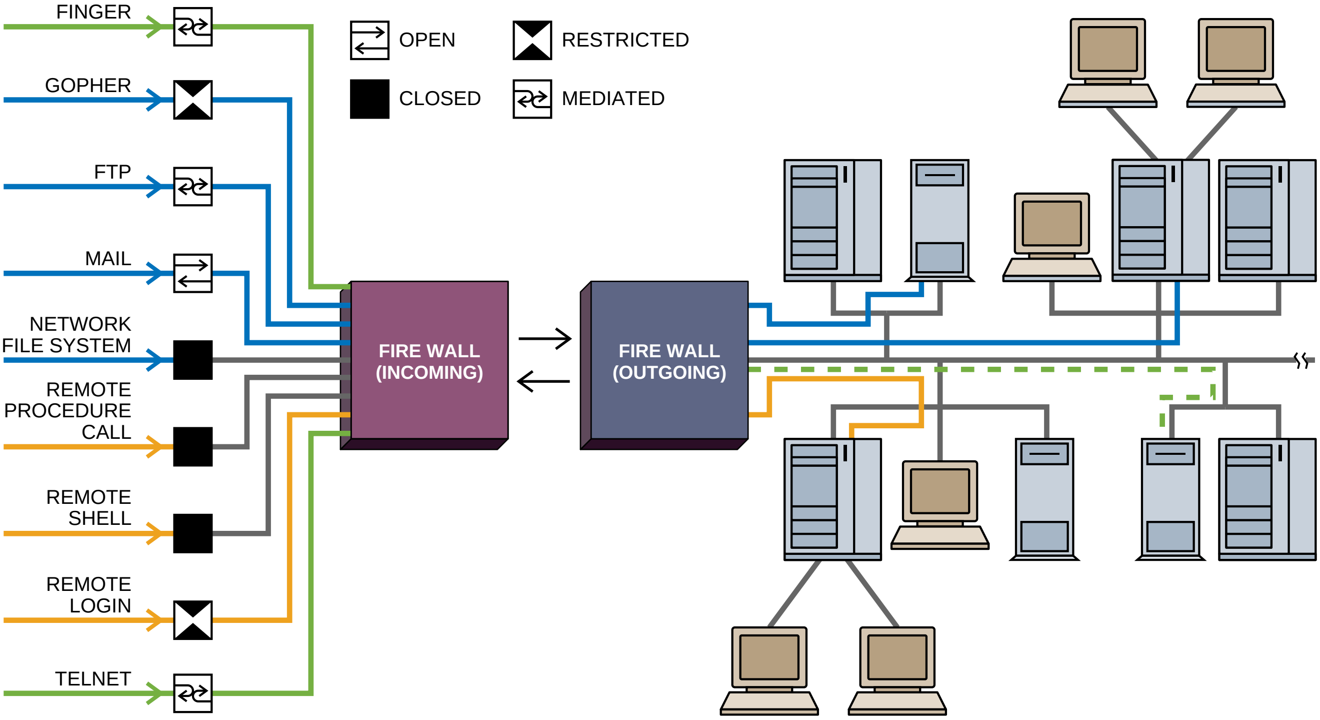
The next level of defense is the so-called fire wall, a computer that protects internal networks from intrusion. Most major companies have long since installed fire walls, and many universities are adopting them as well. Fire walls examine all the packets entering and leaving a domain to limit the kinds of connections that can be made from the Internet at large. They may also restrict the information that can be passed across those connections. “Anyone who would connect a corporate network directly to the Internet should be fired,” Farber asserts.

Proposing a fire wall and constructing it are two different matters. Users would like to have access to all possible Internet services. But that desire encounters a harsh reality: “Some things you can’t do securely,” maintains Marcus J. Ranum of Trusted Information Systems. Ranum, who helped to install the fire wall for “whitehouse.gov,” names Gopher and Mosaic as two programs whose trusting nature defies the attempts of a fire wall design to provide safety. In such cases, he argues, security experts must be content to minimize risk rather than eliminate it entirely.

At a bare minimum, a fire wall must pass mail, according to Bellovin (even though mailers may be demonstrably insecure). After that, users want to be able to log in to machines elsewhere on the Internet and to retrieve files from public archive sites or from the directories of colleagues at other institutions.

To perform these functions, AT&T built a fire wall consisting of two dedicated computers: one connected to the Internet and the other connected to the corporation’s own network. The external machine examines all incoming traffic and forwards only the “safe” packets to its internal counterpart. In addition, it will accept outgoing traffic only from the internal gateway machine, so an attacker attempting to transfer information illicitly out of AT&T’s domain would be unable to do so without subverting the internal gateway. The internal gateway, meanwhile, accepts incoming traffic only from the external one, so that if unauthorized packets do somehow find their way to it, they cannot pass.

Other services are more problematic. Workers would like to be able to log in to their office computers from anywhere on the Internet, for instance. Any intermediate computer relaying traffic over the Internet might have been com-



NETWORK FIRE WALL prevents malicious data packets from wreaking havoc on trusting computers. In some cases, the fire wall may allow access only from trusted locations on the Internet to particular machines inside the fire wall. Or it may allow only demonstrably “safe” information to pass (perhaps permitting users to read the e-mail from remote locations but not to run privileged programs). In other cases, it is impossible to distinguish safe use of a particular network service from unsafe use, and so all requests must be blocked. The fire wall may also provide a substitute for some network services (such as finger, telnet or ftp) that performs most of the same functions but is not as vulnerable to penetration.

promised, however, and could be reading packets (including those containing passwords) as they go by. In fact, in two separate incidents in October 1993, hackers gained access to Panix, a public-access Internet site in New York City, and to BARRNet, an Internet carrier in California, and installed "packet sniffers." These programs watched all the data going by and recorded user names and passwords as people logged in to (at least) hundreds of other computer systems, according to system administrators at Panix.

Such attacks render conventional passwords "obsolete," Ranum asserts. Instead safe connections to machines inside a fire wall require a different kind of authentication mechanism, one that cannot be recorded by a sniffer and then replayed to gain unauthorized access. Two methods are already in limited use: the "one-time" password and "challenge-response."

To use one-time passwords, a worker simply carries a list of them. Reuse indicates that an intrusion attempt is in progress. Challenge-response systems have no list of passwords; instead they require an answer to a random query before allowing access. Most often the query consists of a number that must be mathematically transformed by a secret key known only to authorized users. Most people cannot multiply 100-digit numbers in their heads, so commercial challenge-response equipment usually employs a "cryptographic calculator," primed with the key and activated by a shorter sequence that a person can remember.

Encryption Is Key

If passwords should traverse the Internet only in encrypted form, what about other sensitive information? Standardization efforts for "privacy-enhanced" e-mail have been under way for more than five years, but widespread adoption lies well in the future. "The interoperability problem is nasty," Ware says, unless everyone has software that can handle encrypted messages, it is of little use to anyone.

Encryption could provide not only privacy but authentication as well: messages encoded using so-called public-key ciphers can uniquely identify both recipient and sender. But encryption software in general remains at the center of a storm of political and legal controversy. The U.S. government bars easy export of powerful encoding software even though the same codes are freely available overseas.

"At the civil level, constraints. At the criminal level, constraints. At the technological level, constraints."
—Donn B. Parker, SRI

Within the U.S., patent rights to public-key encryption are jealously guarded by RSA Data Security, a private firm that licensed the patents from their inventors. Although software employing public-key algorithms has been widely published, most people outside the U.S. government cannot use it without risking an infringement suit.

To complicate matters even further, the government has proposed a different encryption standard, one whose algorithm is secret and whose keys would be held in escrow by law-enforcement agencies. Although many civil libertarians and computer scientists oppose the measure, some industry figures have come out in favor of it. "You can't have absolute privacy," Parker says. "A democracy just can't operate that way."

The question is not whether cyberspace will be subjected to legislation but rather "how and when law and order will be imposed," Parker says. He predicts that the current state of affairs will get much worse before the government steps in "to assure privacy and to protect the rights people do have."

Others do not have Parker's confidence in government intervention. Ranum foresees an Internet made up mostly of private enclaves behind fire walls that he and his colleagues have built. "There are those who say that fire walls are evil, that they're balkanizing the Internet," he notes, "but brotherly love falls on its face when millions of dollars are involved."

Denning counts herself among the optimists. She lends her support to local security measures, but "I don't lose any sleep over security," she says. Farber, also cautiously optimistic, sees two possible directions for the Internet in the next few years: rapid expansion of existing services, or fundamental reengineering to provide a secure base for the future. He leaves no doubt as to which course he favors. Spafford is likeminded but gloomier. "It's a catch-22," he remarks. "Everyone wants to operate with what exists, but the existing standards are rotten. They're not what you'd want to build on."

Even if computer scientists do redesign the Internet, he points out, putting new standards in place may be impossible because of the enormous investment in old hardware and software. So much of the Internet rests on voluntary cooperation, he observes, that making sweeping changes is almost impossible.

Then again, Ware counters, perhaps piecemeal evolution may be the only possibility. No single organization understands the idea of a national information infrastructure well enough to be put in charge, he contends: "There's no place to go and say, 'Here's the money, work out the problems.' There aren't even three places, and I'm not sure there should be."

In the meantime, the network grows, and people and businesses entrust to it their knowledge, their money and their good names.

FURTHER READING

PRACTICAL UNIX SECURITY. Simson Garfinkel and Gene Spafford. O'Reilly and Associates, 1992.

ZEN AND THE ART OF THE INTERNET. Brendan P. Kehoe. Prentice-Hall, 1993.

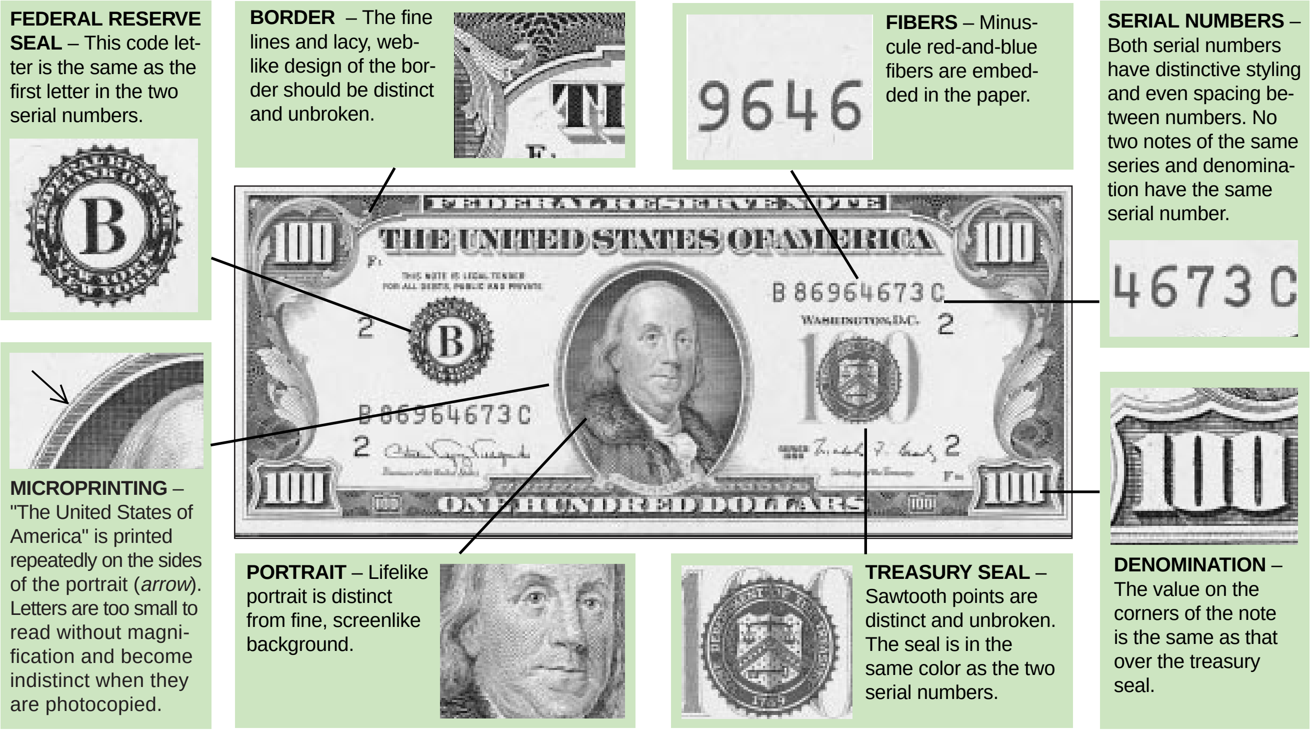
IMPROVING THE SECURITY OF YOUR SITE BY BREAKING INTO IT. Dan Farmer and Wietse Venema. Available by ftp from win.tue.nl as /pub/security/admin-guide-to-cracking.Z

FORUM ON RISKS TO THE PUBLIC IN COMPUTERS AND RELATED SYSTEMS. Available via usenet newsgroup comp.risks or by e-mail from risks-request@csl.sri.com. Back issues are available by ftp from crvax.sri.com in the RISKS: directory.

