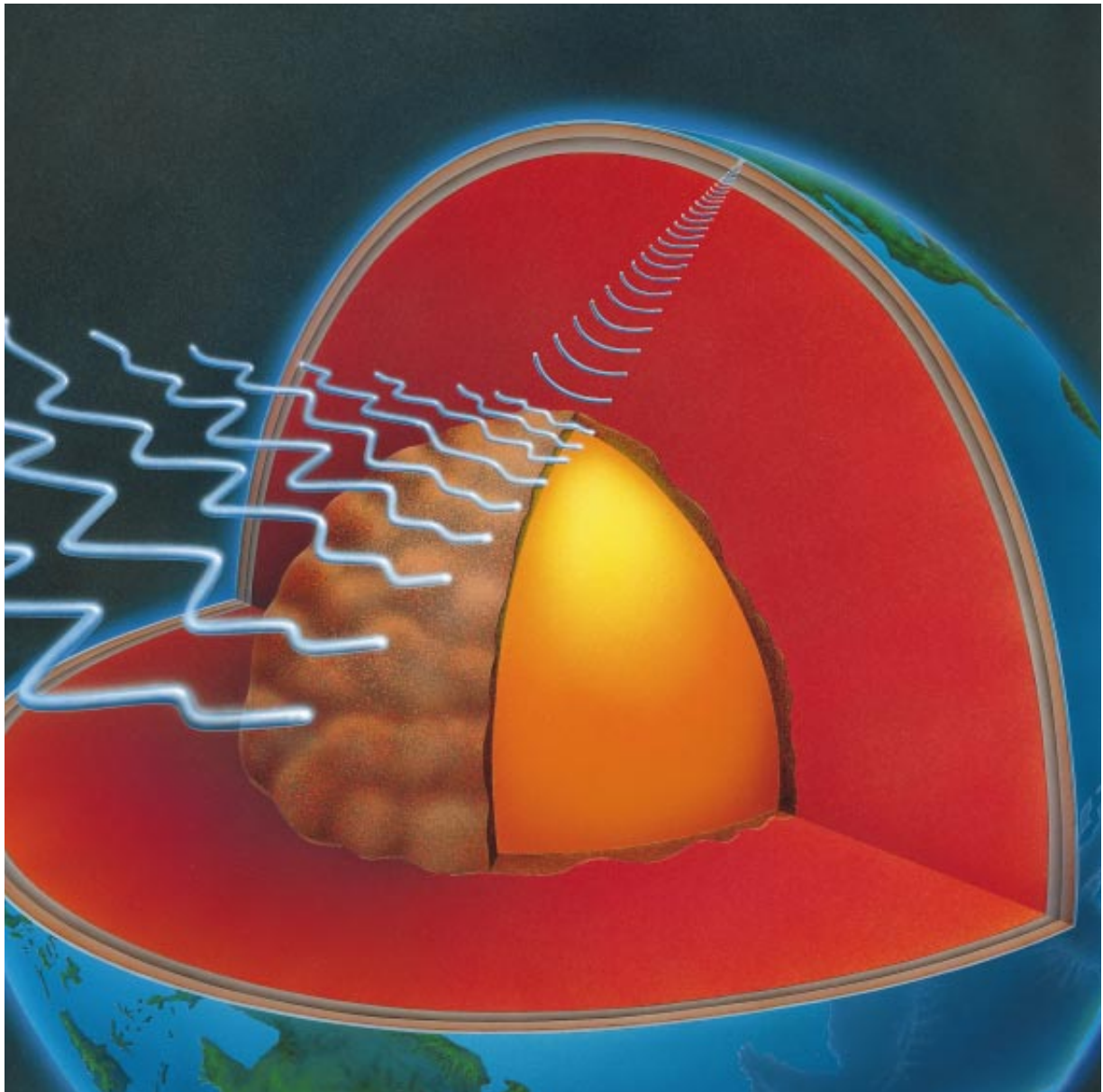


SCIENTIFIC AMERICAN

MAY 1993
\$3.95

*Building soft machines from smart gels.
The neurological pathways of fear.
Life and death as economic indicators.*



*Seismic waves trace the turbulent boundary
between the earth's rocky mantle and molten core.*

40

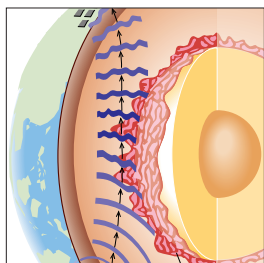


The Economics of Life and Death

Amartya Sen

The health of nations is normally charted in statistics that reveal only the wealth of nations: financial indicators such as gross national product and the balance of payments. Yet such statistics say little about human well-being, especially where famine and hunger persist. But if economists supplement such figures with mortality data, the social benefits and deficiencies of alternative strategies can be assessed.

48

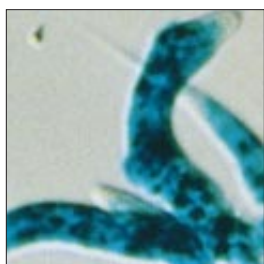


The Core-Mantle Boundary

Raymond Jeanloz and Thorne Lay

The region with the most intense geologic activity is not on the earth's surface. It lies 2,900 kilometers down, where the rocky mantle meets the planet's molten core. This turbulent interface has been found to influence the earth's rotation and its magnetic field. Advances in seismology and high-pressure experiments have enabled geophysicists to elucidate the boundary's physical and chemical interactions.

56

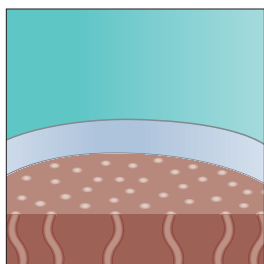


How Cells Respond to Stress

William J. Welch

Thirty years ago biologists discovered that cells defend themselves from heat damage by producing a group of specialized proteins. These protective molecules have now been shown to play an important role in helping cells withstand a broad range of assaults, from disease to toxins. Exploring this mechanism may provide new ways to combat infection, autoimmune disease and even cancer.

82

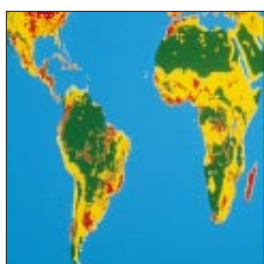


Intelligent Gels

Yoshihito Osada and Simon B. Ross-Murphy

Industrial designers usually prefer materials that are tough, hard and dry. But a few researchers are exploring applications for substances that are soft and wet. Gels that swell or shrink in response to a stimulus can deliver controlled doses of medicine or act as selective filters and valves. They may even result in "soft" machines that work, as muscles do, by contracting and relaxing.

88



SCIENCE IN PICTURES

The Power of Maps

Denis Wood

Even the most accurate of modern maps incorporate assumptions and conventions from the society and the individuals who create them. An awareness of the cartographer's bias is essential to interpreting the information that maps contain.

94



The Neurobiology of Fear

Ned H. Kalin

Studies of monkeys have begun to reveal the neurological pathways that underlie fear-related behavior. The work may lead to an understanding of the ways in which the various brain systems contribute to inordinate fear in humans; eventually they may open up new approaches to easing and preventing anxiety and depression.

104



P.A.M. Dirac and the Beauty of Physics

R. Corby Hovis and Helge Kragh

To this towering figure in 20th-century theoretical physics, the effort to describe natural phenomena was a search for mathematical perfection. Between the ages of 23 and 31, Dirac achieved his goal through a series of important theories in quantum mechanics, including the prediction of the existence of antimatter.

110



TRENDS IN ASTROPHYSICS

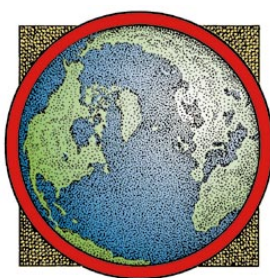
Inconstant Cosmos

Corey S. Powell, staff writer

Recent satellite observations of the cosmos in the high-energy spectrum would startle most earthbound stargazers. Some objects suddenly flare, then fade to obscurity; others flicker or flash on and off like neon signs. Astronomers are increasingly convinced that the engines powering many of these violent and baffling entities are the most mysterious denizens of the universe: black holes.

DEPARTMENTS

18



Science and the Citizen

Premature rumors of an AIDS treatment? ... Immune imbalance.... Venus in the eye of the beholder.... Final thoughts of a dying computer.... When anybody can get public data.... PROFILE: Science philosopher Paul K. Feyerabend.

12



Letters to the Editor

Asian schools.... Coming to an understanding.... Linguistic spat.

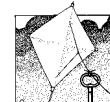
16



50 and 100 Years Ago

1943: Insurers seek the ideal weight for longevity.

134



The Amateur Scientist

Charting a watershed to make a cartographer's point.

138



Book Reviews

Richard Leakey continues his search for humanity's origins.

144



Essay: W. Brian Arthur

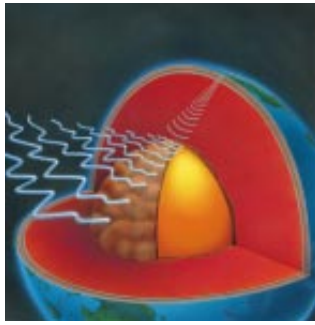
Complexity: the force that keeps things simple.

122



Science and Business

An activist administration tackles technology policy.... Success for Silicon Glen?... Battling MS.... Flat screens from light-emitting polymers.... Waste to slag.... THE ANALYTICAL ECONOMIST: Why the same job pays more (or less).



THE COVER painting provides a cutaway view of the earth's interior to reveal how a seismic wave is reflected and distorted by the unusual D'' layer. Such seismic-wave perturbations indicate that the region, which lies between the mantle and outer core, varies markedly in composition and thickness. Experiments simulating the conditions of the deep earth suggest that the zone between mantle and core may be the most chemically dynamic part of the planet (see "The Core-Mantle Boundary," by Raymond Jeanloz and Thorne Lay, page 48).

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Alan Hall, *Executive Editor*; Michelle Press, *Managing Editor*; John Rennie, Russell Ruthen, *Associate Editors*; Timothy M. Beardsley; W. Wayt Gibbs; Marguerite Holloway; John Horgan, *Senior Writer*; Philip Morrison, *Book Editor*; Corey S. Powell; Philip E. Ross; 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*; Nisa Geller, *Photography Editor*; Johnny Johnson

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*; Tanya DeSilva, *Prepress*; Carol Hansen, *Composition*; Madelyn Keyes, *Systems*; Leo J. Petrucci, *Manufacturing & Makeup*; Carl Cherebin

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

ADVERTISING: Robert F. Gregory, *Advertising Director*. OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager*; William Buchanan, *Manager, Corporate Advertising*; Peter Fisch, Elizabeth Ryan, Michelle Larsen, *Director, New Business Development*. CHICAGO: 333 N. Michigan Avenue, 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; Kate Dobson, *Advertising Manager*; Tonia Wendt, Lisa K. Carden, Lianne Bloomer, San Francisco. 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; Barth David Schwartz, *Director, Special Projects*, Amsterdam. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

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

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

PRESIDENT AND CHIEF EXECUTIVE OFFICER: John J. Hanley

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

CHAIRMEN OF THE BOARD: Dr. Pierre Gerckens
John J. Hanley

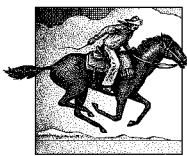
CHAIRMAN EMERITUS: Gerard Piel

PRINTED IN U.S.A.

THE ILLUSTRATIONS

Cover painting by Tomo Narashima

Page	Source	Page	Source
40-41	© 1986 Raghbir Singh		Libraries; Ken Pelka (<i>photograph</i>);
42	Johnny Johnson		<i>bottom</i> : National Anthropological Archives, Smithsonian Institution;
43	Les Stone/Sygma		Victor Krantz (<i>photograph</i>)
44-46	Johnny Johnson		William F. Haxby
47	A. Tannenbaum/Sygma	92	<i>top</i> : Conservation International;
49	Adam M. Dziewonski, Harvard University, and John H. Woodhouse, University of Oxford; photoshop by Dimitry Schidlovsky	93	<i>middle</i> : W. T. Sullivan, Hansen Planetarium Publications; Beth Phillips (<i>photograph</i>); <i>bottom</i> : © Stuart L. McArthur; Beth Phillips (<i>photograph</i>)
50-54	Ian Worpole		Ned H. Kalin
55	Ian Worpole (<i>left and right</i>), Douglas L. Peck (<i>center</i>)	95	Carol Donner (<i>top</i>), Ned H. Kalin (<i>bottom</i>)
56-57	E.P.M. Candido and E. G. Stringham, University of British Columbia; <i>Journal of Experimental Zoology</i> , © John Wiley & Sons, Inc.	96-97	Carol Donner
		98	Ned H. Kalin (<i>top</i>), Carol Donner (<i>bottom</i>)
58	J. Bonner, Indiana University; Dimitry Schidlovsky (<i>top</i>), Dale Darwin/Photo Researchers, Inc. (<i>middle</i>), S. Lindquist, University of Chicago (<i>bottom</i>)	101	Ned H. Kalin
		105	AIP Meggers Gallery of Nobel Laureates
59-62	Dimitry Schidlovsky	106	Courtesy of AIP Emilio Segrè Visual Archives (<i>left</i>), UPI/Bettmann Archive (<i>center</i>), courtesy of AIP Niels Bohr Library; Francis Simon (<i>right</i>)
83	Yoshihito Osada		Courtesy of AIP Emilio Segrè Visual Archives; Francis Simon
84	Ian Worpole; Yoshihito Osada (<i>photograph</i>)	107	Courtesy of Florida State University, Tallahassee
85	Ian Worpole		George Retseck
86-87	Ian Worpole (<i>top</i>), Yoshihito Osada (<i>bottom</i>)	108	Dennis Bracke/Black Star (<i>left</i>), COMPTEL team (<i>right</i>)
88-89	Tom Van Sant/GeoSphere Project (<i>bottom</i>), NASA (<i>top right</i>)	110-111	Michael Goodman
		112-113	Max Planck Institute for Extraterrestrial Physics, Garching, Germany
90	John W. Williams, University of Pittsburgh; Gabor Kiss (<i>top</i>), <i>Commentary on the Apocalypse of Saint John</i> , by Beatus of Liebana, Pierpont Morgan Library (<i>bottom</i>)	114	Lund Observatory; data courtesy of Gerald Fishman, NASA Marshall Space Flight Center
		115-116	Robert Prochnow
91	<i>top</i> : from <i>Geography</i> , by Claudius Ptolemy, The Murray Collection; <i>middle</i> : from <i>Cary's New Universal Atlas</i> , Smithsonian Institution	117	Westchester Land Trust
		118	
		135-136	



LETTERS TO THE EDITORS

Mathematics in Motion

I was delightfully surprised by "A Technology of Kinetic Art," by George Rickey [SCIENTIFIC AMERICAN, February]. It was an excellent choice to complement "Redeeming Charles Babbage's Mechanical Computer," by Doron D. Swade, in the same issue.

From a picture, we can visualize how the intricate, gleaming brass cams, linkages, gears, levers and dials in Babbage's difference engine work in unison. Yet even with time-lapse photography and knowledge of pendulums and balance beams, it is more difficult to visualize the beautifully random motions that Rickey's sculpture traces with only a whisper of wind.

I imagine that for many the article was an intriguing introduction to the technology of Rickey's art. For a mesmerizing feast for the eyes, try to locate one of Rickey's shows and see the art of the technology.

GEORGE SHERWOOD
Ipswich, Mass.

Failing Marks

We are disturbed and disappointed by Harold W. Stevenson's article, "Learning from Asian Schools" [SCIENTIFIC AMERICAN, December 1992]. As educators living in Japan who also have experience with elementary schools in the U.S., we are sure that the study he describes is neither good science nor useful scholarship.

The Sendai area is not representative of Japanese elementary schools as a whole, nor does it have much in common with Chicago. Sendai is a rural community recently inundated by suburban development and its attendant demographic changes. The uses and social meaning of university education in Japan are far different from those in the U.S. Regional differences among U.S. schools were also ignored.

The result of Stevenson's efforts is a set of dubious facts that doesn't match our own or our associates' teaching experience. The article omits that the vast majority of fifth graders in Japan attend *juku* (cram school) as many times each week as they attend regular school. How many "seat hours" does a kid in Osaka rack up on the average day when

he gets out of *juku* sometime between 6 and 10 P.M.? Why are our Japanese colleagues so worried about this idyllic system?

ROBIN AND THOMAS KITE
Osaka, Japan

Stevenson replies:

The Kites' informal observations fail to be convincing in the face of data from a series of major studies conducted during the past decade. That work involved 20,000 students and many of their parents and teachers in Sendai, Taipei, Beijing, Chicago, Minneapolis, Fairfax County in Virginia, Szeged in Hungary and Alberta, Canada.

The vast majority of Japanese elementary school students do not attend *juku*: even by sixth grade, no more than a third do so, even in Japan's largest cities. *Juku* attendance is a high school phenomenon among students seeking entrance to universities. Sendai is not a recently populated rural community; it has been one of the major cities of Japan for centuries. The Japanese may be more critical of their schools than Americans because they believe even a good product can be made better.

Sharp Words over Linguistics

I must protest the publication of "Linguistic Origins of Native Americans," by Joseph H. Greenberg and Merritt Ruhlen [SCIENTIFIC AMERICAN, November 1992]. The Greenberg classification of Native American languages has been rejected over and over in peer review. By Greenberg's own account, 80 to 90 percent of linguistic specialists reject his proposals. Criticisms of his work include the stunning number of errors in his data, languages classified on the basis of little or no data and the mistaken classification of a scholar's name as a language. He groups some words on the basis of accidental similarities while also missing true cognates. He stops after assembling similarities among compared languages—but that is where other linguists begin.

Greenberg's methods have been disproved. Similarities between languages can be the result of chance, borrowing, onomatopoeia, sound symbolism and other causes. For a proposal of remote family relationship to be plausi-

ble, one must eliminate the other possible explanations.

LYLE CAMPBELL
Department of Geography
and Anthropology
Louisiana State University

Greenberg and Ruhlen reply:

Although many Americanists reject our findings, the same tripartite classification has been discovered independently by geneticists. Many Russian linguists and others do accept our results. As for the methodology having been disproved, Greenberg's universally accepted classification of the African languages demonstrates just the opposite. In fact, our methods are the only way to discover language families: nonobvious cognates can generally be recognized only after the language families have been identified on the basis of their similarities. Campbell and his colleagues have never discovered a single family or a single new linguistic relationship. Their methods are apparently so precise that they have no results.

The Science-Reader Barrier

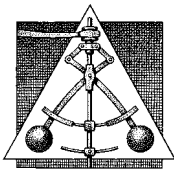
I want to commend Elaine Tuomanen for "Breaching the Blood-Brain Barrier" [SCIENTIFIC AMERICAN, February]. How rare it is to read an article by a scientist that is clear to the many of us who are interested in her area of expertise but are not knowledgeable enough to understand its complexities. Tuomanen sets an excellent example with her writing.

GLENN C. WATERMAN
Bainbridge Island, Wash.

ERRATA

On page 41 of "Environmental Change and Violent Conflict" [February], the population densities in Senegal and Mauritania should have been stated as 38 people per square kilometer and two people per square kilometer, respectively.

The color key for the chart of bridge condition versus age on page 72 of "Why America's Bridges Are Crumbling" [March] was not printed. The colors are: brown, timber; blue, steel; green, reinforced concrete; and red, prestressed concrete.



50 AND 100 YEARS AGO

MAY 1943

"All in all, longevity is probably the best single index of 'ideal' weight. A large-scale study by the Metropolitan Life Insurance Company has shown definitely that at the young adult ages a moderate degree of overweight was beneficial, but that beginning at about 35, the advantage lay with women of average weight. In middle age and beyond, the underweights had the best longevity record. Even in young people, the advantage of a moderate degree of overweight has been diminishing, because two important diseases—tuberculosis and pneumonia—which have largely accounted for the excess mortality among young underweights in the past, have been brought under control."

"In a recent discussion of helicopters, Igor Sikorsky revealed that his present model has flown at a maximum speed of 80 miles an hour, has carried two people, and has extreme ease of control and smooth riding qualities. He has estimated that during early production of helicopters the price would probably be comparable to that of a medium-priced airplane; in quantity production the cost would undoubtedly approach that of a medium-priced automobile."

"Preliminary tests have revealed that the powerful X-rays from the betatron

have the special advantage of producing their greatest effect about 1½ inches below the surface of the body. With X-ray therapy as used up to the present time, the effect is greatest on the surface, and decreases with depth. Direct use of the high-speed electrons from the betatron may be even more valuable than the use of the X-rays. Most of the X-rays continue beyond the point of treatment to pass entirely through the patient. The electrons would not do this. At 20 million volts they will penetrate as far as four inches, and no farther. The region of maximum effect should be about three inches beneath the surface, according to calculations by Philip Morrison, of the University of Illinois physics staff."

"Our search for human origins is complicated by the possibility that a varied assemblage of human types simultaneously existed in the lower (earlier) Ice Age. Which of these types is truly ancestral to modern man? Or have several played their part and was *Homo sapiens* from the start something of a mongrel breed? To none of these questions can science as yet provide an exact answer. But the bones from the Barnfield Pit at Swanscombe, if the rest are ever found, may indicate the solution to a major question in human prehistory: Whether, that is, a form approximating our own species in appearance had attained such status far back

in the dim vistas of the earlier Ice Age or whether, on the other hand, we, as individuals, derive from a big-browed human line, like Neanderthal, which remained primitive in all its major aspects down into the period of the last ice advance."



MAY 1893

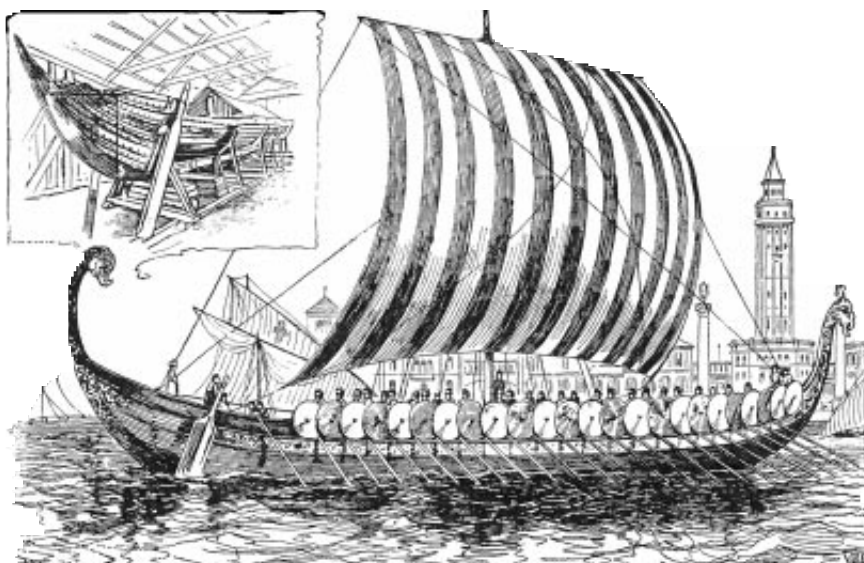
"In an interview on the subject of the extensions and alterations of the elevated railway system by a *Tribune* reporter with one of the directors, the latter evidently expressed himself somewhat differently from what he intended.

"Reporter: 'Do you think the present elevated structure strong enough to support the further weight of three tracks and more rapid trains?'

"Mr. Sloan: 'Certainly; you have no idea of the anxiety with which our engineers watch the present structure. It is carefully examined continually.'"

"From the experiments recently performed in electrical oscillations, the conclusion that light and electrical oscillations are identical is very strongly substantiated. The principal parts in which they practically agree are the velocity, rectilinear propagation, laws of reflection, interference, refraction, polarization and absorption by material substances. In fact, the sole certain difference appears to be the wave length. In the domain of wireless telegraphy this subject is of prime importance. Although existing methods are far from perfect, we can confidently expect that in the near future we will be able to telegraph on land and sea without wires by means of electrical oscillations of high power and frequency."

"Within a comparatively recent period the remains have been dug up, at various places in Norway, of ancient Scandinavian vessels, models of which are to be exhibited at Chicago. Our illustration (*left*) represents one of these models, which has recently sailed for America, after visiting most of the towns on the Norwegian coast. It is an exact copy of an old Viking vessel, the remains of which were discovered in 1880, near Sandefjord, Norway."



Model of a Viking ship



Triple Whammy

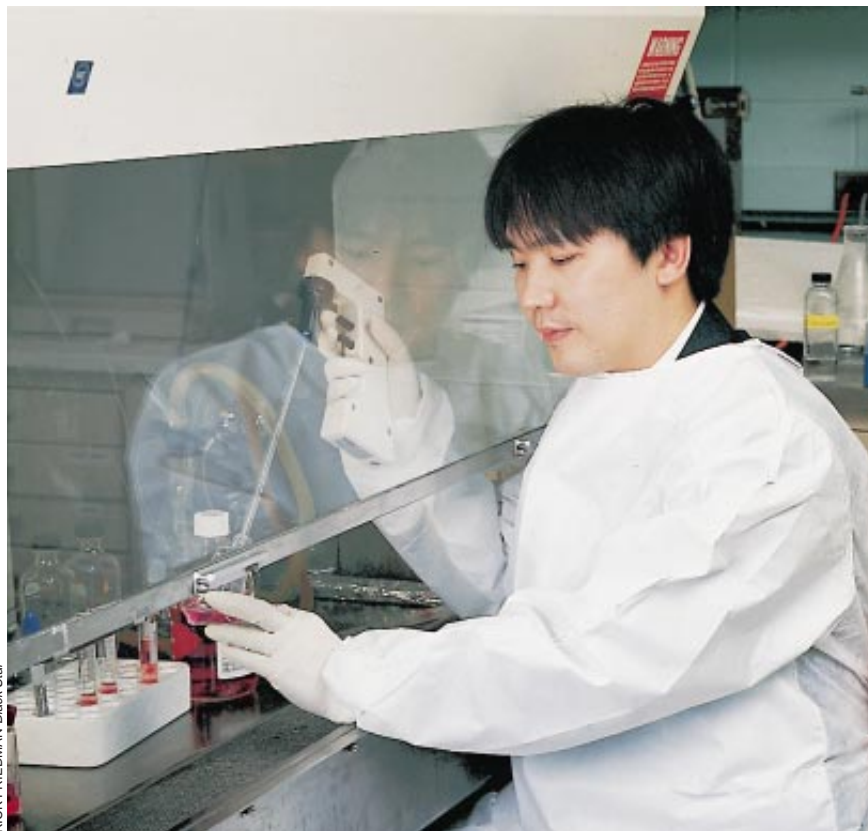
Will an AIDS therapy live up to its advance billing?

The whole hullabaloo is completely out of proportion," fumes Douglas D. Richman, an immunologist at the University of California at San Diego. He is troubled by the message that he feels the public is getting about convergent combination therapy, an experimental AIDS treatment discovered by Yung-Kang Chow, a 31-year-old student at Harvard Medical School. Following widespread press coverage, desperate AIDS patients are reportedly clamoring for places in the imminent clinical trials of the new therapy this spring.