Usenet newsgroups that cover security: comp.security.misc; comp.security. unix; alt.security

Computer Emergency Response Team advisories are available by ftp from cert.org

Security-related information available by ftp from tis.com (Trusted Information Systems) and research.att.com



FEDERAL RESERVE SEAL – This code letter is the same as the first letter in the two serial numbers.

BORDER – The fine lines and lacy, web-like design of the border should be distinct and unbroken.

FIBERS – Minuscule red-and-blue fibers are embedded in the paper.

SERIAL NUMBERS – Both serial numbers have distinctive styling and even spacing between numbers. No two notes of the same series and denomination have the same serial number.

MICROPRINTING – "The United States of America" is printed repeatedly on the sides of the portrait (*arrow*). Letters are too small to read without magnification and become indistinct when they are photocopied.

PORTRAIT – Lifelike portrait is distinct from fine, screenlike background.

TREASURY SEAL – Sawtooth points are distinct and unbroken. The seal is in the same color as the two serial numbers.

DENOMINATION – The value on the corners of the note is the same as that over the treasury seal.

C NOTE PROTECTION is afforded by a variety of design features for this denomination as well as other notes. Because of the prospect of a wave of "casual" counterfeiting done on color copiers and desktop publishing systems (similar to the one used to produce this likeness of a bill), the treasury department may have to consider further steps to combat funny money.

Making Money

Desktop counterfeiting may keep the feds hopping

The era of the desktop counterfeiter is coming. By 1995 the ability to manipulate and print color images on personal computer–based publishing systems that cost less than $5,000 could mean that someone who is pinched for cash might forgo a trip to the nearest automatic teller. True professionals may lament this growing trend toward amateurism. In past years illicit money manufacture required a craftsman with skills at executing lithographic reproduction. That care was needed to convey the right nuances of highlight and detail so that a salesclerk wouldn't give Andrew Jackson's mug a second glance.

The U.S. Secret Service joins in the lament. It has found that the fastest-growing means of making illegal tender is by ink-jet printer. A computer attached to the printer can be equipped with an electronic scanner to capture the image of a note, which can then be enhanced with software. The value of such bills discovered by the Secret Service was a piddling $66,000 or so from October 1992 to the end of June 1993. But these notes, all of low denominations, turned up at 205 places throughout the country. One forgery apparently had nothing to do with another. "The low number and value of the notes indicate that most of them can be attributed to casual counterfeiters," observed a recent National Research Council report that recommends changes in currency to combat counterfeiting.

The document, *Counterfeit Deterrent Features for the Next-Generation Currency Design,* also ponders the disturbing prospect that, left unchecked, the growth of funny money made on both color copiers and computer printers could continue to double every year, as it has since 1989. Notes from these sources accounted for only $6 million to $8 million in 1992. But such growth would mean bills with a total value of $1.6 billion to $2 billion could be infused into the economy during the year 2000. That amount would overshadow the roughly $160 million of U.S. counterfeit bills estimated to have been minted during the 1993 federal budget year.

All these sums are paltry compared with the more than $300 billion in cash that now circulates. But a billion-dollar crop of counterfeit bills produced by thousands of hobbyists making money in their spare time could paralyze enforcement efforts by the Secret Service. The agency currently removes about $25 to $70 million a year before it ever reaches circulation. That task would become more arduous should the agency be confronted with flocks of weekend counterfeiters instead of a few illicit mints cranking out stacks of bills.

The U.S. already incorporates a number of features to discourage counterfeiters. Special paper, engraving techniques and printing methods distinguish real bank notes from fake ones. Since 1990 the treasury has also incorporated a "security thread," a 1.5-millimeter-wide metallized polyester thread on which is printed the letters "USA" and the denomination of the bill. Still, U.S. currency has not undergone a fundamental redesign since 1929.

The NRC panel concluded that the threat from color printers and copying machines warrants a new round of copy-protection measures: the use of a pattern that would produce a series of wavy lines (a moiré) when copied or inks whose color changes on the original when viewed from different angles (the color on a copy would remain unchanged). At the same time, a law might be considered that would require copier manufacturers to place an identifying number in any image reproduced. (In fact, legislation floating around the House Banking Committee suggests that the government give priority to redesigning the dollar.)

The NRC report examines further measures, ranging from holograms and laminated papers to insertion of small metal particles that would reflect a radar beam. The treasury might also talk to some foreign central bankers. Dutch currency now has bar codes. Australia prints bills on a plastic substrate and plans to employ diffractive optical elements called kinegrams.

Even after careful study, a good deterrent will be hard to find—as it has been for centuries. Fear of death is not enough to thwart the efforts of an intrepid moneymaker.

As the NRC report noted, counterfeit-

Inside Story

Doctors like to look inside things. So do scientists and engineers. Yet the computer graphics many rely on to make sense of the ever growing mountains of data spewed from their instruments and simulations have been annoyingly superficial. Despite tremendous increases over the past decade in speed and realism, nearly all graphics software can still draw only geometric surfaces—fine for designing a turbine blade but inadequate for peering inside a three-dimensional ultrasound image of a fetus. One solution, a technique known as volume rendering, is finally wending its way out of academia and into the marketplace.

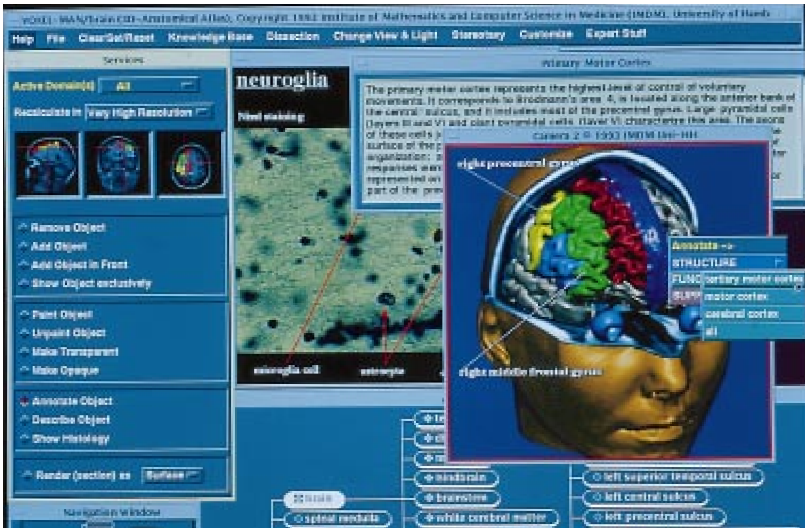
Volume rendering enables researchers to view three-dimensional data sets in ways that would choke even the most powerful supercomputer running surface-rendering software. A computed tomographic (CT) scan of a patient's head, for example, produces a stack of two-dimensional images that are then melded together to form a digital 3-D model. Surface graphics programs would wrap contour lines around the model and connect them to create a detailed but empty shell composed of thousands of polygons.

Volume graphics software, such as that developed by Advanced Visual Systems (AVS) in Waltham, Mass., works like a fluoroscope to reveal inner detail. It stores the entire model in memory as a collection of "voxels" (volume elements akin to pixels), then draws a screen image by casting imaginary rays from behind the model. Each ray accumulates data from every voxel it passes through and then splats on the screen in the appropriate color. The software lets the user change the angle of view, strip away certain parts—the skull, say—or cut away a section and look inside.

"Volume rendering is relatively new to the market," says Hambleton D. Lord, AVS's director of product marketing. University researchers have used the technique for several years, however. Jeanne B. Lawrence, a cell biologist at the University of Massachusetts Medical Center, made the cover of *Science* last year when she used volume visualization with images of cell nuclei to discover that RNA is collected into a few dozen "hot spots" within the nucleus. Thomas R. Nelson, a physicist-turned-radiologist at the University of California at San Diego, has been working with T. Todd Elvins of the San Diego Supercomputer Center (which has developed volume graphics software for Unix machines that it distributes free of charge over the Internet) on a system to produce useful 3-D ultrasound images.

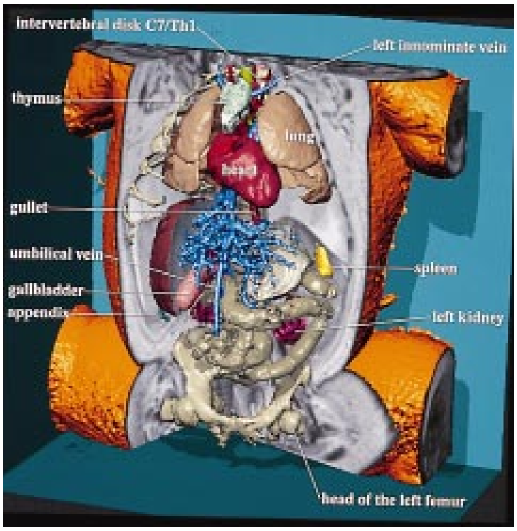
Lord sees several potentially lucrative markets for commercial volume-rendering packages. "The medical imaging market is growing very rapidly," he says, as 3-D diagnostic techniques like CT scanning, magnetic resonance imaging and positron emission tomography grow ever more commonplace. "Another heavy user is the oil and gas exploration market. They take their thumper trucks out into the field or use towed sonar arrays to collect massive amounts of seismic data." Extracting and rendering the surfaces of such large data sets is unfeasible.

AVS is competing with some of the giants in the computer industry, notably IBM and Silicon Graphics, for these estab-



The primary motor cortex represents the highest level of control of voluntary movements. It corresponds to Brodmann's area 4, is located along the anterior bank of the central sulcus, and it includes most of the precentral gyrus. Large pyramidal cells (layers III and V) and giant pyramidal cells (layer V) characterize this area.

KARL HEINZ HÖHNE ET AL. UNIVERSITY OF HAMBURG



KARL HEINZ HÖHNE ET AL. *University of Hamburg*

***VOLUME GRAPHICS** builds computer models out of 3-D "voxels" rather than 2-D surfaces. Viewers can then dissect complex volumetric data, such as this 3-D ultrasound image* (right) *of an 18-week-old fetus. Doctors could compare such images with a 3-D* ing was a capital offense in England from 1697 to 1832. More than 300 English counterfeiters died at the end of a rope, and many others were remanded to penal service in Australia. Yet forgery continued. The law relaxed only under the pressure of public indignation provoked by the Bank of England's failure to get on with the work of devising a bank note resistant to counterfeiting. Take note, Lloyd Bentsen. —*Gary Stix*

lished customers even as new markets emerge. Environmental analysis will be one. The U.S. Geological Survey has purchased software from AVS to help it build a three-dimensional computer model of the flow from a 14-kilometer-long tunnel designed to carry Boston's sewage to the sea. Nondestructive structural testing could be another. Failure Analysis Associates in Menlo Park, Calif., has used volume visualization to look for flaws in everything from coffeemakers and valve assemblies to the solid rocket boosters of Minuteman missiles. And voxel-based imaging is practically a prerequisite for laser scan confocal microscopes, instruments that let biologists peer into living cells in a manner analogous to the way CT lets doctors scan the interior of patients.

"By the end of this century, volume-rendering engines will be as popular as surface-based engines," predicts Arie Kaufman, a computer scientist at the State University of New York at Stony Brook, who helped pioneer volume graphics. "In the next century, they'll beat them." —*W. Wayt Gibbs*

THOMAS R. NELSON, DOLORES PRETORIUS *University of California, San Diego*

anatomical atlas **(left *and* center)** ***assembled by researchers at the University of Hamburg.***

The Last Frontier

Researchers explore fiber's outer reaches

If the much heralded multimedia revolution ever comes to pass, the telecommunications infrastructure needed to carry *Jurassic Park* on demand, along with the image of a Harvard history professor lecturing to children in a Harlem high school, could quickly be overwhelmed. The capacity required for digital video services is 25 to more than 5,000 times that required for voice conversations routinely channeled through telephone lines.

For this reason, a few researchers at places like AT&T Bell Laboratories continue to work on problems other than devising hardware and software that will let a teenager in Miami take on pimply counterparts all over the country at real-time games of Sonic the Hedgehog.

One of the highest rates of optical transmission over a single fiber on a commercial network is 3.4 gigabits a second. That rate, achieved on AT&T's trunk-line transmission links, is less than 1 percent of the theoretical capacity of 20 or more trillion bits per second. Squeezing more bits into each individual fiber remains one of the grand challenges of telecommunications. It is also one that is being approached with caution: if no one wants multimedia, there will be no need for a multibillion-bit optical pipe.

Bell Labs is still involved in developing advanced fiber-optic networks. But it is also hedging its bet by seeking the shelter of a government-funded consortium for research on sending information at blinding speeds of 100 gigabits or more each second over a single fiber. "In the new climate, this far-term research is not going to be heavily supported by any company, even AT&T," says Vincent W. S. Chan, a researcher at the Massachusetts Institute of Technology, who directs the consortium in which AT&T participates. It is one of several consortia in optical communications that are funded by the Advanced Research Projects Agency.

Besides AT&T, the collaboration also includes M.I.T. and Digital Equipment Corporation. The group, funded at $32 million, is exploring the boundary at which electronic communications must give way to networks through which traffic not only moves but also is processed in the form of light. Today fiber is essentially a passive medium. Electronic signals are transferred to an optical format for transfer over a network. But each time a signal goes on or off a fiber, it has to be processed electronically. Many of the individual optoelectronic transmission and receiving components will probably wink out at rates around 50 gigabits per second.

Some all-optical technologies for harnessing light waves are being deployed by telecommunications companies. Last summer AT&T began commercial use of optical amplifiers doped with the rare-earth element erbium on a short 150-mile link in California. The amplifiers revivify an optical signal without converting it to an electronic one and are capable of boosting higher-speed data signals than electronic systems can handle.

Erbium amplifiers will be essential to the high-speed networking research undertaken within the ARPA-funded collaborations. They will help the consortia grapple with the problem of how to get more capacity from a fiber by apportioning streams of video, voice or data onto many different wavelengths of an optical signal, a technique called wavelength division multiplexing (WDM).

An erbium amplifier can boost different wavelengths simultaneously. The consortium in which AT&T has been involved is also helping to develop a semiconductor laser that can rapidly be tuned to different wavelengths and a device for switching data from one wavelength to another.

The use of different wavelengths is a form of parallel transmission. WDM assembles separate wavelengths into a high-capacity signal. The arrangement is analogous to a bundle of wires in a sheath: each wavelength carries a lower number of bits than does the aggregated signal. Communications equipment at a telephone company's switching office need handle only the lower carrying capacity of each wavelength, not the entire bandwidth of the whole bundle.

A number of laboratories are also developing a kind of network that uses the full capacity of any signal that may be racing along at tens or hundreds of gigabits per second. Instead of WDM's parallel transmission, this time-division multiplexed network operates in a serial fashion. Each modulated pulse follows the other, forming a string of tiny photonic boxcars. One boxcar may transport video; another may carry computer data. A pulse, corresponding to a bit of information, must be pulled on and off this fast-moving freight train (multiplexed or demultiplexed) at intervals of perhaps 10 trillionths of a second. In September, Nippon Telegraph and Telephone (NTT) announced that it had demultiplexed a 100-gigabit-per-second signal that traveled across 50 kilometers into 16 channels, each carrying 6.3 gigabits per second.

NTT used an interferometer, called a nonlinear optical loop mirror. In it, light pulses interfere constructively in a loop of fiber to emit a pulse that corresponds to a demultiplexed 6.3-gigabit signal.

Researchers at British Telecom Laboratories last year handled another piece of the all-optical signal-processing problem. They demonstrated how the pulsations of a 40-gigabit-per-second optical stream of data transmitted over a fiber could be synchronized with the pulsing of a laser at a receiving point through the interactions of the two light beams. This laser, turning on and off at a few trillionths of a second, could then serve as a timing source, or clock, to control the optical circuitry required for demultiplexing the data.

Much more work is needed. For optical networks to match their electronic cousins, the receiving equipment must determine the destination, or address, for which the digital information is intended. It must ascertain whether incoming information should be plucked from the network or routed along to a subsequent receiving point. Mohammed N. Islam, a professor at the University of Michigan, has devised a rudimentary optical processing device that uses solitons to read an address. Solitons are light pulses that do not become distorted when transmitted even over great distances. They may form a key part of networks that will let the participants in the Sonic Hedgehog game compete through the rapid-fire oscillations of their forefingers on a videogame controller. *—Gary Stix*

King Cotton

W. R. Grace now controls all transgenic cotton in the U.S.

When chemicals giant W. R. Grace & Co. took control in 1990 of Agracetus, a small, money-losing agricultural biotechnology company just west of Madison, Wis., it might have seemed foolish to lay out cash for a firm that, despite nine years of research, still had no products. But two years later the purchase paid off. On October 27, 1992, the U.S. Patent and Trademark Office awarded Agracetus the rights to all genetically engineered cotton produced in the U.S. by any means, of any variety, in any form, until April 2, 2008. Thanks to its timely acquisition, Grace is now entitled to a royalty on every American cotton seed or plant that carries a foreign gene, regardless of how the gene was put there.

The patent's broad scope, unparalleled in plant biology, effectively gives Grace control over much of the research and development on the nation's fourth-largest crop. After nearly a year during which the patent went mostly unnoticed, news of its breadth has spread and begun to worry some researchers and lawyers, who believe it may have set a dangerous precedent.

"It certainly seems to me to be much too broad a claim," says John H. Barton, a professor of patent law at Stanford University School of Law. "It may be a hindrance to research." Jerry Quisenberry, director of the U.S. Department of Agriculture's Cropping Systems Research Laboratory in Lubbock, Tex., shares that fear. "We always had freedom in the public sector to develop a better variety and release it to a seed company for production," he explains. "All of a sudden that's restricted."

NEIL MacDONALD

U.S. PATENT ON ALL TRANSGENIC COTTON was awarded to Paul F. Umbeck of Agracetus, a W. R. Grace subsidiary. Umbeck's broad patent effectively allows Grace to control genetic research on the nation's fourth-largest crop.

Quisenberry also complains that Agracetus's reward is disproportionate to the innovation it actually published. Agracetus perfected the technique of using a microorganism, *Agrobacterium tumefaciens,* to carry a foreign gene into cotton cells and insert it into the genome in such a way that the transformed cells can be identified and cultured into full-grown plants.

But, Quisenberry asserts, "USDA researchers already knew that *Agrobacterium* was going to infect cotton, we knew how to load the bacterium with particular genes and we knew how to regenerate cotton plants" from just a few cells. It was just a matter of putting it altogether, he grumbles.

"That's patent law," responds Kenneth A. Barton, vice president for research and development at Agracetus. Their experiment, he maintains, "was the first time that anyone ever showed convincingly that you could make transgenic cotton. That's why those claims were issued by the patent office."

Grace has promised that it will do nothing to impede public research on transgenic cotton. "We will provide free licenses to any academic or government researcher," Barton pledges. That generous licensing policy makes good business sense. "If Agracetus tried to circle the wagons and not grant licenses, then it would be ripe for attack," says John L. Callahan, senior vice president for cotton at Calgene, a biotech firm in Davis, Calif. Indeed, John Barton of Stanford doubts that Agracetus's patent would stand up in court against cotton transformed using techniques other than the *Agrobacterium* method.

Agracetus has so far managed to avoid legal conflict over its patent. "Nobody is going to take them to court so long as they are reasonable about licensing it," Liz Lassen says. She evidently believes they are: as chief patent counsel for Calgene, Lassen made the decision to license Agracetus's patent so that Calgene can introduce a transgenic cotton variety of its own construction later this year. Monsanto has also purchased a license, and both companies will pay royalties to Agracetus on every transgenic cottonseed they sell.

Currently that is exactly zero. But with those two licenses, Agracetus already has a strong foothold in the future: of 50 applications for field tests of genetically engineered cotton submitted to the USDA from 1988 to 1993, 42 came from Monsanto and Calgene, according to Hope Shand of Rural Advancement Foundation International. Those corporations also happen to own all or part of the two biggest cottonseed companies in the world, giving them control over 61 percent of the U.S. cottonseed market.

Monsanto and Calgene are anxious to replace much of the natural seed sold in that market with more expensive transgenic seed that offers built-in herbicide, insect or fungus resistance. Calgene's BXN cotton, which the company plans to plant on 4,000 acres this spring, is resistant to bromoxynil, a weed killer produced by Rhône-Poulenc. Calgene and Monsanto are both testing cotton that has been given a gene for a toxin found in *Bacillus thuringiensis,* a bacterium that kills pests like the bollworm while sparing helpful insects.

Agracetus's scientists are more interested in tinkering with the genes that code for the cotton fiber itself. "By increasing the strength of the individual fibers, we can increase spinning speeds, and that improves production economics," vice president Barton explains. The company also hopes to develop cotton that holds dye more easily.

Here Agracetus's benevolence runs out. "For the time being, we've decided not to license the patent for fiber modifications in order to give us a competitive advantage," Barton says, although he adds that "anything is negotiable." That decision feeds the fear of those who worry that more monopolistic companies might gain similarly broad patents on other important crops.

Whether the patent office intends to approve such claims is unclear. Some patents filed during the mid-1980s and still under consideration "applied for broad claims similar to those we received on cotton," Kenneth Barton reports. But recently granted patents suggest that the strength of Agracetus's precedent should not be overestimated. Calgene, according to Lassen, set its sights as high as Agracetus did when it applied for claims on all transgenic *Brassica,* a genus that includes rape, broccoli, cauliflower, cabbage and brussels sprouts. But the patent office denied the broadest claims and awarded the company rights only to *Brassica* cells transformed using Calgene's method.