Richman is not a critic of the work itself—in fact, he wrote a favorable commentary on the possibilities of combined convergent therapy that accompanied the February report in *Nature* by Chow, Martin S. Hirsch, Richard T. D'Aquila and their colleagues at Harvard Medical School and Massachusetts General Hospital. "I think the authors of the paper were perfectly honest and straightforward in saying what they had to say," he explains. "It's just that the paper was taken out of context, which I think is bad for everybody."

He is not alone. Although most AIDS investigators praise Chow's group for having achieved an interesting result in the test tube, they express concern that—as has happened with other new leads in AIDS research—serious reservations about efficacy and safety are being ignored. "The kind of play that it is getting runs the risk of creating incentives for patients to leave proven therapies to try unproven therapies," warns Daniel F. Hoth, director of the Division of AIDS at the National Institute of Allergy and Infectious Diseases (NIAID).

The essence of Chow's announcement was that by using a combination of three drugs, he and his colleagues stopped a strain of human immunodeficiency virus (HIV) from replicating in cultures of isolated blood cells. In itself, that result is not new. "This is not the first time that HIV has been eliminated from cultures," notes Anthony S. Fauci, director of NIAID. Nor is the use of more than one drug an innovation: combination approaches are under study in many laboratories. Used individually, anti-



RICK FRIEDMAN/Black Star

NEW AIDS THERAPY devised by Yung-Kang Chow, a student at Harvard Medical School, relies on three drugs that converge on a viral molecule.

ral drugs gradually lose their potency against HIV, probably because mutant forms of the virus become resistant.

But in recent years, when researchers have tried to develop combination therapies against HIV, they heeded the grandmotherly advice "Don't put all your eggs in one basket." They used drugs that attacked the virus at different stages of its life cycle because the odds of a virus simultaneously developing resistance to diverse drugs are slight.

Chow's inspiration was to contradict that orthodoxy. He used three drugs—zidovudine (also called AZT), dideoxyinosine (ddI) and either nevirapine or pyridinone—that all act against the enzyme reverse transcriptase, which is essential to the replication of HIV. Viruses can become resistant to any one of those drugs by developing small mutations in their gene for reverse transcriptase. Chow noticed, however, that the mutant forms of reverse transcriptase are slightly less enzymatically efficient.

Convergent combination therapy cap-

italizes on the accumulation of those inefficiencies: in viruses resistant to all three drugs, a mutant reverse transcriptase cannot do its job. Chow showed in the test tube that viruses exposed to his drug combination died or became unable to replicate. After the infected cells died, workers could detect no virus in the cell cultures.

"The concept of using multiple drugs targeting the same enzyme has been around for a very long time," notes Warner C. Greene, director of the Gladstone Institute of Virology and Immunology at the University of California at San Francisco. AZT and ddI, for example, have been used together in clinical trials for several years simply because they are both good antiviral agents. The genetic rationale behind convergent combination therapy does mark a conceptual advance. Nevertheless, on a practical level, the approach only means using three drugs instead of two.

The clinical trials will be a critical test of convergent combination therapy. So

far it is completely uncertain how well—if at all—it will work in people. The viruses in Chow's cultures did not find a useful defense against the drug trio, but the amount of HIV inside a person is much greater. "It's a question of probability," explains Mathilde Krim, co-founder of the American Foundation for AIDS Research in New York City. "I think if you waited long enough, you probably would see resistance to even three drugs." Moreover, HIV infection in the body is not restricted to short-lived blood cells like those in Chow's cultures. HIV can hide inside neurons and other cells that might serve as viral reservoirs for the recurrence of infections. Therefore, convergent therapy would likely be only another way of maintaining a patient's health until a cure can be found.

The individual and combined side effects of the drugs must also be taken into account. In all combination therapies, as Greene notes, the hope is that the synergistic effect of the drugs will be so great that the dosages and side effects of each one can be minimized. AZT can cause anemia and damage to peripheral nerves; ddi can produce severe inflammation of the pancreas. Small doses can often moderate the harmful effects, but some patients still have severe reactions and cannot bear to take those drugs. Nevirapine, an unapproved drug under development by Boehringer Ingelheim Pharmaceuticals in Ridgefield, Conn., seems to have relatively few or mild side effects, but it has been taken by only a tiny handful of patients so far.

According to Maureen Myers, a nevirapine researcher at Boehringer Ingelheim, the company has been reluctant to expose a large clinical population to the drug until more was known about it. Yet that is exactly what will happen in the upcoming trials of convergent combination therapy. The accelerated schedule for the start of the trials "is putting serious compromises on the question of how much safety data we'll have on the drug interactions," she says. "It's on a pretty fast track, and it got on a faster one when the publication appeared in *Nature*."

In some researchers' eyes, NIAID may be partly responsible for the attention that Chow's report received. On the heels of the *Nature* paper, NIAID announced that it was "accelerating the trial design process" with the intention of starting clinical trials of convergent therapy during the spring. Initially the trials were to involve 200 people at 10 research centers throughout the U.S.; later they were expanded to include 400 people at 16 centers. D'Aquila and Hirsch will oversee the trials.

The results will probably determine

how the move for expedient testing is viewed. If those patients seem to benefit from convergent therapy, the decision to test without hesitation may be hailed for its humanitarianism. On the other hand, the rush to the clinic "adds quite a bit of credibility that wasn't there in the absence of Tony Fauci's action," Greene observes.

Fauci denies that he has exaggerated the importance of Chow's work and points out that the clinical trials will quickly settle many of the unresolved questions about the therapy. Hoth elaborates that the larger the trials, the sooner a reliable verdict on the therapy will be available. When asked whether the outpouring of public interest had affected the size of the trials, Hoth replied, "You'd have to ask Marty Hirsch that question." Neither Hirsch, D'Aquila nor Chow was available for comment.

Whatever the results of convergent combination therapy, many researchers remain convinced that combination therapy in some form will be the most fruit-

ful approach to treatment. If nothing else, investigators point out, any renewal of interest in combination therapies also reinvigorates the research programs for all drugs, including ones such as nevirapine that were dogged with resistance problems when used alone.

Nevertheless, those same researchers also emphasize that the need to develop new drugs and vaccines against HIV is as great as ever. Greene expresses doubts about "whether or not one can combine imperfect agents and make a more perfect therapy—I think the future of AIDS therapy rests with the development of new agents."

In the meantime, however, Greene decries the harm that excessive optimism about preliminary research does to AIDS patients. "It's just a roller-coaster ride for these folks. We buoy them up, and then we drop them," Greene says sadly. "I think we have to be a lot more circumspect about how we handle these small, incremental increases in our knowledge."
—John Rennie

Balanced Immunity

Would killing some T cells slow the progress of AIDS?

The death of the white blood cells called *T* lymphocytes leaves AIDS patients vulnerable to lethal infections. Paradoxically, however, some researchers now suspect that decimating the ranks of those *T* cells might extend the health of people infected with human immunodeficiency virus (HIV). They believe that by struggling to maintain the quantity rather than the variety of its cells, the immune system sets itself up for disaster. "The homeostatic mechanism that maintains the *T* cell count is blind," says Leonard M. Adleman of the University of Southern California, one of the idea's originators.

All *T* cells are not alike: they are morphologically uniform, but their behavior and molecular markings differ. One large set of *T* cells, called killer lymphocytes because they attack infected tissues, carries a surface protein known as CD8. A second set, the helper *T* cells that seem to coordinate the immunologic assault, bears the protein CD4 instead.

As medical researchers have known for more than 10 years, HIV hits the CD4 *T* cells particularly hard. Healthy and newly infected persons have more than 800 CD4 *T* cells in each cubic millimeter of their blood plasma, but that number gradually declines during the decade-long latency period usually associated with AIDS. The infections char-

acteristic of AIDS often set in after the CD4 *T* cell count drops below 200.

But, in Adleman's words, "losing a *T* cell is not like losing an arm or a leg." The body routinely replaces *T* cells lost through bleeding or disease by making new ones. Even HIV-infected people can generate new *T* cells, at least until late in their illnesses. Why the CD4 *T* cell population shrinks in people who have HIV has therefore been a mystery.

Adleman and others have recently suggested that a flaw in the immune system's approach to self-repair may aggravate the damage done by the virus. The problem, they say, is that the homeostatic mechanism monitoring the levels of the *T* cells does not distinguish between those bearing the CD4 protein and those bearing CD8. Consequently, when CD4 cells die, "it detects the loss and causes the generation of new *T* cells until the total *T* cell count is back to normal," Adleman explains. "But it does that by producing both CD4 and CD8 *T* cells." In effect, the addition of the CD8 cells suppresses the production of new CD4 cells. As the virus continues to kill cells selectively and the immune system replaces them generically, the population of CD4 *T* cells declines.

This past February in the *Journal of Acquired Immune Deficiency Syndromes*, Adleman and David Wofsy of the University of California at San Francisco described their test of that hypothesis. Using monoclonal antibodies, they eliminated the CD4 *T* cells from mice. As predicted, the total number of *T* cells soon returned to normal, but the pop-

ulation consisted entirely of CD8 cells.

In the same issue, Joseph B. Margolick of the Johns Hopkins School of Hygiene and Public Health and his colleagues also advanced that idea, supporting it with data from the Multicenter AIDS Cohort Study. Margolick found that the *T* cell population did shrink slightly during the first 18 months after HIV infection but that thereafter it stayed fairly steady for years: increases in the number of CD8 cells had offset the drop in CD4 cells. "The total change in *T* cells is not very much compared with the change between those populations. That suggests there is some sort of compensation going on," he notes. "It may be that the people who are the longest-term survivors are the ones with the best compensatory mechanisms."

The Adleman and Margolick findings build on similar observations by other researchers working with genetically engineered mice and with cancer patients who have received bone marrow grafts. "I think the concept of *T* cell homeostatic mechanisms being at work has been pretty well established," says Anthony S. Fauci, director of the National Institute of Allergy and Infectious Diseases, who wrote an editorial accompanying the Adleman and Margolick papers. "Whether or not that is going to explain some of the phenomena we see in HIV is unclear at this point."

Indeed, many aspects of the blind homeostasis model, as Adleman calls it, are still hazy. Immunologists are still in the dark about how the immune system counts or regulates the number of *T* cells. "We're viewing it as a black box," he concedes.

Nevertheless, even at a broad conceptual level, the model does raise new therapeutic possibilities. One is that physicians might be able to rebalance the im-

mune system by eliminating 10 to 15 percent of a patient's CD8 *T* cells every six months or so. If the model is correct, the immune system should respond by producing both CD4 and CD8 cells. Pruning the CD8 cell cadre might briefly weaken the immune responses, Margolick acknowledges, but most of the eliminated cells would probably not be relevant to the patient's infections. "You have to weigh the balance," he says. "If you get more CD4 cells back, that may compensate for the loss of the few HIV-significant CD8 cells."

Fauci thinks that approach deserves further investigation in animals, particularly in monkeys infected with the related simian immunodeficiency virus (SIV). One technical obstacle to pursuing such experiments in monkeys—or in humans, for that matter—is that no one has yet developed monoclonal antibodies or other agents that can selectively kill CD8 *T* cells. "But those can be developed; that's not totally prohibitive," Fauci adds.

A gentler approach might be to stimulate the production of more CD4 cells. If researchers can discover the chemical cues that signal an immature *T* cell to differentiate as either a helper or a killer cell, Adleman believes there is at least a possibility that those cues could be used "to trick the immune system into pumping out new CD4 cells."

Immunology is Adleman's adopted field: he is best known as a computer scientist and a co-inventor of an encryption system for electronic mail. He was first drawn to immunology because the subject "stimulated" him and because its unsolved problems "had the kind of beauty mathematicians look for." Leave it to a mathematician to notice when something in the immune system does not add up.

—John Rennie

Honest Advertising

Why ostentatious antlers are like an expensive car

What do a Porsche and the antlers of a red deer stag have in common? Both are impressive, certainly. And according to a once unpopular theory that has made a remarkable comeback, that is the key to why a red deer stag grows antlers and to why people who can't really afford them buy expensive cars.

By virtue of price alone, the car delivers an unmistakable message: the owner of this indulgence must have economic power and the status that goes with it. Antlers, despite their size, are not much use for fighting, and the effort of growing them and carrying them around is substantial. But they presumably indicate to other stags—as well as to does—that their owner has a healthy constitution. After all, the bearer can sustain the waste of a lot of protein that could be made into useful things, such as muscles.

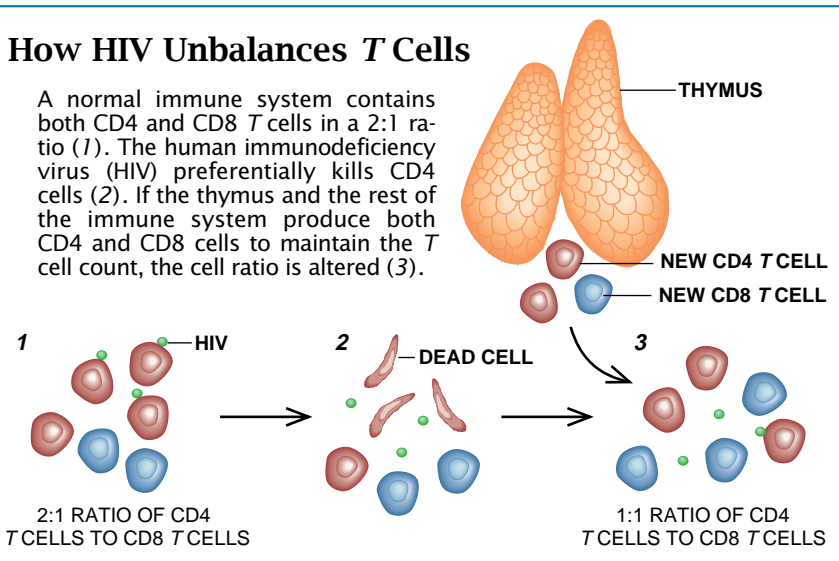
The notion that the extravagant features of many animal displays might be advantageous precisely because they lower viability was first proposed in 1975 by Amotz Zahavi, a researcher at Tel Aviv University. Because the idea, known as the handicap principle, is so paradoxical, it attracted a lot of attention. Consider, for example, the handicap explanation for "stotting." Some antelopes stot, or jump vertically into the air, if they spot a lion. Zahavi's explanation is that the antelope is trying to persuade the lion that the chase would not be worth it: that a prey animal that can deliberately waste time and effort stotting instead of running would be too swift to catch.