Of course, patents, even broad ones, tend to become obsolete before they expire. Someone inventing a way to control the precise location at which a foreign gene is inserted into the cotton genome would have "a very good chance," predicts John Barton, of getting a patent that bumps Grace's off the map. Callahan of Calgene evidently has thought along the same lines. "Agracetus thinks its patent covers everything," he muses. "That premise may have to be tested." —*W. Wayt Gibbs*

Biocatalysts Turn Rings around the Competition

When the going gets tough, the tough turn to biology. That is what a number of chemists are doing in an effort to solve some extremely difficult problems in chemical synthesis. In at least a few laboratories, the disease-fighting proteins called antibodies are starting to moonlight as catalysts, and they are sometimes succeeding where more conventional agents have failed. In particular, researchers in California have recently shown that catalytic antibodies have a knack for making the cyclic molecules that are the basis of many plastics, synthetic rubbers, insecticides and other materials.

Industrial chemists often need to synthesize organic compounds that contain rings of carbon. In 1928 Otto P. H. Diels and Kurt Alder found that a pair of reactants, known as a diene and a dienophile, would spontaneously combine into rings; that discovery brought them a Nobel Prize in 1950. But from a practical standpoint, the Diels-Alder reaction has some shortcomings.

One is that it can proceed along either an *endo* pathway or an *exo* pathway. Each pathway produces a stereoisomer, or mirror image, of the molecule made by the other, and those stereoisomers have different chemical properties. Thermodynamic constraints favor the *endo* pathway because in the *exo* pathway the reactants must assume a more energetic transition state. Making *exo* products by the Diels-Alder reaction is therefore inefficient at best and nearly impossible at worst.

Enzymes and other catalysts can sometimes help chemists sidestep such problems, but nature has not provided any catalysts for the Diels-Alder reaction. "The *exo* reaction has been rather intractable," explains Richard A. Lerner of the Scripps Research Institute. So he and his colleagues at Scripps and at the University of California at Los Angeles harnessed antibodies for the job.

As catalysts, antibodies have several advantages. If they bind to a transition state, they lend it energy (about 20 kilocalories per mole), which is often more than enough to overcome thermodynamic obstacles. "Antibodies can also control multiple features of the reaction simultaneously," Lerner explains. "So they can catalyze it, direct it down a disfavored pathway and control the stereochemistry all at once." Moreover, the fantastic versatility of the mammalian immune system allows it to manufacture antibodies against virtually any molecule. In effect, biochemists can ask the immune system to invent antibodies with the desired catalytic features simply by injecting mice with molecules that resemble a particular transition state.

Last October in *Science,* the researchers announced that they had created a pair of antibodies that could steer the Diels-Alder reaction along either pathway. In theory, synthetic chemists could use one antibody to make the energetically disfavored *exo* products and the other to make cleaner batches of the *endo* products. Lerner believes the antibody approach should be applicable to many other types of reactions that are not normally favored to occur. "A lot of chemistry never happens but for the want of a few kilocalories," he quips.

In the past, industry has not raced to adopt catalytic antibodies (including some previous antibodies for the *endo* Diels-Alder reaction) because of uncertainty about whether they could produce materials in bulk. That may be about to change. According to Lerner, a forthcoming paper will soon show that antibodies can catalyze the formation of products in gram quantities—not much by industrial standards but a good start, Lerner thinks.

The antibodies that control the *exo* and *endo* Diels-Alder pathways are also just a start. The work is still at "the show-and-tell stage," Lerner notes. His group is now working on antibodies that will afford more control over the reaction so that it yields a single type of molecule rather than just a set of *exo* or *endo* isomers. The team is also trying to make catalytic antibodies for a difficult reaction called a cationic cyclization, which Lerner hopes will be helpful for making steroid drugs inexpensively. —*John Rennie*

Prosthetic Vision

Workers resume the quest for a seeing-eye device

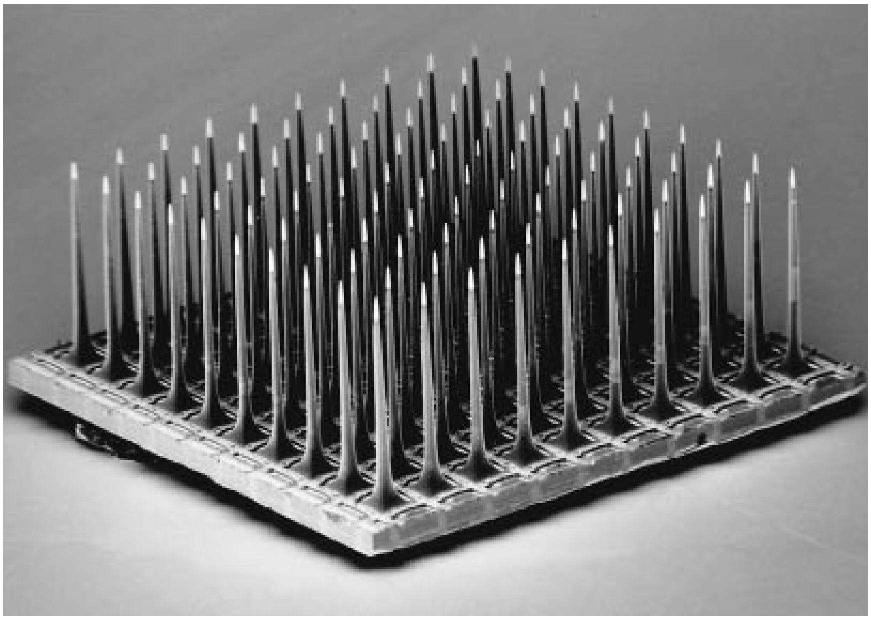
The cochlear implant provides hearing for deaf people by directly stimulating the auditory nerve. Scientists feel a similar prosthesis may one day enable blind individuals to see. Teams of researchers across the country are investigating the prospect of developing implants that could deliver electrical signals to neurons at various points along the visual pathway or even to the visual cortex itself. Assisted by such instruments, enthusiasts hope, a blind person could acquire limited, though useful, vision.

The idea has a long, controversial history. In the 1960s Giles S. Brindley of the University of Cambridge conducted a series of highly publicized experiments. Brindley placed 80 electrodes on the surface of a sightless volunteer's brain. He wired the electrodes to 80 miniature radio receivers and sewed the entire apparatus under the patient's scalp. When he transmitted signals to the device, the subject reported perceiving points of light, known as phosphenes. Similar trials were conducted elsewhere until mounting skepticism concerning safety and usefulness ended them.

After a lull of more than 20 years, the experiments have resumed. "Times have definitely changed," says David J. Edell, who works with Joseph F. Rizzo on an artificial-vision project at the Massachusetts Institute of Technology and Harvard University. "People are a little more cautious about stepping off into the unknown with someone else's body than Brindley was." Edell reasons that scientists have a much better understanding of the nervous system now, an awareness that instills respect for how difficult the task of creating a visual prosthesis is. "The idea is conceptually very simple," Edell notes. "But to do it in a way that's compatible with the biological system is inordinately difficult."

Nevertheless, careful research has led to some real, if modest, success. In 1992 a blind volunteer, who had electrodes implanted in her visual cortex by surgeons at the National Institute of Neurological Disorders and Stroke, recognized phosphene letters. Furthermore, the NIH team observed that mere microamps of current could evoke phosphenes. Artificial-vision researchers give credit for their achievements to new materials, developed primarily for consumer and military electronics. Modern electrodes, such as those made from silicon, are less than one hundredth the size of those used by Brindley, the workers say. Because they are so small, they can penetrate the visual cortex and excite a highly localized population of neurons.

Unfortunately, our visual world does not consist of discrete points of light. "It has yet to be shown that patterned stimulation via a large number of electrodes will evoke a complex perception rather than just a diffuse blob of light," says Richard A. Normann, who heads the visual prosthetics project at the University of Utah. To answer that question, Normann's group has built a three-dimensional array of 100 electrodes intended to stimulate the visual cortex. To date, only subunits of 10 electrodes have been tested at a time.



RICHARD A. NORMANN *University of Utah*

SILICON ELECTRODE ARRAY, built at the University of Utah, may one day be used to stimulate neurons in the visual cortex. The three-dimensional array contains 100 electrodes, each 1.5 millimeters long and 0.08 millimeter wide at the base.

If patterned stimulation does indeed generate adequate images, then what could the blind expect to see? Kenneth W. Horch of the Utah team has shown that a small number of electrodes might produce a diminished, but valuable, visual sense. In simulation experiments, Horch blocked volunteer students' vision using perforated masks. The participants saw 625 points distributed over a 1.7-degree angle, a span roughly the size of your thumbnail at arm's length. Under these conditions, the students were able to read text from a computer screen at two thirds the rate they normally could.

Teams at M.I.T. and Harvard and at Duke University are designing artificial-vision systems to stimulate the retina. Such an aid could capitalize on the natural image-forming processes of the eye. Still, the resolution delivered from a retinal implant would be far from ideal, and this intervention would work only for those patients who had not sustained severe damage to the retina or to the optic nerve. All the same, some patients could avoid the risk of brain surgery.

The designs for retinal implants, like their cortical cousins, are challenged by a host of problems, among them the issue of how to attach an instrument to the retina itself. "Anything that's placed against the eye will tend to flail around when the eye moves," Normann notes. Because the eye moves very quickly, an object of any real mass could cause serious damage. "The retina is like wet tissue paper," Edell explains. "The slightest tear can cause a retinal detachment." To avoid such an outcome, the M.I.T.-Harvard team hopes to devise a silicon implant no more than 20 microns thick. The researchers would communicate with the device via a laser diode mounted on an eyeglass frame attached to a video camera.

"I'm not willing to say we're home-free with artificial vision at all," Normann says. "But the technologies we have developed recently will begin to allow us to ask whether it could become practical or not." Normann suggests that clinical testing of cortical implants may be possible within the next five years and that commercial artificial vision systems could be a reality in the 21st century. —*Kristin Leutwyler*

THE ANALYTICAL ECONOMIST

Is Bigger Still Better?

Buried somewhere in every basic economics text lies a sleepy section on economies of scale. The titans of economic theory—from Marx to Smith to Schumpeter—gravitated toward one conclusion: capitalism inexorably leads to mass production by big enterprises. More than a century of analysis has served only to contrast Marx's maxim "The larger capitals beat the smaller" to textbook mogul Paul Samuelson's adage "Large size breeds success."

Discussions of economies of scale often focus on the role of technology and how it makes vast production runs possible. On a graph the cost per widget drops as the quantity produced rises. Economists, in fact, borrow an example of scale economies from engineers who design gas pipelines. The cost of a pipeline is roughly proportional to its circumference, which determines the amounts of steel needed to build it, but its gas-carrying capacity is equivalent to its area. Wider pipelines have disproportionately increased capacity and thus lower unit costs.

The invention of the microchip suggests to some modern pundits the possibility that the classic relationship can be inverted. As the circuitry shrinks, smaller computers can handle the same amount of work that had been assigned to a hulking cybernetic ancestor. Unlike the gas pipeline, the efficiencies come from dramatic reductions in the size of the means of production. "Technologies are less lumpy than they used to be," says Harvard University economist Dale Jorgenson. "The economies associated with a mainframe are mimicked by a local-area network [LAN]. But the LAN is less indivisible. You can add or subtract units."

Such a concept becomes particularly potent when markets demand increasing customization. The management and control efficiencies that computers make possible mean that capitalists can deploy many assembly lines, each turning out a different variant of the same good, instead of one huge line stamping out identical products. Henry Ford has been preempted: you can now get a widget of any color you want, including black.

In burgeoning service industries, meanwhile, the new technology minimizes "transaction costs." A computer network, for example, makes it possible to enter a piece of information such as an airplane reservation just once. The goal is not to pump an ever increasing amount of bits through an electronic pipeline but rather to make the most use of each bit.

A more speculative notion is whether the smaller means of production alter the scale of the organizations that use them. In other words, do companies scale down along with the size of the wire widths in their computers? If they do, then relatively small firms might sometimes compete with larger ones on equal terms (or at least until the more substantial rival responds by instituting its own efficiencies).

Is production of fewer numbers of more specialized products better carried out by a small company? Erik Brynjolfsson and Thomas W. Malone of the Massachusetts Institute of Technology have found evidence that suggests smaller companies may be better at producing short runs of specialized products than are their oversized cousins. The two led a study that correlated a 20 percent decline in the number of employees in industrial firms with a tripling in the investment in information technology over the decade of the 1980s.

Economists ponder the notion that a firm's size may shrink along with the circuits in its computers.

This trend, Brynjolfsson says, cannot for the most part be explained by the prevailing idea that machines replace people. Instead the investment in computers and networks often makes it cheaper to farm out work to low-cost specialists rather than to retain the rigid organizational structure needed to manage the entire process under one roof. Small automotive-parts companies grow, and General Motors lays off workers. "There is a silent majority of very small enterprises hiring as fast as big firms are laying off," Brynjolfsson says.

Other economists look at the statistics and draw a nearly opposite conclusion. "The small-firm myth is not true. It's a myth that's partly ideology and partly wishful thinking," says Bennett Harrison, a professor of political economy at Carnegie Mellon University. "After adjusting for ups and downs of the business cycle, the share of jobs attributed to small and medium-size firms has stayed almost exactly flat since 1960."

True, large companies have slimmed down, but they still account for a disproportionate share of economic activity, Harrison argues. The notion that diminishing costs lower the break-even point for investing in a new technology is often valid at the level of the individual computer-controlled machine but not at the scale of a business. A big company will have the resources and capital to invest in using the equipment most effectively. And as for flexibility, "the technology can produce customized small lots, but that can be done in the corner of a Ford Motor Company plant as easily as in a small business," he contends.

Attempts to settle the argument in favor of either big or small firms may ultimately be futile. Instead of trying to determine an optimal firm size in a vacuum, some academics are looking at the relationships among small firms, large firms and the political and economic environment around them.

These studies are often as much the discipline of the geographer as they are of the economist. At the University of California at Los Angeles's Lewis Center for Regional Policy Studies, researchers are studying the means by which industrial districts such as Silicon Valley, the southern California aerospace industry or the agglomeration of textile firms near Florence (known as the Third Italy) achieve "external" economies of scale. Once a regional industry has begun to grow, it attracts a skilled labor force and a supporting infrastructure of city streets, expressways and airports. Small specialty firms then may coexist with giant production houses. The scale factors can involve hundreds of thousands of workers spread over a few square miles rather than single plants or pipelines.

The simple cost-curve portrayal of traditional economies of scale may gradually fade from future texts. Perhaps the only certainty in the big versus small debate is that to capture the subtleties of these arguments the textbooks themselves are bound to grow in size. —*Gary Stix and Paul Wallich*

MATHEMATICAL RECREATIONS

by Ian Stewart

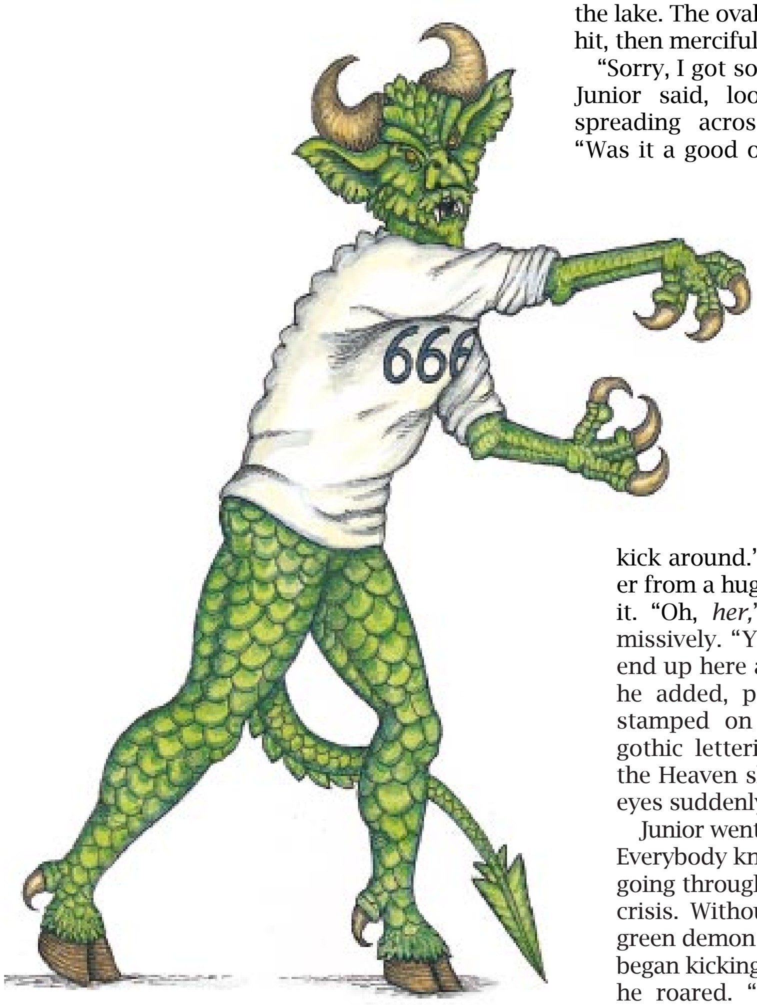
The New Merology of Beastly Numbers

Let him that hath understanding count the number of the beast: for it is the number of a man; and his number is Six hundred threescore and six.

—Revelation of St. John the Divine 13:18

The flames of hellfire flickered amid towering pillars of smoke, scorching the leaden sky. In the distance the iron walls of Dis glowed a fiery red. A green demon in a sweatshirt was lounging beside a lake of boiling oil, idly tossing a brown oval object from claw to claw. The shirt bore the number 666 on the front, and when he swiveled around for a practice throw the name "THE BEAST" became visible on the back.

A second demon, which was a rather attractive shade of pastel blue, tapped him on the shoulder. "Uh—excuse me, Mr. Beast, sir."



"Who are you?" The Beast turned around and looked at the blue demon's sweatshirt. "What pansy kind of a number is $-847\ ^1/_2$?"

"It's what they gave me when I took the job you advertised, sir. It was the only shirt that fit me."

"Job? What—oh, I remember. Turn around." The back of the blue demon's shirt bore the words "JUNIOR SOULS ASSISTANT." "Okay, Junior, let's see how good you are. Go out for a run, and I'll throw you a pass." The Junior Souls Assistant set off at a gallop along the shoreline. The Beast drew back a bony arm and fired one of the oval objects at Junior's head. Junior made a grab for it, missed and watched it bounce past into the lake. The oval object screamed as it hit, then mercifully sank from view.

"Sorry, I got some sulfur in my eye," Junior said, looking at the ripples spreading across the burning lake. "Was it a good one?" Then he caught himself. "Silly question. We don't get the good ones down here."

"It wasn't even a bad one," the Beast remarked. "Only a cheap rubber soul, not even leather. But you don't need to worry about losing it—there are always plenty of condemned souls to kick around." He picked up another from a huge pile and scrutinized it. "Oh, *her,*" said the Beast dismissively. "Yes, she was bound to end up here all right. But at least," he added, pointing to the name stamped on the soul in uneven gothic lettering, "she knows who the Heaven she is. Unlike me." His eyes suddenly welled up.

Junior went a paler shade of blue. Everybody knew that the Beast was going through a mid-death identity crisis. Without warning, the huge green demon leaped to his feet and began kicking things. "WHO AM I?" he roared. "Thousands of years, and nobody ever tells me who I am!" Then he burst into tears.

"I heard that in Aramaic—the original language of the Book of Revelation—the symbols for 666 spell 'Nero.' Do you think you might be Nero?"

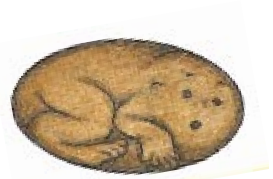
" 'Nero fiddled while Rome burned....' I don't mind the burning bit, but I'm not a very good fiddler, you know."

"Well, the Jesuit Father Bongus decoded your number as 'Martin Luther,' " the Junior Souls Assistant ventured, patting the Beast sympathetically between his horns. "He used the system known as gematria, in which A = 1, B = 2,... up to Z = 26."

"I know all about that. But, on the other claw, Michael Stifel—a German mathematician—'proved' I was Pope Leo X," the Beast sniveled. "He started with 'Leo Decimus' and threw away everything except LDCIMV."

"Why?"

"Those are the letters that correspond to Roman numerals. They add up to 1,656, so he added another X because it was Leo X and deducted M because it was the initial letter of 'mystery.' " The Beast grimaced. "Did you ever hear of such a dumb argument?"



"Never," Junior said. "Obviously contrived to get the desired result."

"Precisely. Everybody's got an ax to grind. And what makes it worse," the Beast went on, "is that they all have different systems for assigning numerical values to letters. Nobody asks whether there's some sensible method that doesn't depend on alphabetical order or other arbitrary choices."

"Funny you should mention that. I've just been reading the February 1990 issue of *Word Ways.*"

"What the angel is that?"

"Well, it's a journal of recreational linguistics."

"We're in trouble then," the Beast said. "This is a mathematical recreations column."

"Ah, but that particular issue had an article that combined both mathematical and linguistic recreations. It was by Lee Sallows, an expert in numerical wordplay, and it was called 'the new merology.' He noted that 'the time-hon-

Perfect Number-Words in English

Assign the values

$E = 3$	$I = -4$	$R = -6$	$V = -3$
$F = 9$	$L = 0$	$S = -1$	$W = 7$
$G = 6$	$N = 5$	$T = 2$	$X = 11$
$H = 1$	$O = -7$	$U = 8$	$Z = 10$

Then

$Z + E + R + O = 0$

$O + N + E = 1$

$T + W + O = 2$

$T + H + R + E + E = 3$

$F + O + U + R = 4$

$F + I + V + E = 5$

$S + I + X = 6$

$S + E + V + E + N = 7$

$E + I + G + H + T = 8$

$N + I + N + E = 9$

$T + E + N = 10$

$E + L + E + V + E + N = 11$

$T + W + E + L + V + E = 12$

oured practice of linking each letter to its position number is an expendable—because profitless—convention. New merology takes this as its starting point.' "

"I don't follow you," the Beast interrupted testily.

"Well, take, for instance, the English word 'one.' With the usual alphabetical ordering in that language, its numerical value—or gematric constant—is $15 + 14 + 5 = 34$. But if numerology has any intrinsic meaning, then the gematric constant clearly ought to be the number that the word denotes, namely, 1. But it's not."

"Do any number-words have totals that equal their gematric constant?"

"According to Dave Morice, no—not in English, at least. Unless you count phrases like 'This is a Beastly text: numerological constant of six-six-six.' "

"So what does this Sallows person propose to do about it?"

"Assign numbers for each letter that aren't given by the position in the alphabet. Say a number-word is perfect if, using those assignments, its gematric constant equals its numerical value. Then try to make as many as possible of the number-words ONE, TWO and so on be perfect. He restricts attention to whole number values and imposes the rule that different letters must be given different values."

The Beast blew his nose loudly. "To what effect?" he sniffed.

"Well, you get a whole pile of equations like

$$O + N + E = 1$$
$$T + W + O = 2$$
$$T + H + R + E + E = 3$$

in algebraic unknowns O, N, E, T, W, H, R…. You see how many of those equations can be solved, and then you adjust your answers so that they satisfy any other criteria you want to impose. For instance, you can see from the equation $O + N + E = 1$ that some of the numbers have to be negative. Then, because N and E both occur in NINE and TEN, it makes sense to assume that N and E have been assigned values and see what happens. So O, for instance, must satisfy $O = 1 - N - E$. From NINE and TEN, you get $I = 9 - 2N - E$, $T = 10 - N - E$. Next, because we want $T + W + O = 2$, we find that $W = 2 - O - T = 2 - (1 - N - E) - (10 - N - E) = -9 + 2N + 2E$.