After a number of thoughtful papers had been written on the subject, however, the consensus among animal behaviorists was that the handicap principle simply could not work. But Alan Grafen, a behavior theorist at the University of Oxford, has recently set a cat among the pigeons. His series of mathematical models, he maintains, shows that under a wide range of conditions Zahavi's idea does indeed make sense. The gist of his conclusion—supported by several other workers—is that a biological signal such as a pair of antlers actually must have a "cost," or deleterious effect on viability, if it is to be taken seriously. Furthermore, the cost must be one that stronger individuals can pay more easily than their weaker brethren.

In Grafen's view, the cost or handicap is a guarantee of the honesty of the dis-

How HIV Unbalances *T* Cells

A normal immune system contains both CD4 and CD8 *T* cells in a 2:1 ratio (1). The human immunodeficiency virus (HIV) preferentially kills CD4 cells (2). If the thymus and the rest of the immune system produce both CD4 and CD8 cells to maintain the *T* cell count, the cell ratio is altered (3).



play. If there were no cost, there would be rampant cheating, and observers would quickly learn to ignore the false advertising. "You can't argue with success," the saying goes, and so it is that paste diamonds will never have the cachet of the real things, even if they glitter just as much. Likewise, evolution produces cumbersome antlers because conveying an unmistakable message about

one's superior constitution more than compensates for the aggravation.

One of the implications of Grafen's work is that animal signals should be, on average, "honest." Because antlers are costly, it would not be worthwhile for a weak stag to produce very large antlers and so try to bluff his way to holding a harem. The expenditure also means that animal signals might often provide some

clue to their meaning. "The best way to show you are very rich would be to burn a million-dollar bill," Grafen says. "Actually sending the signal is cheap because it takes no time or effort." Similarly, the best way for a peacock to show that he has been healthy—an important consideration for an interested peahen—might be for him to show off an elaborately patterned tail that takes months

Three Faces of Venus

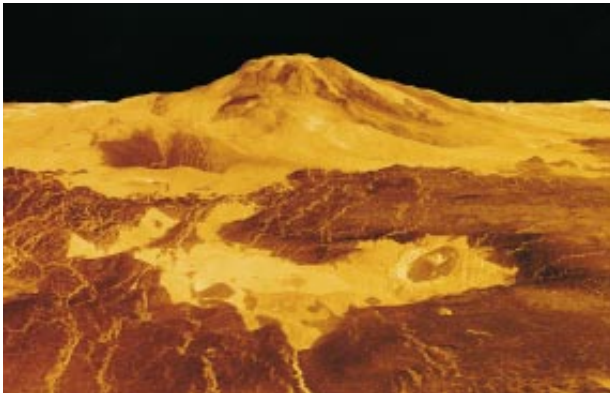
For centuries, astronomers squinted and stared through their telescopes in the vain hope of catching a glimpse of the surface of Venus, Earth's cloud-enshrouded planetary neighbor. The National Aeronautics and Space Administration's *Magellan* probe has changed all that. Since *Magellan* began to orbit Venus in 1990, planetary

scientists have been practically drowning in a sea of images.

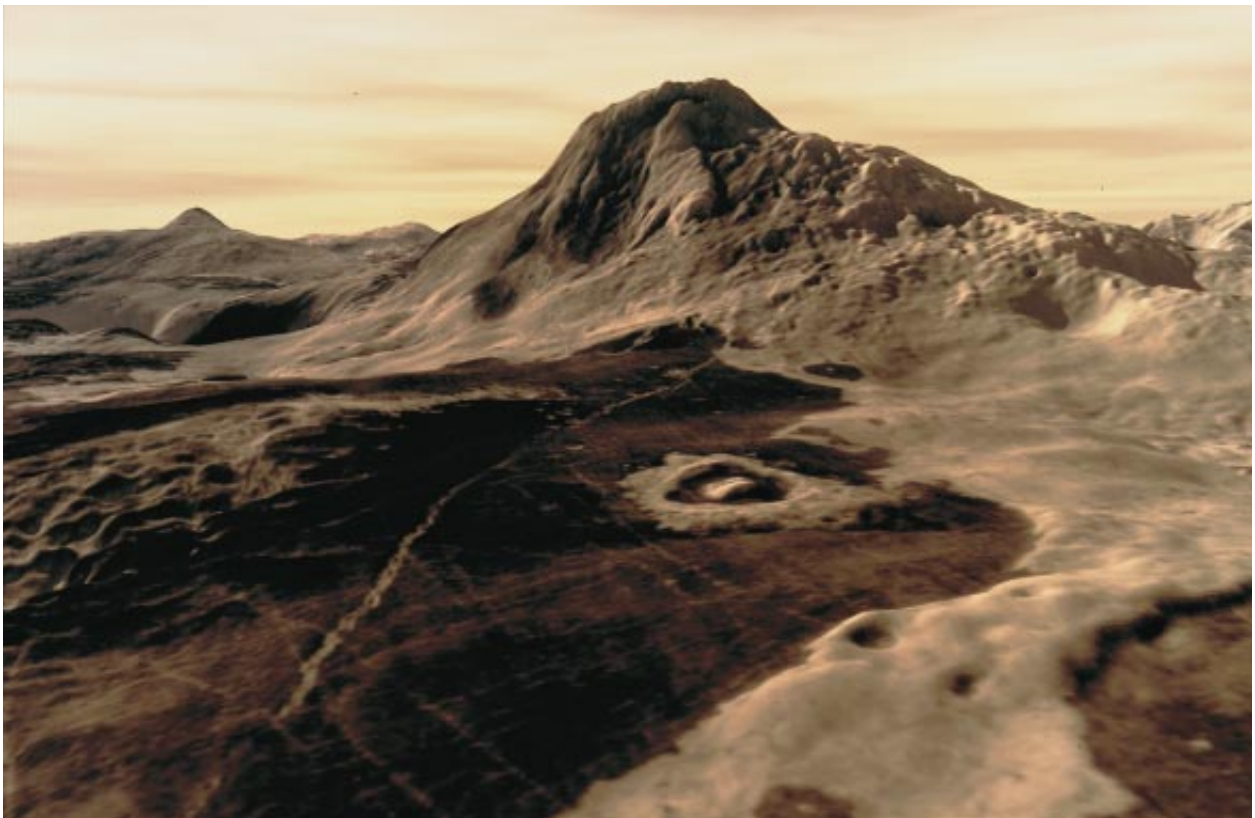
Magellan's completed radar map of Venus will contain roughly three trillion bits of data, thousands of times as much information as is contained in the entire *Encyclopaedia Britannica*. Converting that giant catalogue of radar echoes into intuitively meaningful pictures posed a challenge to researchers at the Jet Propulsion Laboratory in Pasadena, Calif., which issues the official NASA images.

The laboratory team has now received a creative helping hand from other scientists who are taking advantage of the wide dissemination of the *Magellan* data and the ready availability of powerful computer graphics programs. The images shown here demonstrate three different philosophies about how best to display *Magellan's* scientific bounty—and to depict an unveiled Venus.

The now familiar NASA image (*top left*) shows a view of the five-kilometer-high Venusian volcano known as Maat Mons. The brightness of each part of the image simply indicates how well the local terrain reflects *Magellan's* radar, which is influenced both by the roughness of the surface and by its inclination. To clarify the topography, workers



NASA/JET PROPULSION LABORATORY



DAVID P. ANDERSON/Southern Methodist University

to grow but requires little exertion to display.

Critics are still considering the implications of the resurgent handicap principle. Marian Stamp Dawkins and Tim Guilford, also at Oxford, point out that the handicap principle does not necessarily mean that every individual instance of a biological signal is honest, even if signals are truthful on the whole.

In addition, they believe that when the receiver as well as the transmitter of a signal has to pay a penalty, cheating or bluffing might occur more frequently. For example, red deer stags hold roaring matches to determine who gets access to a harem. But both challenger and harem master end up exhausted after such a contest.

Similar situations are common, Daw-

kins and Guilford note, and they think this and other complications—such as the psychology of the receiver—will often lead to the evolution of inexpensive signals that are open to cheating. Grafen accepts that his revamping of handicap theory will not be the last word on animal signaling. But, he says, “at least now we have competing theories to evaluate. That’s healthy.” —*Tim Beardsley*

at the Jet Propulsion Laboratory magnified the relief by a factor of 10 and inclined the image to simulate a perspective view. More controversial is the electric orange coloration, chosen to mimic how the surface might appear when illuminated by the reddened sunlight that filters through Venus’s thick atmosphere. Of course, the *Magellan* images are produced by radar, not visible light, and the jet-black skies contradict the illusory sense of realistic color.

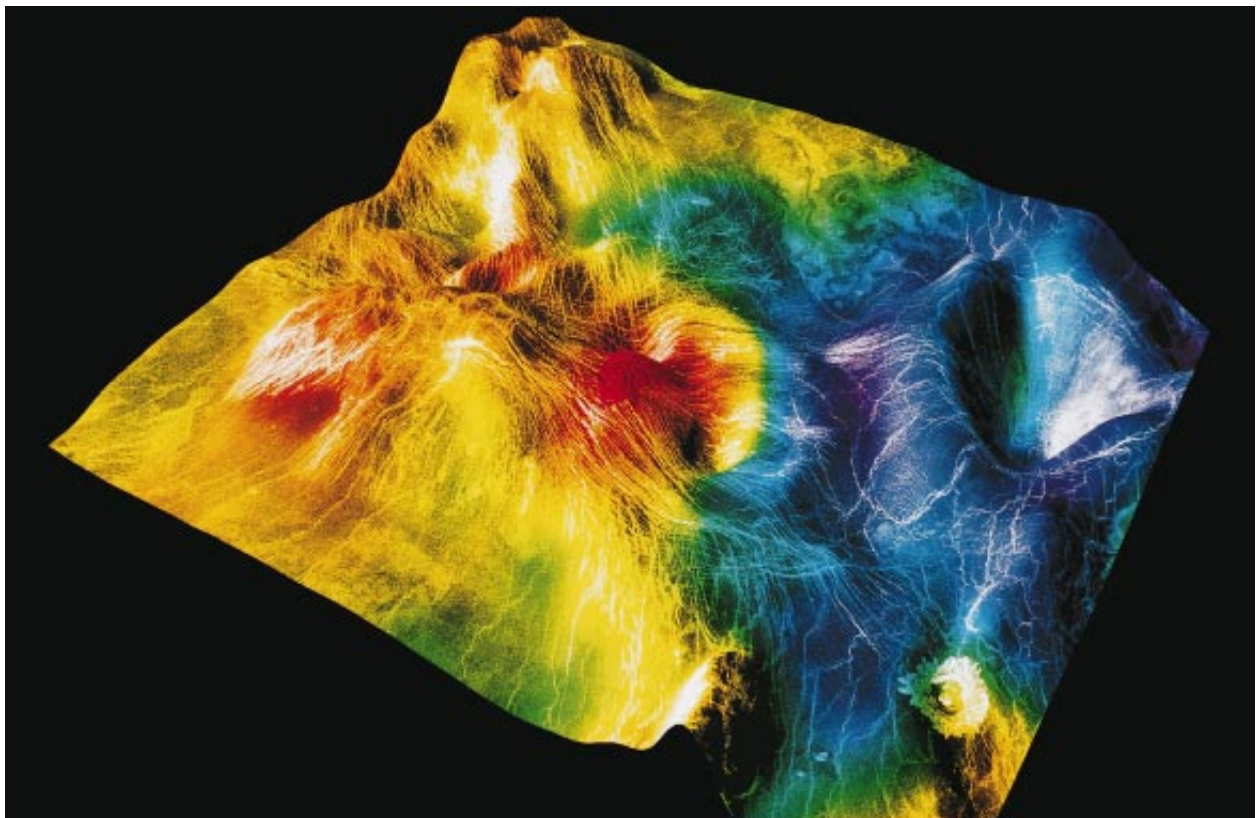
David P. Anderson of Southern Methodist University has produced a more “Earth-like” view of Maat Mons using the same *Magellan* data set (*bottom left*). The most noticeable difference is Anderson’s palette. He based the hues of the ground on the color of basalt, the kind of rock thought to cover most of Venus’s surface. The clouds were introduced “mostly for aesthetic reasons,” he explains but adds that they provide a background that enhances the sense of depth perception. The form of the clouds was based on educated guesses about the appearance of the Venusian sky.

Such window dressing is of secondary importance to Anderson, however; “the hardest part is getting the topography right,” he says. Employing techniques derived from fractal geometry, Anderson has produced topography that he considers to be more realistic than that in the NASA im-

ages; he then used a sophisticated ray-tracing program to give the resulting landscape a plausible, solid appearance.

Given that the *Magellan* radar images have no inherent color, Randolph L. Kirk, Laurence A. Soderblom and Ella M. Lee of the U.S. Geological Survey in Flagstaff, Ariz., have experimented with applying tints to depict a property known as emissivity, the degree to which the hot rocks on the Venusian surface naturally emit microwave radiation (*below*). Emissivity is lowest for rocks that are smooth and electrically conductive. Here rocks having the lowest emissivity appear violet, and those having the highest emissivity are colored red; intermediate values move through the spectrum.