"You have to watch out, though. Suppose you decide to start with $E = 4$, $N = 2$. Then T works out as 4, too—the same as E. Because that's not allowed, you have to choose different values for E and N."

The Beast perked up. "I get it! Suppose you try $E = 1$, $N = 2$. Then you have to have $I = 4$, $T = 7$, $O = -2$ and $W = -3$. What if we want THREE to be perfect, too? Well, that's two new letters, H and R. If I guess that $H = 3$ and use the fact that $T + H + R + E + E = 3$, then R has to be -9. And then FOUR gives two more new letters, F and U. If I choose $F = 5$, then $F + O + U + R = 4$ leads to $U = 10$. We're making good progress, Junior."

"Right. Then $F + I + V + E = 5$ leads to $V = -5$. Because SIX contains two new letters, let's try SEVEN first and fix the value of S. If $S + E + V + E + N = 7$, then $S = 8$. And then we can fill in X from $S + I + X = 6$, getting $X = -6$. Then looking at the equation for EIGHT, we get $G = -7$."

"So now all the number-names from ONE through TEN are perfect. Can we go further?" the Beast asked.

"Maybe. The only extra letter in ELEVEN and TWELVE is L. It would be quite a coincidence if it could be chosen to make them both perfect."

"Yes, but in fact $L = 11$ does make them both perfect."

"Amazing. Now $T + H + I + R + T + E + E + N = 7 + 3 + 4 + (-9) + 7 + 1 + 1 + 2$, which is 16. Oh, salvation take it!"

"I suppose we might get further by making different choices earlier," Junior observed. "Several of the values were just arbitrary guesses."

"I'm not sure it helps," the Beast countered. "Look, if you take the following equation

$$\text{THREE} + \text{TEN} = \text{THIRTEEN}$$

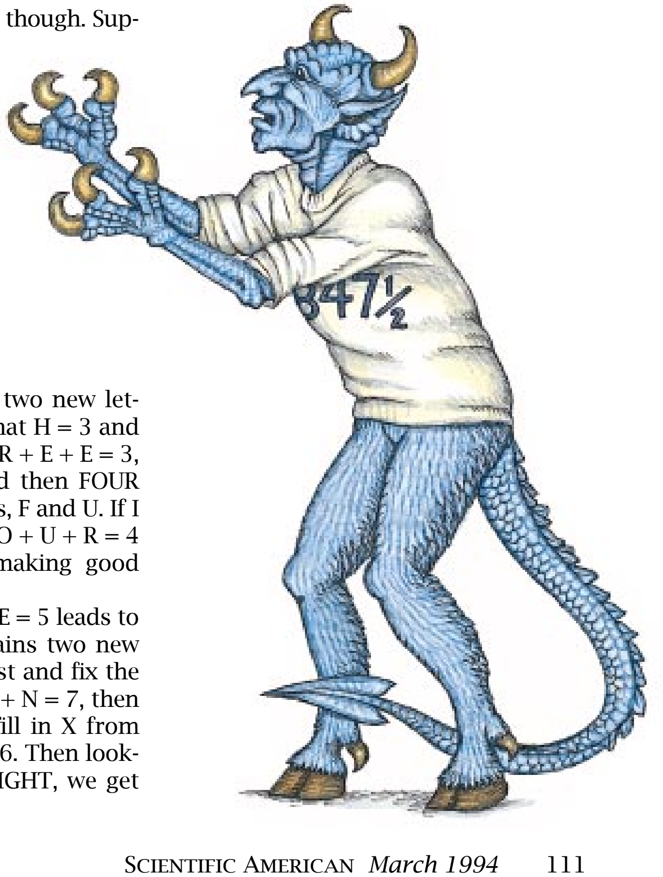
and cancel a letter if it occurs on each side, then you end up with $E = I$. And that violates the rule that different letters get different values."

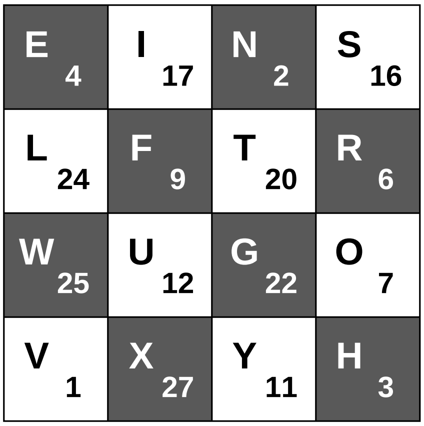
The Junior Souls Assistant nodded. "I remember now, Sallows discovered that argument. He says it's a new merological proof that 13 is unlucky."

"We could still go backward. Let's see: if $Z + E + R + O = 0$, then $Z = 10$."

"That's the same as U. Bother."

"Yes, but we made a lot of assumptions about the values of letters. Maybe





▨ E 4	I 17	▨ N 2	S 16
L 24	▨ F 9	T 20	▨ R 6
▨ W 25	U 12	▨ G 22	O 7
V 1	▨ X 27	Y 11	▨ H 3

NUMERICAL SPELL of Lee Sallows. Select any number on the board. Spell it out, letter by letter. Add together the corresponding numbers (subtracting those on black squares, adding those on white squares). The result will always be plus or minus the number you choose.

we can fix things up by changing them." And they discovered they could [*see box on preceding page*].

"Sallows describes several magic tricks based on similar ideas," Junior offered. For instance, if you use a different assignment,

E = 0	I = 1	R = 5	V = 14
F = –10	L = –7	S = –11	W = –1
G = 9	N = 4	T = 6	X = 16
H = –8	O = –3	U = 12	Z = –2

then ZERO through TWELVE are perfect, and so are FOURTEEN, SIXTEEN, SEVENTEEN and NINETEEN. The idea is to make a set of cards, each having a letter and the corresponding number written on it. Make three cards with E/0, two with N/4 and one each for the rest, a total of 19 cards. Ask someone in the audience to spell out a number-word. Add up the numbers on the cards, and lo and behold it is the same value. Except for sign—it could be positive or negative."

"And what if some smart aleck spells out THIRTEEN or FIFTEEN?" the Beast inquired.

"Can't be done. There's only one T and only one F in the pack."

"Devilishly clever. That Sallows guy would fit in well down here."

"True. He invented a similar trick that involves a 4 × 4 chessboard [*see illustration above*]. Choose any number on the board. Spell out its name, and add up the numbers associated with its letters, counting them as positive on white squares but negative on black. Again, the result is the number selected."

The Beast laughed, a horrible sound. Joyously, he booted a dozen condemned souls out of sight over the horizon and bounced a 13th up and down to the accompaniment of muted howls whenever it hit the ground. Then the demon's face fell.

"What's the matter, Mr. Beast?"

"It's all very well using English names. *Mais en Français, par exemple?*"

"Ah, yes. That opens up an entirely new range of questions. In French the number-words are UN, DEUX, TROIS, QUATRE, CINQ, SIX, SEPT, HUIT, NEUF, DIX, ONZE, DOUZE, TREIZE, QUATORZE and so on. Now, you can get as far as 13 but not 14. Because canceling common letters on both sides of QUATRE + ONZE = UN + QUATORZE leads to E = U, which is forbidden since different letters must get different numbers. The whole game is remarkably rigid in French. If you tackle the number-words in the right order from ZERO to TREIZE, you will find that 11 letters are all determined by the value of N and that the only other free choice is A. That is, you are forced to have

A = *	P = 2
C = A – 5N – 4	Q = 2N + 5 – A
D = 2N	R = N – 11
E = 3N – 5	S = 2N – 4
F = 13 – 3N	T = 14 – 5N
H = 4N – 11	U = 1 – N
I = 2N + 4	X = 6 – 4N
N = *	Z = 16 – 4N.
O = 0	

You get different (and small) values for all the letters if you take A = 20, N = 7. You end up with the assignments

A = 20	H = 17	Q = –1	X = –22
C = –19	I = 18	R = –4	Z = –12.
D = 14	N = 7	S = 10	
E = 16	O = 0	T = –21	
F = –8	P = 2	U = –6	

Every number-name from ZERO to TREIZE is then perfect."

"*Und auf Deutsch?*"

"Oops. Sallows didn't mention German in his article. I guess we'd better work that out for ourselves."

"Me first. Omitting zero—because I have a hunch it will be better that way—the number-words go EINS, ZWEI, DREI, VIER, FÜNF, SECHS, SIEBEN, ACHT, NEUN, ZEHN, ELF, ZWOLF, DREIZEHN, VIERZEHN, FÜNFZEHN, SECHSZEHN, SIEBZEHN, ACHTZEHN, NEUNZEHN, ZWANZIG. I suppose we'd better consider Ü to be the same as ordinary U."

"If you say so, Mr. Beast. Hey, I think we're on to something here. Look, because of the way German deals with the teens, if you get the number-names up to ZEHN working, then DREIZEHN up to NEUNZEHN are automatic."

"Except from SIEBZEHN, which isn't SIEBENZEHN."

"Right. But that tells us that because SIEBEN + ZEHN = SIEBZEHN, then, canceling, we must have E + N = 0. Then E + I + N + S = 1 dictates that I + S = 1. Suppose we let E and I have any values we like. Then we must have

$$\begin{aligned} E &= * \\ I &= * \\ N &= -E \\ S &= 1 - I. \end{aligned}$$

Continuing with ZWEI, we can make Z arbitrary and deduce that

$$\begin{aligned} Z &= * \\ W &= 2 - E - I - Z. \end{aligned}$$

And so on. It gets complicated, but I think if we work through it systematically we can—"

"Yes. And I've just had a thought. German for 'twenty-one' is EINUNDZWANZIG and so on. If we make UND have a total of 0, then we ought to get the numbers from 21 to 29 for free, too."

For several hours, Junior and the Beast scribbled in the sulfur sands of Hades with their pointed tails. Eventually they came up with this:

A = –10	F = –2	N = 1	U = 8
B = 7	G = 33	O = –6	V = –8
C = –18	H = 17	R = 16	W = 13
D = 9	I = –3	S = 4	Z = –7.
E = –1	L = 14	T = 19	

Now all number-names from EINS to NEUNUNDZWANZIG (29) were perfect.

"Good work, Mr. Beast," Junior exclaimed. "I wonder if we can get DREIZIG (30) working as well? If so, we'd get 31–39 for free. And what about Italian, Spanish, Russian, Greek, Japanese, Pidgin?"

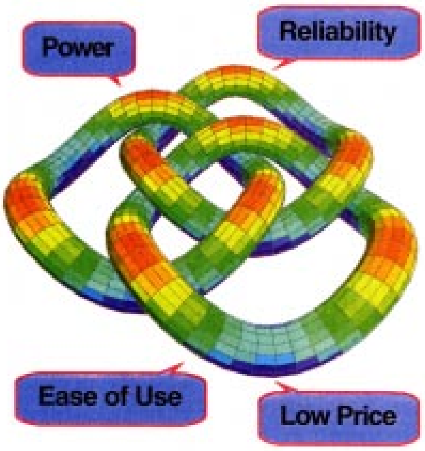
"Best left for the torment of earthly souls, I would judge," the Beast snorted.

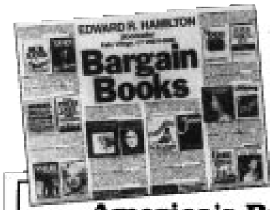
FURTHER READING

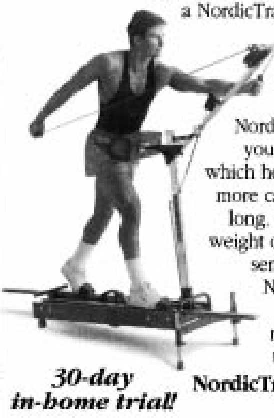
A SHORT HISTORY OF MATHEMATICS. Vera Sanford. Houghton Mifflin, 1930.

PERFECT NUMBER NAMES IN NEO-ALPHABETS. Dave Morice in his regular column *Kickshaws,* in *Word Ways,* Vol. 22, No. 4, page 238; November 1989.

THE NEW MEROLOGY. Lee Sallows in *Word Ways,* Vol. 23, No. 1, pages 12–19; February 1990. [*Word Ways* is available from Faith W. Eckler, Spring Valley Road, Morristown, NJ 07960.]







BOOK REVIEWS by Philip Morrison

What Is Humankind?

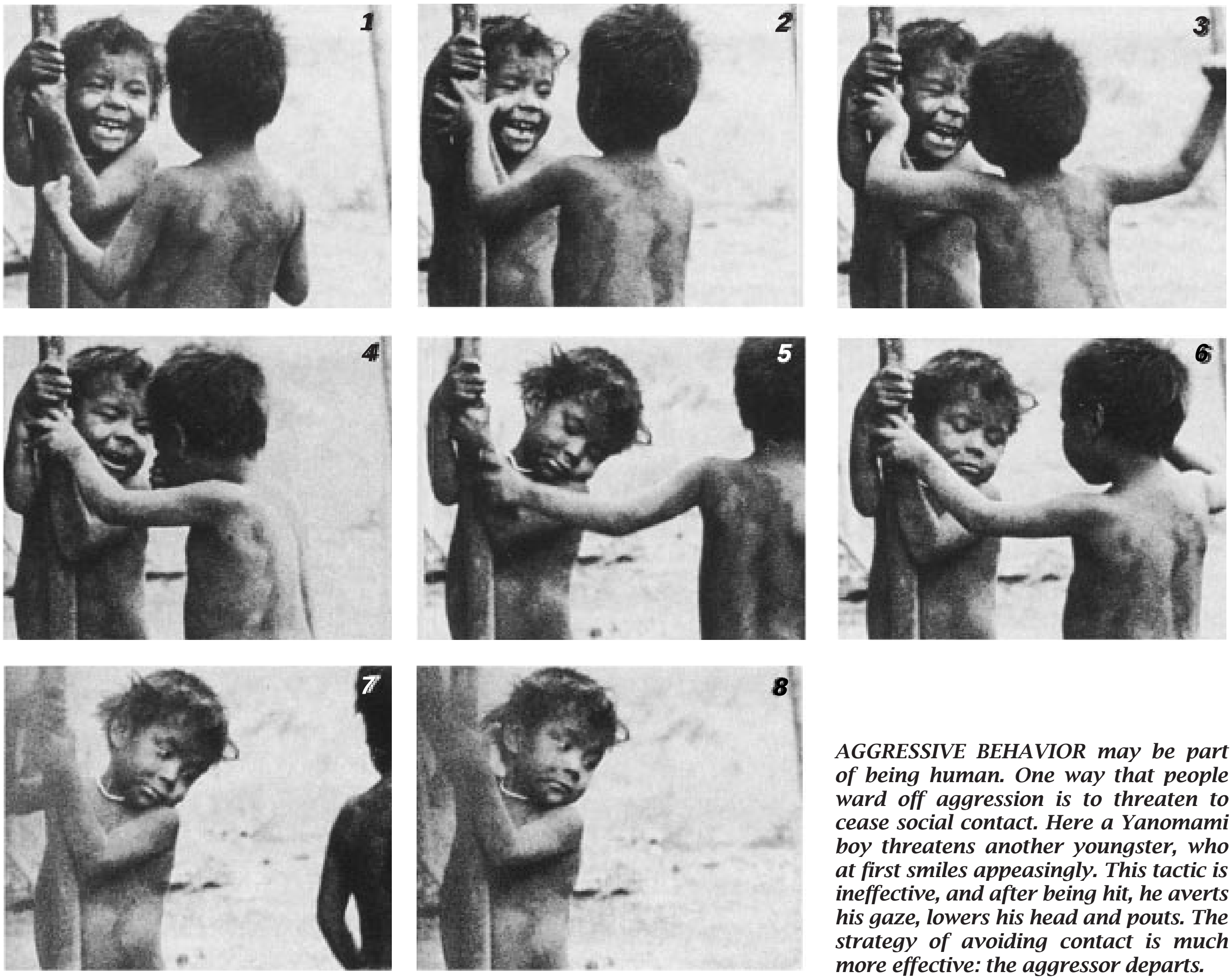
The First Humans: Human Origins and History to 10,000 B.C., by Göran Burenhult, general editor. The Illustrated History of Humankind, Vol. 1. HarperCollins, 1993 ($40).

What we call the Old World extends over nearly a hemisphere, an unbroken tract of land whose far-flung cornerposts lie near the present cities of London, Capetown, Beijing and Surabaya. It is the indisputable relics of our forebears that define the Old World, the ancestral home. Scavengers, foragers and hunters, they spread on foot, probably from the Rift Valley of Africa, for a million years and more, until expansion paused at the barrier of open sea.

The true New World is the other, watery hemisphere, with three seagirt continents. "Australia was *Homo sapiens'* first New World." The best dating suggests the ocean barrier was first broken at the deep, narrow passage eastward from Indonesia, by men and women in outrigger log canoes. Our resourceful species likely first entered the Americas by traveling eastward from Siberian tundra along the shallow Arctic seas.

A fine map lists by name some 70 Old World sites where bones and artifacts of "our fossil forebears" have been found; firm evidence of premodern species dots two dozen regions of Africa, Eurasia and Indonesia. In all the New World, not one such site offers the bones or tools of any hominids but our own modern species. The New World (don't forget Australia!) is a world wholly of sapient humankind, confident on arrival, with fire, language and widely varied styles of high craftsmanship. No site in the New World is anything near a million years old, an order-of-magnitude distinction from many an old site in the other hemisphere. More precise dating remains disputable, but that sharp division by great age and by the resident human species rests on sure ground.

Fresh and open design, large pages with many colorful photographs, diagrams and reconstructions, and an appealing text unafraid of controversy mark this, the first volume of an ambitious five-volume set. The work was conceived of and produced in Australia, although the effort is cosmopolitan, with a general editor in Stockholm and two dozen anthropologist contributors



AGGRESSIVE BEHAVIOR may be part of being human. One way that people ward off aggression is to threaten to cease social contact. Here a Yanomami boy threatens another youngster, who at first smiles appeasingly. This tactic is ineffective, and after being hit, he averts his gaze, lowers his head and pouts. The strategy of avoiding contact is much more effective: the aggressor departs.

from around the world. There are 10 lively chapters by as many authors to carry the topic systematically from human origins on to the archaeology of the circumpolar Arctic. Within these narratives, 40 relevant spreads are embedded, each one a specialist's concise account, a bouquet of insights and surprises that includes Neanderthal religion, Venus figurines, thermoluminescence-dating techniques and tests of old spearpoints in elephant hunts today.

A couple of the 300 eye-and-mind-catching images impel mention. Here are two grand flint points made for the mammoth hunt, beautifully worked in Washington State some 10,000 years ago, each one twice the length of a human hand. Or look at the great auk masterfully drawn 18,000 years back on an upper wall of a half-flooded limestone cave not far from Marseilles. The image is found 500 feet shoreward from the submerged cave entrance 120 feet underwater. That unique cave was found in 1991 by a diver.

While in China, "imagine a huge cavern...existing for hundreds of thousands of years, and consider a few of the things that would take place." Large numbers of hyenas, wolves, cave bears, a few at a time, live and sometimes die within; owls roost and regurgitate their pellets of fur and bones. Rocks break loose and fall from the roof, sand and dust wash in, plants grow from the seeds carried in by scurrying rodents. Finally, people enter, too. Much later the scholarly excavation is no single coherent campaign, but a long struggle of many directors against time and penury, often under stress from war outside. That is the history of the great Chinese cavern Zhoukoudian, studied for two generations.

Many quartz tools and many incomplete skeletons of Peking Man have been found there. But the noise of background events is loud, the human signal faint, even though the fossil bones recovered are "immensely important." About all we can be sure of is that local groups of *Homo erectus* lived there, left some bones and some tools, and made fire. They were much rarer occupants than the hyenas, their firemarks far less visible than thick black layers of owl droppings. This eloquent and realistic passage goes a long way to make clear what we do and do not know about our ancestors; even more, perhaps, it clarifies what we know that ain't so.

For balance and timeliness, for its rich images and compact lively text, this is a most welcome book. Of the four more volumes promised soon are one on the rest of the Stone Age, one on the civilizations of the Old World and one on those of the New (with the Pacific Islands), and a final book in late 1995 on "traditional peoples today." The books to come have here a prepossessing forerunner.

Taming the Gases

HISTORY AND ORIGINS OF CRYOGENICS, edited by Ralph G. Scurlock. Oxford University Press, 1992 ($165).

Recognition of the material nature of impalpable, invisible gases was no small part of the revolution in chemistry in the days of Antoine Lavoisier. A generation later Michael Faraday began to link gases fully to the world of matter by turning them to liquids or solids. All the apparatus artful Faraday needed was a stock of thick-walled glass tubes a foot or two long that he could seal. At one end he placed the reactants he would heat to form the gas; the other end might be ice-cooled. His instrumentation was an internal pressure gauge—a drop of mercury that sealed room air into the closed end of a length of glass capillary tubing—and an external thermometer. At once he made decisive progress. He returned to the task much later with somewhat heavier weapons, even hoping to turn hydrogen into a metal! By 1844 he could report condensation of a dozen or two gases, including a few exotics such as cyanogen or fluosilicon. Pressures rose to 50 or 100 atmospheres; glass explosions were common, and masks and goggles became de rigueur at the Royal Institution.

"Permanent gases" remained a tough challenge. Hydrogen, nitrogen, carbon monoxide and oxygen refused to liquefy, even at the highest pressures. By the 1870s it became plain both from theory and from experiment that gases have a critical temperature. Above that temperature, gas and liquid have equal density and form in fact only a single uniform phase. Once any compressed gas has been cooled enough to turn it liquid, the liquid can be evaporated and thus cooled simply by lowering the pressure. That colder liquid can be used to cool a second gas below its own critical point, and so a cascade of many cooling steps can be organized using different substances. Free expansion can cool some gases by lowering internal energy, and gas expansion to do external work against the mechanical load of piston or turbine wheel cools even better.

Steam-powered mechanical refrigeration, using various compressed and expanded fluids, had begun in mid-century. Its purpose was to replace the winter ice customarily stored in Europe and America for summertime use and shipped as far as India. Refrigeration for brewing lager beer in summer and for overseas transport of frozen meat prospered by the 1870s. By then the thermodynamic lessons had been well learned in the laboratories, too.

Oxygen was first liquefied in 1877. In Geneva, success came from a cascade of heat exchange stages, first with liquid sulfur dioxide, then with liquid carbon dioxide. The goal was independently reached at the same time in Paris by cooling oxygen at high pressure by surrounding it with cold liquid ethylene and suddenly releasing the pressure. Both labs made only a transient liquid, a mist or a small jet of liquid oxygen that neither could catch or store before it vaporized. Yet the first permanent gas had been tamed; cryogenics had been born.

A ubiquitous resource, atmospheric oxygen, supports breath itself; a ton of oxygen is taken up each second by the breaths of the population of the U.S., far less than our fires use. Industry separates purified oxygen from American air at about half the rate we breathe and purifies a ton a second of nitrogen besides. Those once permanent gases are distilled in tall columns that enclose a stack of vertically spaced shallow trays. Liquefied air is fed in halfway up. Nitrogen-rich vapor bubbles upward tray by tray while the heavier, oxygen-rich liquid drips down. For air, however, all temperatures are well below the ambient. The cryogenics here is not at all in the product but in the process.