Kirk and his colleagues exaggerated the topography of a volcanic region called Sigrun Fossae by a factor of 100. The patterns of emissivity may indicate surface weathering or variations in the composition of the local lava flows, Soderblom notes. Kirk’s group opted to portray the emissivity data in bright, saturated colors that the eye can easily decode. The surreal beauty of the resulting landscape testifies to just how far astronomical images have moved beyond the literal, magnified vistas witnessed by the observer crouching at the end of the eyepiece. —*Corey S. Powell*



U.S. GEOLOGICAL SURVEY

Make, Model and . . .

A privacy advocate puts license plates on line

If you drive a car in Massachusetts, Simson L. Garfinkel probably knows who you are. This past March, David Lewis of the Massachusetts Registry of Motor Vehicles told a session at the Computers, Freedom and Privacy Conference in San Francisco that the agency is required by law to sell its registration file for the cost of copying. "So how much does it cost?" asked Garfinkel, a computer journalist and technical adept. "What fields does it contain?"

The answer: \$77 for a magnetic tape containing nine million registration records with the make, model and year of each car, plus the name and address of the owners, the date of registration and any liens against the vehicle.

Garfinkel hopes to make the file—all two gigabytes or so—available to one and all for searching via computer network as an exercise in freedom of infor-

mation. A data-base consultant at the conference estimated that a high-end personal computer could process several requests per second from car thieves, stalkers, marketers, the merely curious and other agents of social and economic change. California restricted access to its motor vehicle files four years ago, after an aberrant fan tracked down actress Rebecca Schaeffer through her automobile registration and killed her. But registration and license records are open to the public in most states. So are court records, real-estate title listings and even, in some cases, the files of public gas and electricity companies.

So, what has been protecting our privacy? Mainly time and trouble. In the past, those wishing to search public records either had to pore through stacks of documents or find a mainframe computer. Garfinkel's plan, however, highlights the growing conflict between the presumption of open public records and citizens' desire for privacy. Desktop computers can now assemble a dossier of financial, medical and other information at the touch of a few keys.

Advocates of free access to such information assert that it can be used to lubricate the wheels of commerce, aid in medical care or improve the quality of government. For example, everyday credit card transactions rely on financial data bases. In some hospitals, physicians can retrieve patients' records in seconds instead of an hour or more (about half the time, paper records arrive too late to be of any use, notes Eunice Little of the American Health Information Management Association). And shortly after traffic citation records became available in Massachusetts, Lewis pointed out, newspaper reports exposed an appeals commission that was letting off up to two thirds of the drunk drivers who appeared before it. Such an investigation would have been virtually impossible without a computerized search.

The privacy-minded rebuttal by pointing out the hazards that accompany easy access to information. Although murders aided by public and private data bases are rare, tales of financial damage are widespread. Indeed, Jack H. Reed, chairman of Information Resource Ser-

Attractive and Demure

The devil is in the details. Although for decades physicists have understood how the fundamental forces of nature influence some of the most esoteric elementary particles, they have suddenly realized that they do not know what actually holds the nucleus of an atom together. "For a long time, we have had a very simple picture, but now it seems too simplistic," comments George F. Bertsch, a nuclear theorist at the University of Washington.

Physicists had assumed that the protons and neutrons that make up the core of an atom attract one another by exchanging a particle known as a pi meson, or pion. But recent results from particle accelerators show that the pion is responsible only for conveying the nuclear force over long distances. And no one has figured out what is happening over the short range.

To be sure, a vast distance in this context is, by any conventional scale, close to nothing. Because the diameter of a proton is only one fermi—that is, a millionth of a billionth of a meter—nuclear physicists consider a distance of a few fermi to be a long haul.

The idea that a particle carries the nuclear force can be traced back to the work of Nobel laureate Hideki Yukawa in the 1930s. His theory was confirmed in 1947, when British physicist Cecil Frank Powell and his co-workers discovered the pion. Yukawa originally predicted that the pion would mediate all nuclear interactions.

But things got complicated during the 1970s, when investigators demonstrated that protons, neutrons and pions are themselves composed of elementary particles known as "up" quarks, "down" quarks and gluons. A proton is made of two up quarks and one down quark; a neutron is one up and two down. A pion can consist of an up quark and the antimatter counterpart of a down quark, but pions can also be made of certain other pairs of quarks. In pions,

neutrons and protons, the quarks are held together by gluons, which convey the strong force, just as photons carry the electromagnetic force.

Gluons and quarks must ultimately be the carriers of the nuclear force, but the question is what combination of gluons and quarks really do the job. By the early 1980s physicists had figured out that various pairs of quarks could carry nuclear forces, but pions, they believed, played the most important role.

Then, in 1986, researchers at Los Alamos National Laboratory tried to observe the exchange of pions by bombarding atomic nuclei with protons. The Los Alamos group found that pions did not seem to be involved in short-range nuclear interactions. After a series of experiments that culminated last summer, physicists have been forced to conclude that pions carry the nuclear force only over distances of 0.5 fermi or more. "Although a fraction of a fermi does not seem like very much, that distance scale is crucial to all nuclear processes," says Joel M. Moss, one of the principal investigators on the Los Alamos team.

Unfortunately, the new findings do not give physicists many clues about how protons and neutrons do interact at close range. The nuclear force could, quite possibly, be conveyed over short distances by a particle heavier than the pion. A more intriguing idea is that gluons are directly involved in carrying nuclear forces over short distances. Researchers have established only that gluons exist inside protons and neutrons; if gluons do jump between protons and neutrons in an atomic nucleus, physicists would be forced to rewrite nuclear theory.

"We need to know much more about the internal structure of protons and neutrons before we can really say we understand the forces that bind nuclei together," Bertsch explains.

—Russell Ruthen

vice Company, a personal-data seller, told the conference audience how he had been denied a mortgage because of a misleading credit report. Insurers, who subscribe to a centralized medical-information data base, have been accused of denying coverage to people who have had themselves tested for HIV, even if they test negative, on the theory that being worried enough to take the test implies risky behavior.

These kinds of potential abuses are becoming more important as lawmakers (and private companies) put personal data to uses for which it was never intended. Federal law, for example, now supports using motor vehicle information to track parents whose child-support payments are late; a single database entry can cause computers to issue a warrant for the alleged deadbeat's car to be seized. In a striking mismatch of crime and punishment, Massachusetts legislators recently proposed blocking license renewal for citizens with unpaid library fines. "We told them they were crazy," Lewis notes. If automotive files, containing only name, address, vehicle identification number and a few other bits of information, can spur such controversy, what of medical information? Clinton administration policymakers regard automated medical records as a

crucial ingredient in cutting health care costs—Rene C. Kozloff, a project officer at Kunitz and Associates, a health-management information firm, anticipates a "conception to death record" stored on either smart cards or a central data base.

Yet there are minimal controls over the five or six dozen people who may handle those records as a result of a visit to a hospital or clinic. Given the problems that have been caused by disclosure of medical records kept on paper, opening such information to massive, uncontrolled computer searches seems unwise, says Janlori Goldman of the American Civil Liberties Union.

Privacy advocates have been working for nearly 20 years for a so-called Fair Information Practices Act that would give the subjects of public and private data bases power over how personal information on them is used. Although pro-privacy forces have thus far been unsuccessful in the U.S., they have had more luck in Europe. The British enacted "Data Protection" rules in 1984, and a privacy directive for the European Community is in draft form.

British law requires businesses that keep data bases to register them with the Data Protection Registrar, to ask for people's consent before gathering information about them and not to use those

data for a purpose different from the one for which they were collected. "Information about others is held in trust" rather than being owned by data-base compilers, says Rosemary Jay, legal adviser for the registrar. Jay has brought court challenges against credit-reporting agencies; she has also had to deal with direct marketers seeking access to the registrar's list of data bases. "Cheeky," she comments. —Paul Wallich

"Daisy, Daisy"

Do computers have near-death experiences?

What does a computer do when it starts to die? The HAL 9000 in the film *2001: A Space Odyssey* burst into a rendition of "A Bicycle Built for Two," a song it had been taught early in life. The memorable scene may not be too far off the mark. That's what one researcher found out when he began to "kill" a type of computer program known as an artificial neural network. As the network approached death, it began to output not gibberish but information it had previously learned—its silicon life flashed before its eyes, so to speak.

The analogy to a so-called near-death experience is irresistible because the creators of artificial neural networks design them to mimic the structure and function of the biological brain. A neural network relies on "units" to serve as the cell body of a neuron and "links" between the units to act as the interconnecting dendrites and axons. The units are typically organized into several layers. A consequence of such an architecture is that the network, like the brain, can learn. In a real brain, learning is thought to occur because of changes in the strength of synaptic connections among neurons. Similarly, a neural network alters the strength of the links (specifically, the weighting between units) to produce the correct output. Typically a programmer teaches a network by repeatedly presenting training patterns to it [see "How Neural Networks Learn from Experience," by Geoffrey E. Hinton; *SCIENTIFIC AMERICAN*, September 1992].

Properly trained neural networks can handle diverse tasks, from compressing data to modeling dyslexia. Stephen L. Thaler, a physicist for McDonnell Douglas, began to explore neural networks a year ago as a way to optimize the process control of diamond crystal growth. But curiosity led him to start annihilating neural nets as an evening avocation.

AnthroCart®

LOTS OF SIZES AVAILABLE

LIFETIME WARRANTY

HOLDS 150 LBS.

ADJUSTABLE

ANTHRO

Call for a free catalog:
800-325-3841
6:30 AM - 5:00 PM, PST, M-F

Anthro Corp. • 3221 NW Yeon St., Portland, OR 97210 • (503) 241-7113 • FAX: (503) 241-1619

He devised a program that would gradually destroy the net by randomly severing the links between units. "The method was meant to emulate the depolarization of the synapses in biological systems," Thaler says. After each successive pass, he examined the output.

When about 10 to 60 percent of the connections were destroyed, the net spat out nonsense. But when closer to 90 percent of the connections were destroyed, the output began to settle on distinct values. In the case of Thaler's eight-unit network, created to model the "exclusive or" logic function, much of what was produced was the trained output states 0 and 1. The net sometimes generated what Thaler terms "whimsical" states, that is, values that neither were trained into the net nor would appear in a healthy net. In contrast, untrained networks produced only random numbers as they died.

That an expiring net would produce meaningful gasps is not entirely far-fetched. "It makes sense in terms of a network that has made some stable patterns," says David C. Plaut, a psychologist and computer scientist at Carnegie Mellon University who uses artificial neural nets to model brain damage. Indeed, Thaler has a detailed explanation. In a fully trained, functioning network, all the weighted inputs to a particular unit are about the same in magnitude and opposite in sign. (In math speak, the weights follow a Gaussian distribution, or bell-shaped curve.) The odds are, then, that the sum of several weighted inputs to a unit equal zero. Hence, when the links are broken, the unit might not "feel" the loss, because it may have been receiving a total zero signal from them anyway. The few surviving links will often be sufficient to generate reasonably coherent output.

But concluding that this artificial experience can be extrapolated to human brushes with death is a stretch. "Neural networks have got to be a rough approximation at best," Plaut notes. The brain is far more sophisticated than neural nets. Furthermore, it is not entirely clear how collections of real neurons die. The death of a few neurons, for instance, may kill off some nearby ones. And the method used to train the neural nets—an algorithm called back-propagation—is dissimilar to the way the brain learns.

Still, the observations suggest that some of the near-death experiences commonly reported might have a mathematical basis. "It may not just be fancy biochemistry," Thaler asserts. He is currently working on more complex networks, including one that will produce visual images. Any wagers for a light at the end of a long tunnel? —Philip Yam



The best Kentucky
folklore has always been
passed down orally.



WILD TURKEY

101 proof, real Kentucky.

Wild Turkey® Kentucky Straight Bourbon Whiskey, 50.5% Alc./Vol. (101°), Austin, Nichols Distilling Co., Lawrenceburg, KY. © 1998 Austin, Nichols & Co., Inc.



UNCONVENTIONAL WISDOM

They're The McLaughlin Group. Each with a view that's contentious and contagious. (clockwise from left) Jack Germond, Clarence Page, John McLaughlin, Eleanor Clift, Morton Kondracke and Fred Barnes.

Made possible by a grant from GE.

The McLaughlin Group

Check your local listing for station and time.



We bring good things to life.



PROFILE: PAUL KARL FEYERABEND

The Worst Enemy of Science

In 1987 *Nature* published an essay in which two physicists deplored a growing public skepticism toward science. The physicists blamed this insidious trend on four philosophers who have attacked traditional notions of scientific truth and progress: Karl R. Popper, who proposed that theories can never be proved but only falsified; Imre Lakatos, who contended that scientists ignore falsifying evidence; Thomas S. Kuhn, who argued that science is a political rather than rational process; and Paul K. Feyerabend.

The physicists singled out Feyerabend as “currently the worst enemy of science.” Photographs published along with the essay seemed to confirm that view. Popper, Lakatos and Kuhn wore sober expressions, as if this business of pointing out the shortcomings of science somehow pained them. Not so Feyerabend: smirking at the camera over glasses perched on the tip of his nose, he was clearly either anticipating or relishing the perpetration of some great mischief. He looked like an intellectual Loki.

Which he is. For decades, the Austrian-born Feyerabend (pronounced fire-AH-bend) has waged war against what he calls “the tyranny of truth.” By deconstructing such scientific milestones as Galileo’s trial before the Vatican and the development of quantum mechanics, he has insinuated that there is no logic to science; scientists develop and adhere to theories for what are ultimately subjective and even irrational reasons. According to Feyerabend, there are no objective standards by which to establish truth. “Anything goes,” he says.