What is most striking is that this modern industry goes straight back to la belle époque. Professor Carl von Linde in Munich was first to produce liquid air in 1895; inventive young Georges Claude in Paris was only a few years behind. Linde used the expansion of precooled high-pressure air through a simple valve, depending on internal work only. Claude instead made a piston engine to take up gas expansion and needed much lower pressure. One key invention was Claude's use of engine lubricant that would not freeze; in the last expansion stage, that was liquid air itself, an idea, he wrote, "which was by no means a stroke of genius." The first large market was oxygen for welding. Oxyacetylene welding—acetylene was in wide use for lighting at the time—was a European innovation of 1900.

The original firms of Linde and of Claude still dominate world production. L'Air Liquide, the bigger of the two, operates plants in 100 countries. The separation process is conceptually unchanged, though remarkably improved; plant size has multiplied 200-fold.

In 1877 Sir James Dewar was newly professor at the Royal Institution, where Faraday had worked all his life. Within a year Dewar had made a mist of liquefied oxygen in a public lecture using apparatus from Paris. He was not first to liquefy hydrogen, however; a transient hydrogen jet was made in Kraków in 1884. But in 1898 he was the first to hold hydrogen liquid, using his celebrated silvered vacuum flasks, still called dewars. In 1901 Sir James, now expert with hydrogen, set out to liquefy the new, very scarce element helium. It was another race, with eager entrants in London, Leiden and Kraków.

James Dewar had always worked rather alone, even secretively, with a couple of loyal technicians. But H. Kamerlingh Onnes—professor of experimental physics at Leiden—pioneered the big lab of the future. His team included younger academic peers, including engineers and theorists, and drew a steady procession of lively students. The entire building full of eager cryogenic researchers at Leiden could outstrip any lone professor and a couple of assistants at the bench. And it was in Leiden that helium was first liquefied, stored and put into research use.

Traveling the unending road toward absolute zero had thrown up the first great cryogenic surprise, bulk superconductivity. By now we understand it pretty well; after all, in every atom a few electrons share unending quantum motion. The even stranger atomic superflows in isotopic liquid helium are familiar, too, with a wealth of phenomena. The search now is on for ceramics that will superconduct even without cryogenics. Thermodynamics led the cryogenics of the last century; in this one, it was quantum mechanics that finally unlocked the wonders of superfluidity. Cryogenics today imagines its future around hydrogen and helium, around useful supercurrents and new fuels, not in the secure past of liquid air.

Powerful and precise superconducting magnets are now in place under helium at 1,000 medical imaging sites and a couple of accelerators (though not in Texas) and are in late stages of R&D for levitated trains and ship propulsion. Maybe one day they will store magnetic energy in bulk for peak hours of power demand. Meanwhile myriad smaller uses of cryogens are signaled by the dewars we see at every size, from thermos bottles to tank cars.

The cryoengineers have not lagged. There are many tankers afloat to carry liquid methane across the seas beyond the span of pipelines. Experts dream of liquid hydrogen to replace the hydrocarbons as the future fuel of a clean and sustainable transportation economy whose primary energy is solar or nuclear. (The learned editor regards liquid hydrogen as too dangerous for automotive fuel.) Safe professional handling of liquid hydrogen reached large scale, mainly for rocketry, although cryogenic rocket fuel is no longer so dominant over solid propellants. Indeed, liquid hydrogen as motor fuel might also give way to hydrogen stored as metal hydrides.

This expensive volume was prepared for insiders. But it is full of good general reading, although it is clearly a collection of opportunity by its many expressive authors. They offer variety from complete national histories to brief narratives of a single important firm or laboratory, in 18 chapters that tell a story covering Europe east to Poland, the U.S., China and Japan.

The Joy of Slime Molds

LIFE CYCLES: REFLECTIONS OF AN EVOLUTIONARY BIOLOGIST, by John Tyler Bonner. Princeton University Press, 1993 ($19.95).

I have devoted my life to slime molds." So opens this funny, self-deprecating, charmingly personal book. It lies between an informal memoir of a life of teaching and fruitful research and an unconventional portrait of biology, a broad-brushed picture painted on the easel of evolution.

Bonner found his way to slime molds as a graduate student. Another student's thesis led the way. He had pulled it idly off the shelf in his supervisor's office while "talking to the pretty secretary." The cellular slime molds described from topsoil worldwide were a young embryologist's dream. Begin with new spores. Given the needed substrate, these quicken into scattered amoebalike cells, scarfing away at the rich bacterial lawn they enjoy. After a few days, the cells have divided mightily, and no food is left; the hungry cells stream together into groups. Each grouping becomes a translucent millimeter-size sausage of many cells, able to mimic a tiny sluglike organism, to ooze slimily about with distinct front and hind ends. These slugs wander a while, until one by one they settle in place and point upward. Their cells flow and change about until the slugs have turned into implanted fruiting bodies with elegantly tapered stalks a millimeter or two high. The ripened stalks come to bear a small sphere, out of which are released new spores da capo. The cycle of life has turned once; it is the organism over time that is this book's grand theme.

Every chapter has as title and topic a way not of being but of becoming. We read of three ways of becoming larger: by multicellularity (remarkably diverse in itself; witness the slime molds); by development (the differentiation of cells); and by evolution, now not within one life cycle but over linked myriads of them. Adult forms change more subtly. Having become increasingly aware of the dynamic environment, they can enlarge in effect by lifelong alliances. Eventually they become social, like the termites and ourselves, until they come to bind past with future much more swiftly, growing cultural as well, which may entail intelligence.

But back to slime molds. These part-time multicellulars are aware of one another. Single cells repel. Cut a slug into three and plant the parts nearby. They make stalks that repel again—the stalk of the slug front end forward, the hind end back, the middle piece vertical. Undergraduates tried shuffling the ends, but the result was the same. All fruiting bodies repel, whatever their place of origin. By what means? After nice hints that some gas was involved, a crucial test was done. A fruiting body nosedives right into any nearby fragment of activated charcoal. That stuff absorbs and binds most volatiles, so that its neighborhood is free of vapors. The clincher was an arrangement that held a fruiting body sealed in a lidded plastic dish, one with a tiny hole in the lid. Put the dish in a chamber that held clean, moist air, and the fruiting body would aim straight up for the hole; sometimes it even went through it. The case was the same for each of the vapors slime cells release, save one. Ammonia outside, even at a thousandth of the concentration of the other test gases, induces the fruiting body to flee the hole. "A moment of slime mold eureka!"

If you want to read about Albert Einstein and slime molds, or about national traditions for preparing culture media, or any of the deep generalities touched on, you will find a lot in this small and friendly space. Nor is sweet, cheerful John Bonner always in one mode. He has outgrown idealized *Arrowsmith* as he welcomed the candor of Jim Watson, never accused of sweetness. He is stubborn, with good reason, about evolutionary grand design: forms get bigger if they can. And try any other of his eight or nine books; this reviewer is going to look up *The Cellular Slime Molds,* second edition, Princeton, 1967.

ESSAY by Susan Zolla-Pazner

The Professor, the University and Industry

It used to be easy to be a professor. You would read your professional journals, write your scientific papers, teach and give seminars. But universities are in transition and so, therefore, are their faculties. More and more emphasis is being placed on research rather than on teaching. With the constriction of federal research funding and the influx of support from private industry, some see a transformation from university to research institute to industrial subsidiary. So when today's professors hit the big time, they have to read their professional literature *and Business Week,* write scientific papers *and* patent applications, teach, give seminars *and* sit on the scientific advisory boards of various corporations.

This interaction among scientists, universities and industry is not new. But the decrease in government support for research, combined with the explosion of new biotechnology products, has intensified the relationship. It is now more productive—and more complicated.

At present, both the university administration and private industry must play a role in developing the scientific knowledge that germinates and grows in the academic environment. The university is usually responsible for obtaining patents and for licensing the rights for its professors' inventions. The biotech company, having licensed the product, must provide the considerable financial backing required for its development and marketing.

In the best of all possible worlds, the inventors, the university administrators and the biotech executives work as a well-oiled machine that creates a beneficial product and generates capital to support the academic lab, the scientist, the university and the company's shareholders. In the real world, however, each of these component parts has its own agenda. The goals may not entirely overlap; the priorities may be misaligned.

Nevertheless, advantages accrue to each of the parties when they come together. The scientist often receives significant personal financial rewards as well as funds for research. Collaborations between academic and industrial labs serve to extend the capacity and output of each. The university receives overhead. The company obtains the rights for a potentially lucrative product. And the product, if it survives the obstacle course between the lab and the bedside, will move into clinical use in a much shorter time. But, a cynic might recall, there is no such thing as a free lunch. What price is paid to achieve these benefits? The answer depends on who responds to the question.

The academic scientist finds herself taking a crash course in business and law. The demands of negotiating agreements and writing patents drain time and energy. Some research activities are redirected from basic science toward more immediately practical goals. The promise of continuing industrial support is seductive but inevitably tied to commercial products and the bottom line. The lab may find itself focused on an agenda set by the company. The basic research that sparked the initial effort may lie fallow. The spontaneity of scientific pursuit, so prized by those lucky enough to have investigator-initiated government research grants, may be restricted. The speed with which the professor can share data or new reagents may be slowed. The result, in the worst scenario, would be deleterious for the lab, harmful for science, bad for society.

Happily, such wholesale commandeering of academic labs does not occur if the lab maintains support from several sources. The decreased availability of government and foundation funds, however, makes the worst-case scenario an ominous possibility.

For the university, there are other concerns. If a university stands to gain financially from the commercialization of one of its professor's inventions, for example, the institution may hesitate, out of fear of conflict-of-interest issues, to participate in clinical trials of the product. Such a policy, however, may engender friction and frustration in the relationship between university administrators and faculty members. Distrust can be heightened if the negotiations with companies are handled by an official who represents the university and not the interests of the faculty.

Universities themselves have faced the frustration of licensing their inventions to biotech companies that have then sublicensed them to larger firms for enormous fees. Because these "fees" can be disguised by a variety of accounting procedures, there is no way for the university (or the inventor) to participate in the profits of the sublicensing agreement. Thus, unless the invention becomes a product, the profits made by the biotech companies by "flipping the contract" are not shared by the university or the inventor.

Meanwhile the company writes the checks. Yet, of the three parties involved, it compromises the least in its time-honored modus operandi. It has obtained an idea or a product and will use it to benefit itself and the public good. In the process, of course, it must contend with the touchy issue of academic freedom while controlling access to information about the product—but the firm conducts business as usual.

So the scientist, the university and industry find themselves on a three-way street where ideas from the academic laboratory move into the realm of application. Because the use of this highway has increased dramatically in recent years, traffic jams and collisions have been unavoidable. And, increasingly, basic research is diverted from its path. Inevitably, such sidetracking will slow the movement of basic science discoveries into biotechnical products.

Preventing this slowdown requires some new rules of the road. Increased government funding for research is necessary to restore order by redirecting lab efforts back toward basic research—the wellspring of all applied technologies. The scientist and the university must cease regarding companies as intellectual lightweights with deep pockets and learn from the business world how economic reality must be intercalated into idealistic goals. And the company attitude that "the scientist has done the easy work" has to give way to a more inclusive approach that permits participation by the scientist and the university in deciding on the best road to development. Without these accommodations on all sides, the flow of ideas into products will be slowed, and all parties, including society at large, will suffer from the gridlock.

SUSAN ZOLLA-PAZNER is scientific director of the Research Center for AIDS and HIV Infection at the New York Veterans Administration Medical Center and professor of pathology at the New York University Medical Center.