It is all too easy to reduce Feyerabend to a grab bag of outrageous sound bites. He has likened science to voodoo and witchcraft, and biologists performing experiments on animals to Nazis. He has defended the attempts of fundamental-

ist Christians to have their version of creation taught alongside the theory of evolution in public schools. He ends his *Who’s Who* entry with the statement “Leading intellectuals with their zeal for objectivity . . . are criminals, not the liberators of mankind.”

Beneath these provocations lies a serious message: the human compulsion to find absolute truths, however noble, too often culminates in tyranny of the mind, or worse. Only an extreme skepticism toward science—and open-mindedness toward other modes of knowl-



CHRISTIAN KAENZIG/Black Star

FEYERABEND has been called the Salvador Dali of philosophy.

edge and ways of life, however alien—can help us avoid this danger. Feyerabend expresses this view in a paradox in his 1987 book *Farewell to Reason*: “The best education consists in immunizing people against systematic attempts at education.”

In spite of—or because of—his rhetorical excesses, Feyerabend has found a broad audience. His first book, *Against Method*, has been translated into 16 languages since it was published in 1975

and remains a staple of courses on the philosophy of science. Even some scientists confess to a grudging admiration. The late physicist Heinz R. Pagels called Feyerabend “a punk philosopher” but added, “Probably some of Feyerabend’s views of science are correct if we could but see our science from the perspective of a thousand years hence.”

Oddly enough, Feyerabend, now 69, has always shunned publicity. Even before he retired in 1990 from the University of California at Berkeley and from the Federal Institute of Technology in Zurich, where he held joint appointments, he rarely granted interviews—or even answered his telephone. “You’ll

never reach him,” one former colleague assured me. Although I obtained and repeatedly called his number in Zurich, where he has a home, he never answered.

After I mailed him a letter requesting an interview, however, Feyerabend wrote back. He planned to visit friends in New York City. Perhaps we could meet there? Accompanying the letter was a photograph of Feyerabend, wearing an apron and a grin, leaning over a sink full of dishes. The letter explained: “I would like you to use the enclosed picture, which shows me at my favorite activity: washing dishes for my wife in Rome.”

I finally meet Feyerabend in a luxurious Fifth Avenue apartment belonging to a former student, one who wisely abandoned philosophy for real estate. He thrusts himself from a chair and stands crookedly to greet me, as if he has a stiff

back. His face, even more leprechaunlike in person than in the photograph in *Nature*, is astonishingly animated, as are his voice and hands. He declaims, sneers, wheedles and whispers—depending on his point or plot—while whirling his hands like a conductor.

Self-deprecation spices his hubris. He calls himself “lazy” and “a bigmouth,” and when I ask about his “position” on a certain point, he winces. “If you have a position, it is always something screwed

down," he says, twisting an invisible screwdriver into the table. "I have opinions that I sometimes defend rather vigorously, and then I find out how silly they are, and I give them up!"

Watching this performance with an indulgent smile is Feyerabend's wife, Grazia Borrini, a 40-year-old Italian physicist whose manner is as calm as her husband's is intense. Borrini, who met Feyerabend while studying public health at Berkeley a decade ago and married him six years later, enters the conversation sporadically—for example, after I ask him why he thinks some scientists are so infuriated by him. "I have no idea," he replies, the very picture of wide-eyed innocence. "Are they?"

"I was infuriated at first," Borrini interjects, explaining that she initially heard a caricature of Feyerabend's message from a hostile physicist. Only after meeting him and reading his books did she realize how subtle his views were. "This is what you should want to write about," she says to me, "the great misunderstanding." "Oh, forget it, he's not my press agent," Feyerabend snaps, then begins defending himself. "I go to extremes but not to the extremes I am accused of," he says. For example, he is not opposed to science, as some have claimed. "Science provides fascinating stories about the universe," he remarks. In fact, he asserts, modern scientists are every bit the equal of such ancient entertainers as myth-tellers, troubadours and court jesters.

It should come as no surprise that Feyerabend studied acting and singing as well as science while growing up in Vienna. He envisioned himself becoming both an opera star and an astronomer. "I would spend my afternoons practicing singing, and my evenings on the stage, and then late at night I would observe the stars," he says. Then the war came. Germany occupied Austria, and in 1942 Feyerabend enlisted in an officers' school, hoping—in vain—that his training would outlast the war. While fighting against (actually fleeing from) the Russians in 1945, he was shot in the spine. "I couldn't get up, and I still remember this vision: 'Ah, I shall be in a wheelchair rolling up and down between rows of books.' I was very happy."

He gradually recovered the ability to walk, with the help of a cane. Resuming his studies at the University of Vienna, he switched from physics to history, grew bored, returned to physics, grew bored again, and finally settled on philosophy. His ability to advance absurd positions through sheer cleverness led to a growing suspicion that rhetoric rather than truth is crucial for carrying an argument. "Truth itself is a rhetorical

term." Jutting out his chin, he intones mockingly, "I am searching for the truth." Oh boy, what a great person."

Within a decade after obtaining his doctorate in 1951, Feyerabend came to know all his fellow enemies of science. He and Lakatos both studied under Popper at the London School of Economics in the 1950s. "He was my best friend," Feyerabend says of Lakatos, who died in 1974. Feyerabend met Kuhn after mov-

Scientists are every bit the equal of ancient myth-tellers, troubadours and court jesters.

ing to Berkeley in 1959. Although he absorbed aspects of his colleagues' views, he finally rejected them as too conservative. He earned Popper's eternal hatred by deriding his theory of "critical rationalism" as "a tiny puff of hot air in the positivistic teacup." What Kuhn called "normal science," in which scientists are devoted to a dominant paradigm, Feyerabend called a "fairytale." He also claimed, to Kuhn's horror, that his sociopolitical model of science could apply to organized crime as well.

Feyerabend's skepticism deepened in the 1960s, when a growing number of Mexican, African-American and Indian students began attending Berkeley. "Who was I to tell these people what and how to think?" he recalls musing in his 1978 book *Science in a Free Society*: "Their ancestors had developed cultures of their own, colourful languages, harmonious views of the relations between man and man and man and nature whose remnants are a living criticism of the tendencies of separation, analysis, self-centredness inherent in Western thought." His task, he realized, "was that of a very refined, very sophisticated slave driver."

The solution to this crisis was to show students that knowledge can be judged only in context. So-called primitive societies such as the !Kung in Africa, Feyerabend notes, "survive happily; they don't need any gadgets. They survive in surroundings where any Western person would come in and die after a few days. Now you might say that people in this society live longer, but the question is, What is the quality of life? And that has not been decided."

Feyerabend is both amused and concerned by the belief of some physicists that they are approaching a "theory of everything." "Let them have their belief, if it gives them joy, but to tell the little children, 'That is what the truth is,' that

is going too far." Feyerabend contends that the very notion of "this one-day fly, a human being, this little bit of nothing" discovering the secret of existence is "crazy." "What they have figured out is one particular response to their actions, and the reality that is behind this is laughing, 'Ha ha! They think they have found me out!'"

The unknowability of reality is one theme of a book Feyerabend is writing, whose working title is *The Conquest of Abundance*. "The world is really abundant," he explains, "and all enterprises consist in cutting down this abundance. First of all, the perceptual system cuts down this abundance, or you couldn't survive. Now philosophers and scientists cut it down further." One threatened aspect of human thought is the conviction—embodied in religion—that the universe has some transcendent meaning. "I was brought up as a Roman Catholic," Feyerabend says. "Then for a very short time, I was a vigorous atheist, but now my philosophy has a different shape. It can't just be that the universe just goes 'boom!' and develops. Is there something else? There should be!"

The book may reveal a gentler Feyerabend. "I would plead guilty to being rude" in the past, he says. He regrets, for example, some of the "nasty things" he said about some of his fellow philosophers. "Today I would not be like that, because today I think of the person I am writing about. Unless the guy is a real bastard; then I don't mind." He has even asked *Who's Who* to delete his reference to intellectuals as "criminals."

"I thought so for a long time," he says of the quote, "but last year I crossed it out, because there are lots of good intellectuals." He turns to Borrini. "I mean, you are an intellectual," he says. "No," she replies dryly. "I am a physicist." He waves away her objection. "What does it mean, 'intellectual'? It means people who think about things longer than other people, perhaps."

I mention that another philosopher told me Feyerabend's relationship with Borrini had made him more "easygoing." Husband and wife both laugh. "Well, getting older you don't have the energy not to be easygoing, but she's certainly made a big difference," he says. "I was married three times before, but now for the first time I am so happy to be married."

Borrini beams. But when I ask if Feyerabend really enjoys washing dishes for her, as he claimed, she snorts. "Once in a blue moon," she says. "What do you mean once in a blue moon!" he cries. "Every day I wash dishes!" "Once in a blue moon," Borrini repeats firmly. Yet again, rhetorical excess has gotten Paul Feyerabend into trouble. —John Horgan

The Economics of Life and Death

Mortality data can be used to analyze economic performance. Such information can illuminate critical aspects of the economic organization of society

by Amartya Sen

Economics is not solely concerned with income and wealth but also with using these resources as means to significant ends, including the promotion and enjoyment of long and worthwhile lives. If, however, the economic success of a nation is judged only by income and by other traditional indicators of opulence and financial soundness, as it so often is, the important goal of well-being is missed. The more conventional criteria of measuring economic success can be enhanced by including assessments of a nation's ability to extend and to improve the quality of life.

Despite unprecedented prosperity in the world as a whole, famine and chronic hunger persist in many places. Avoidable disease and preventable deaths also remain widespread in industrialized countries as well as in the Third World. Economic arrangements are cen-

tral to these problems. By supplementing traditional indicators with statistics that relate more directly to well-being, the benefits and deficiencies of alternative economic approaches can be fruitfully assessed. For example, one country can have a much higher gross national product per capita than another; at the same time, it can have much lower life expectancy than its less wealthy counterpart because its citizens have poor access to health care and basic education. Mortality data can be used to evaluate policy and to identify vital aspects of economic deprivation in particular nations and in specific groups within nations.

The relevance and merit of mortality statistics can be illustrated by examining a series of problems chosen from around the world. These problems include devastating famine, which often takes place even though food is readily available; reduced life expectancy, frequently in countries with high GNPs; higher mortality rates for women than

AMARTYA SEN is Lamont University Professor and professor of economics and philosophy at Harvard University. He was educated in Calcutta and Cambridge, England, and has taught in both places as well as in Delhi, London and Oxford. Past president of the Econometric Society, the International Economic Association and the Indian Economic Association, Sen is currently president-elect of the American Economic Association. His research interests include social choice theory, decision theory, welfare economics and development economics as well as moral and political philosophy.

FEMALE STUDENTS pause on a street in the Indian state of Kerala. Kerala, which has one of the lower gross national products in the country, has high literacy rates for both sexes. Despite extreme poverty, public commitment to education and health as well as to improving the status of women has in general made the population of Kerala literate and long-lived. That fact illustrates that certain measures of economic success, such as GNP, can be incomplete.



for men in parts of Asia and Africa; and the very low survival rates of African-Americans, in comparison not only with those of whites in the U.S. but also with those of populations in some extremely poor countries.

Economic explanations of famine are often sought in measures of food production and availability. And public policy is frequently based on a country's aggregate statistics of the amount of food available per person, an indicator made prominent by Thomas Robert Malthus in the early 1800s. Yet contrary to popular belief, famine can result even when that overall indicator is high. Reliance on such simple figures often creates a false sense of security and thus prevents governments from taking measures to avert famine.

A more adequate understanding of famine requires examining the channels through which food is acquired and distributed as well as studying the entitlement of different sections of society. Starvation occurs

because a substantial proportion of the population loses the means of obtaining food. Such a loss can result from unemployment, from a fall in the purchasing power of wages or from a shift in the exchange rate between goods and services sold and food bought. Information about these factors and the other economic processes that influence a particular group's ability to procure food should form the basis of policies designed to avoid famine and relieve hunger.

The Bangladesh famine of 1974 demonstrates the need for a broader appreciation of the factors leading to such a calamity. That year, the amount of food available per capita was high in Bangladesh: indeed, it was higher than in any other year between 1971 and 1976. But floods that occurred from late June until August interfered with rice transplantation (the process by which rice seedlings are moved from the scattered locations where they were established to neat rows in wet fields) and other agricultural activities in the north-

ern district. Those disruptions, in turn, caused unemployment among rural laborers, who typically lead a hand-to-mouth existence. Bereft of wages, these workers could no longer buy much food and became victims of starvation.

Panic exacerbated the situation. Although the main rice crop, which had been only partly damaged by flooding, was not expected to be harvested until December, anticipation of a shortage led immediately to precautionary hoarding and to speculative stockpiling. All over the country, prices shot up sharply. As rice and other grains became more expensive, the food-buying ability of poor Bangladeshis plummeted. When food prices peaked in October, so also did the death toll.

At this point, the government, belatedly, began relief efforts on a large scale. Its response was delayed for several reasons, one being the suspension by the U.S. of food shipments, which resulted from a quarrel about Bangladesh's export of jute to Cuba. Yet one of the biggest obstacles was a false sense



of security evoked by high figures of food supply. Once relief was set in motion, the market began to readjust to a more realistic assessment of the winter harvest: the loss of crops was much more moderate than had been earlier assumed. By November, food prices started coming down; most relief centers were closed by the end of the month. The famine was mostly over before the partly damaged crop was even scheduled to be harvested.

As mentioned earlier, food levels per capita in Bangladesh were high in this year (because an excellent crop had been harvested in December 1973). The occurrence of this famine illustrates how disastrous it can be to rely solely

on food supply figures. Food is never shared equally by all people on the basis of total availability. In addition, private and commercial stocks of produce are offered to or withdrawn from the market in response to monetary incentives and expectation of price changes.

Famine has often taken place when statistics have shown little or no decline in food supply. During the Bengal famine of 1943, for instance, the diminished purchasing power of rural laborers' wages initiated widespread starvation. Similarly, in 1973 a famine in the Ethiopian province of Wollo was caused by a locally intense drought that impoverished the local population but did not substantially reduce food production

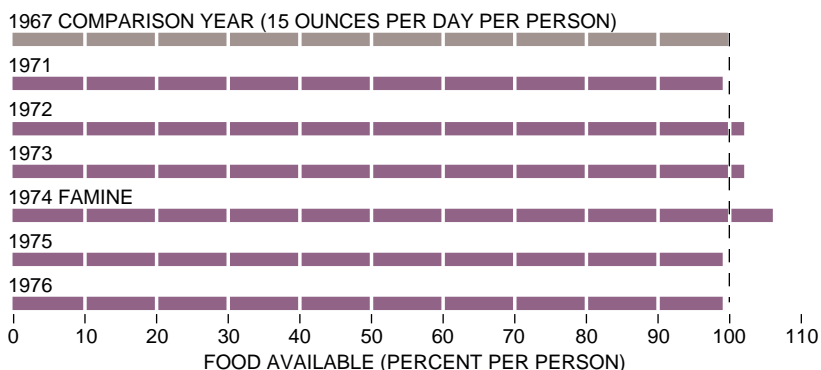
in the nation overall. Prices were often lower in Wollo than elsewhere in the country because the purchasing ability of the province's population was so reduced; some food, in fact, moved out of the famine-stricken region to more affluent areas. (This tragic turn of events also took place during the 1840s, when food was shipped from a starving Ireland to a prosperous England.)

There are several ways to prevent famine. In Africa and Asia, growing more food would obviously help, not only because it would reduce the cost of food but also because it would add to the economic means of populations largely employed in



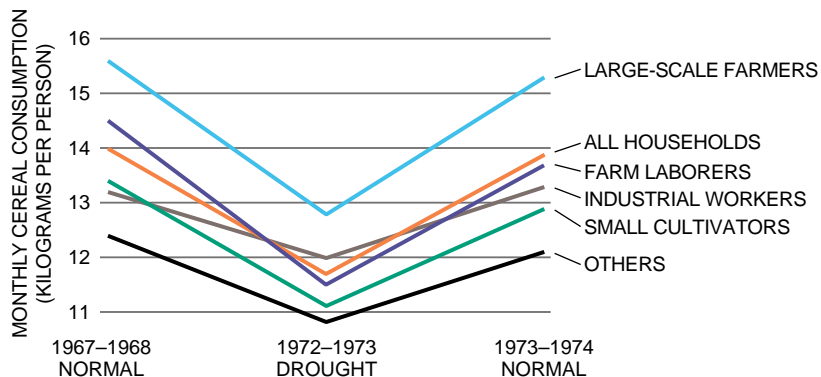
The Bangladesh famine of 1974 took place even though more food was available per person that year than in any other year between 1971 and 1976. (Food availability per year is indexed in relation to the base year of 1967.)

FAMINE AND FOOD SUPPLY IN BANGLADESH



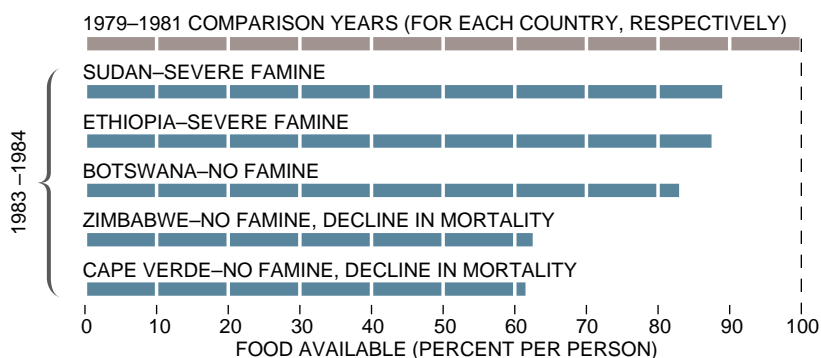
Maharashtra, India, prevented famine during a drought by establishing public works programs, which provided income to the needy. Everyone's consumption of cereal fell; the shortage was shared by all.

DROUGHT AND CEREAL CONSUMPTION IN MAHARASHTRA



Botswana, Zimbabwe and Cape Verde produced less food in 1983-1984 than in earlier years but did not experience famine, because they implemented public programs. Sudan and Ethiopia, which had less severe declines, did far less and suffered more.

FAMINE AND FOOD AVAILABILITY IN FIVE AFRICAN NATIONS



producing food. Enhancing production would require providing incentives to make investments in farming worthwhile. It would also necessitate policies such as expanding irrigation and encouraging technological innovation (which is much neglected in Africa).

Augmenting food production, however, is not the only answer. Indeed, given the variability of the weather, concentrating too much of a nation's resources on growing more food can increase the population's vulnerability to droughts and floods. In sub-Saharan Africa, in particular, there is a strong need for the diversification of production, including the gradual expansion of manufacturing. If people have the economic means, food can be purchased—if necessary, from abroad.

No matter how successful the expansion of production and diversification may be in many African and Asian countries, millions of people will continue to be devastated by floods, droughts and other disasters. Famine can be averted in these situations by increasing the purchasing power of the most affected groups—those with the least ability to obtain food. Public employ-

ment programs can rapidly provide an income. The newly hired laborers can then compete with others for a share of the total food supply. The creation of jobs at a wage does, of course, raise prices: rather than letting the destitute starve, such practice escalates the total demand for food. That increase can actually be beneficial, because it brings about a reduction in consumption by other, less affected groups. This process distributes the shortage more equitably, and the sharing can deter famine.

Such public works projects to avert famine would not typically impose an extraordinary financial burden on the government of a poor nation. Even though the absolute number of famine victims can be high, they tend to make up a small proportion of society: famine usually afflicts less than 5 to 10 percent of the population. Because those who starve are also among the poorest, their share of income or of food consumption is often between 2 and 4 percent. Thus, the fiscal resources needed to re-create their lost incomes are not impossibly exacting.

The success of the public employment approach to famine prevention is well illustrated. In the Indian state of Maharashtra, a series of severe droughts between 1972 and 1973 led to extensive agricultural unemployment and to a halving of the amount of food yielded. Public works programs—for example, the building of roads and wells—saved the affected laborers from starving. They could then compete with others for limited food. Although the average amount of food available per person in Maharashtra was, at that time, much lower than it was in the Sahel countries (Burkina Faso, Mauritania, Mali, Niger, Chad and Senegal), there was little starvation in Maharashtra. The Sahel, however, experienced widespread famine, because the shortage was not distributed so equally.

India has been able to avoid famine in recent years largely through such methods. Its last severe famine took place in 1943, four years before independence from the British. Although food supplies dropped drastically in 1967, 1973, 1979 and 1987 because of natural disasters,

severe famines were averted by recapturing the lost purchasing power of the threatened segments of the population.

Preventing famine through cash income programs differs from the standard practice of herding people into relief camps and trying to feed them. That approach, often used in Africa, tends to be slower and can put an unbearable organizational burden on government officials. Furthermore, packing people in camps away from home can disrupt normal economic operations, such as cultivation and animal husbandry, which, in turn, undermines future production. Such herding can also upset family life. Finally, and not least, the camps often become breeding grounds for infectious diseases.

In contrast, paying cash wages for public employment does not threaten the economic and social well-being of those being assisted. It builds on the existing production and market mechanisms and draws on the efficiency of traders and transporters. This approach can actually strengthen the economic infrastructure rather than weakening it.

Inevitably, beneficial fiscal policies are closely linked to politics. Although the public works approach relies on the market, it is not a free-market policy: it requires the government to intervene by offering employment. Public ownership of at least minimal stockpiles of food can also be helpful. The stores can give the government a credible threat in case traders attempt to manipulate the market. If merchants artificially withhold supplies in an effort to drive up prices, the government can retaliate by flooding the market to cause collapse of the prices and profits.

Famine is entirely avoidable if the government has the incentive to act in time. It is significant that no democratic country with a relatively free press has ever experienced a major famine (although some have managed prevention more efficiently than others). This generalization applies to poor democracies as well as to rich ones. A famine may wipe out millions of people, but it rarely reaches the rulers. If leaders must seek reelection and the press is free to report starvation and to criticize policies, then the rulers have an incentive to take preemptive action. In India, for instance, famine ceased with independence. A multiparty democratic system and a relatively unfettered press made it obligatory for the government to act. In contrast, even though postrevolutionary China has been much more successful than India in economic expansion and in health care, it has not been able to stave off famine. One occurred



SOMALIAN FAMINE VICTIM stands with an empty bucket, waiting for food. Local wars and the breakdown of law and order have disrupted the economy in Somalia, impoverishing many people. Earlier military dictatorships did little to prevent famines: as a result of the suppression of opposition parties and a muzzled press, these governments were free to be irresponsible.

between 1958 and 1961, after the agricultural program of the Great Leap Forward failed. The lack of political opposition and a free press allowed the disastrous policies to continue for three more years. The death toll consequently climbed to between 23 million and 30 million people.

Many countries in sub-Saharan Africa, among them Somalia, Ethiopia and Sudan, have paid a heavy price for military rule. Conflicts and wars are conducive to famine not only because they are economically destructive but also because they encourage dictatorship and censorship. Relatively democratic

sub-Saharan countries, such as Botswana and Zimbabwe, have, in general, been able to forestall famine. Of course, even an undemocratic poor country can avoid famine through luck: a crisis might not arise or some benevolent despot might implement effective famine-relief policies. But a democracy is a more effective guarantee of timely action.

Famine mortality data draw attention

to the failures of certain economic and political structures. Chronically high mortality rates reveal less extreme, but more persistent, failures. The economic policies associated with low infant mortality and increasing life expectancy vary considerably. Several countries that dramatically reduced infant mortality in the years between 1960 and 1985 experienced unprecedented rapid economic



Wealthy nations do not necessarily have greater life expectancies than do poor countries. For instance, Saudi Arabia is rich but has a lower life expectancy than the Indian state of Kerala. Through public outlays for education, health and nutrition, Kerala has extended life expectancy, despite a very low gross national product.

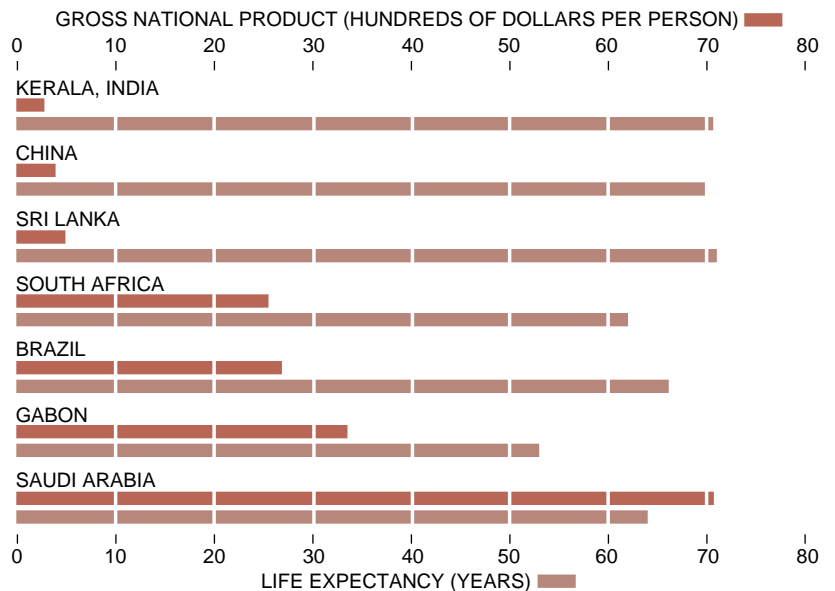


Mortality rates vary by race in the U.S. Black men between the ages of 35 and 54 are 1.8 times more likely to die than are white men of the same age. And black women in this group are almost three times more likely to die than are white women of the same age.

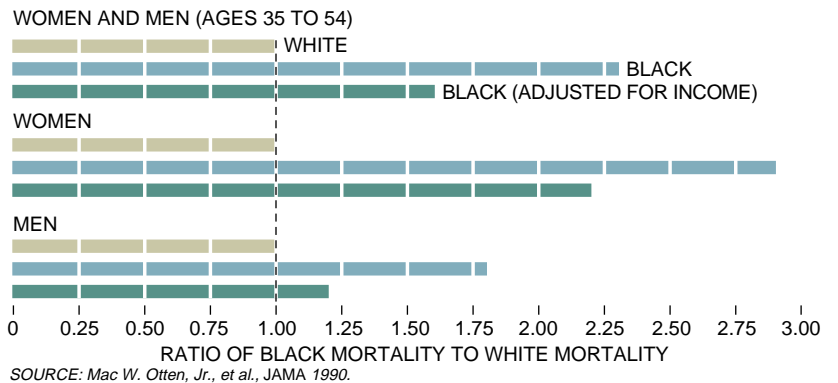


Life expectancy in England and Wales increased most dramatically in the decades of the two world wars largely because of the expansion of health care services and guaranteed food rationing for all citizens.

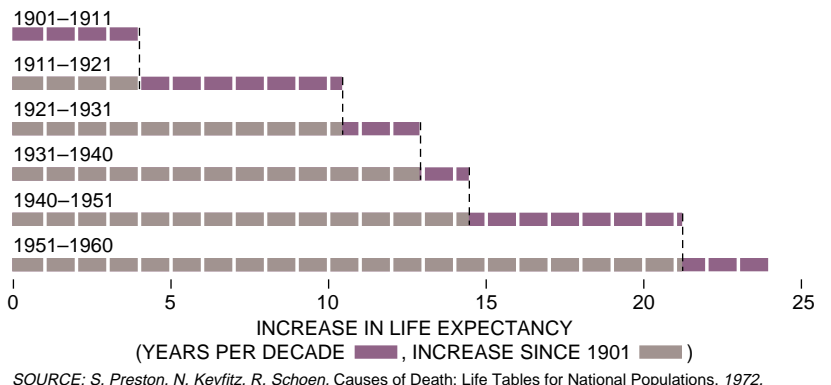
WEALTH AND LIFE EXPECTANCY IN CERTAIN COUNTRIES



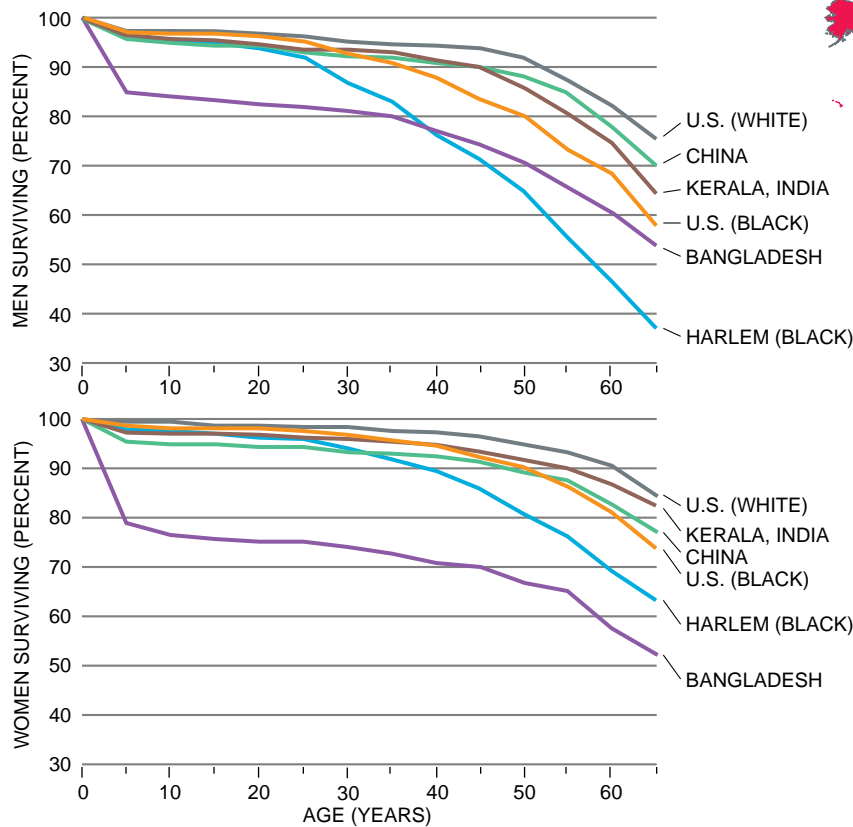
RACE AND DEATH RATES IN THE U.S.



LIFE EXPECTANCY IN ENGLAND AND WALES (1901-1960)



VARIATIONS IN SURVIVAL RATES, BY SEX AND REGION.



SOURCE: Data for Harlem and Bangladesh, Colin McCord and Harold Freeman, NEJM 1990; for others, official population statistics from the 1980s. Data are the most recent available.



The survival chances of the average African-American are better than those of an African-American living in Harlem but are unfavorable when compared with those of U.S. whites and those of the citizens of China and Kerala, who have much lower incomes. Although black women fare better than the men do, they too fall behind women in Kerala and China as they age.

growth. They include Hong Kong, Singapore and South Korea. These nations are now rich in terms of GNP. But also on the success list are several nations that are still poor: China, Jamaica and Costa Rica, among others.

The fact that a poor country can achieve improvements in health care and life expectancy that, in many ways, rival those of wealthier nations has tremendous policy implications. This ability challenges the often-aided opinion that a developing country cannot afford expenditures for health care and education until it is richer and more financially sound. This view ignores relative cost. Education and health care are labor intensive, as are many of the most effective medical services. These services cost much less in a cheap labor economy than they do in a wealthier country. So, although a poor country has less to spend on these services, it also needs to spend less on them.

The long-standing efforts of Sri Lanka and the Indian state of Kerala (whose population of 29 million is bigger than Canada's) illustrate the merits of public spending for education and health. Sri Lanka promoted literacy and schooling programs early in this century. It massively expanded

medical services in the 1940s, and in 1942 it started distributing free or subsidized rice to bolster the nutritional intake of undernourished people. In 1940 the death rate was 20.6 per 1,000; by 1960 it had fallen to 8.6 per 1,000.

Similar changes took place in Kerala. Despite a per capita GNP that is considerably less than the Indian average, life expectancy in Kerala now is more than 70 years. Such an accomplishment in the face of very low income and poverty is the result of the expansion of public education, social epidemiological care, personal medical services and subsidized nutrition.

This analysis does not contradict the valuable contribution that an increasing GNP can make to raising life expectancy. Clearly, economic soundness can help a family obtain better nutrition and medical care. Furthermore, economic growth can augment the government's ability to provide for public education, health care and nutrition. But the results of economic growth are not always channeled toward such programs. Many nations—such as Saudi Arabia, Gabon, Brazil and South Africa—have much worse records on education, health and welfare than do other countries (or states) that have much lower GNPs but more public-oriented policy, Sri Lanka, China, Costa

Rica and Kerala, among them. The crucial point is that poor countries need not wait to get rich before they can combat mortality and raise life expectancy.

The role of public policy in lengthening life expectancy is, of course, not peculiar to the Third World alone. Public intervention in health, education and nutrition has historically played a substantial part in the rise in longevity in the West and in Japan. In England and Wales, the decades of World War I and World War II were characterized by the most significant increase in life expectancy found in any decade this century. War efforts and rationing led to a more equitable distribution of food, and the government paid more attention to health care—even the National Health Service was set up in the 1940s. In fact, these two decades had the slowest growth of gross domestic product per capita: indeed, between 1911 and 1921, growth of GDP was negative. Public effort rather than personal income was the key to increasing life expectancy during those decades.

Analyzing mortality data can help in the economic evaluation of social arrangements and of public policy. This perspective can be particularly useful in elucidating crucial aspects of social inequality and poverty and in identifying policies that can counter them. One of the more immediate problems that must be faced in the U.S. is the need for a fuller understanding of the nature of economic deprivation. Income is obviously a major issue in characterizing poverty, but the discussion of American poverty in general and of African-American poverty in particular has frequently missed important dimensions because of an overconcentration on income.

As has often been noted, two fifths of the residents of New York City's cen-

tral Harlem live in families whose income levels lie below the national poverty line. This fact is shocking, but that poverty line, low though it is in the U.S. context, is many times the average income of, say, a family in Bangladesh—even after correcting for differences in prices and purchasing power. In some ways, a more telling view of poverty in Harlem as compared with that in Bangladesh can be found in mortality statistics. Colin McCord and Harold P. Freeman of Columbia University and Harlem Hospital have already noted that black men in Harlem are less likely to reach the age of 65 than are men in Bangladesh. In fact, Harlem men fall behind Bangladeshi men in terms of survival rates by the age of 40.

These comparisons can be enhanced by scrutinizing the situations in China and Kerala, poor economies that have undertaken much more thorough efforts in public health and education than has Bangladesh. Even though China and Kerala have higher infant mortality rates, their survival rates for teenage and older males are better than Harlem's. The higher mortality of men in Harlem partly reflects deaths caused by violence. Violence is a significant part of social deprivation in the U.S., even though it is not the only cause of the high mortality in Harlem. Women in Harlem fall behind Chinese and Keralan women in survival rates by the ages of 35 and 30, respectively.

Moreover, a similar problem plagues African-Americans in general. Again, black people in the U.S. have lower infant mortality rates than the populations of China and Kerala. But as we move up the age scale, black women and men fall behind the women and men of Kerala and China, in terms of the percent surviving. The nature and extent of the deprivation among Afri-

can-Americans cannot be adequately understood when they are measured by the yardstick of income. According to that scale, African-Americans are poor in comparison with U.S. whites, but they are immensely richer than Chinese and Keralan citizens. On the other hand, in terms of life and death, African-Americans are less likely to survive to a ripe old age than are people in some of the poorest Third World countries.

Another feature of racial inequality revealed by the mortality data is the relative deprivation of African-American women. In some ways, they fare worse than black men. The gaps between white and black mortality for the ages between 35 and 54 years appears to be much wider for black women than for black men. The differences between blacks and whites relate partly to differences in their incomes. But even after correcting for variations in incomes, some of the discrepancy remains. For black women the bulk of the mortality differences cannot be attributed to income gaps at all.

Mortality information can also be used to investigate an elementary manifestation of sexual bias. One striking demographic feature of the modern world is the enormous geographic variation in the ratio of females to males. Medical evidence suggests that, given similar care, women tend to have lower mortality than do men. Even in the uterus, female fetuses

are less prone to miscarriage. Although males outnumber females at birth and at conception, women outnumber men in Europe and North America by about 5 percent.

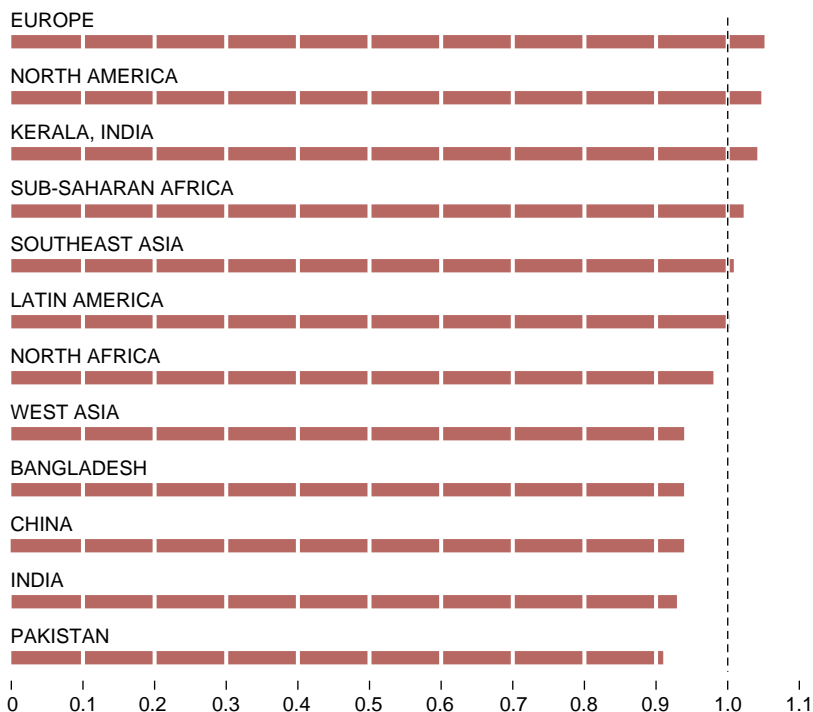
In many parts of the developing world, however, the ratios of females to males are quite different: whereas that ratio is 1.02 in sub-Saharan Africa, it is 0.98 in North Africa, 0.94 in China, Bangladesh and West Asia, 0.93 in India and 0.91 in Pakistan. To form an idea of the magnitudes involved, it is useful to ask such questions as: If countries such as China had the female-male ratio that, say, sub-Saharan Africa has, how many more women would there be? If we do use the sub-Saharan African ratio as the benchmark, as Jean Drèze of the Delhi School of Economics and I did, then it would appear that more than 100 million women were "missing" in the female-deficit countries: 44 million missing in China alone, 37 million in India. Other estimates, using other benchmarks, have placed the number between 60 million and 90 million.

The phenomenon of the missing women reflects a history of higher mortality for females and a staunch antifemale bias in health care and nutrition in these countries. Jocelyn Kynch of the University of Oxford and I examined hospital records in Bombay. We found that women had to be more seriously ill than men did in order to be taken to the hospital. Another study I conducted, with Sunil Sengupta of Visva-Bharati



More males are born than females, but females have lower mortality: thus, they tend to outnumber males if both sexes receive similar health care. In Europe and North America, the ratio of women to men is about 1.05, although this number is inflated because of the loss of men in past wars. In some other countries, women have not had equal access to health care.

RATIO OF WOMEN TO MEN, BY REGION



University, indicated systematic bias in nutritional health care in favor of boys in two West Bengal villages.

Although historical and cultural factors lie behind this bias, economic institutions are involved as well. Evidence suggests that the ability of women to earn an income and to enter occupations, especially in more skilled jobs, outside the home, enhances their social standing and in turn influences the care they receive within the family. Working outside the home also gives women exposure to the world and, sometimes, more of an opportunity to question the justice of the prevailing social and economic order. Literacy, education, land ownership and inheritance can also improve the overall status of women.

In Kerala, economics has helped better the position of women. Not only does the state have a large proportion of working women in occupations that command respect, but, as described earlier, it has a well-developed system of education, with high literacy rates for both sexes, a widespread network of health services and, for a substantial and influential segment of the population, a tradition of matrilineal inheritance. The female-male ratio of the population is now about 1.04 (although it would be reduced by a little if one took into account men working outside the state). Life expectancy in Kerala at birth is 73.0 years for females, 67.5 years for males.

That average life expectancy is nearly matched by China, but women fare relatively better in Kerala. The Chinese government has strived to eradicate sexual inequality, and China does have a high rate of female employment. The level of female literacy is, however, much lower than that in Kerala. The high female infant mortality in China may also be partly connected with the impact of compulsory birth control measures—the partial imposition of the so-called one-child policy—in a society in which male preference is overriding.

This article is not directly concerned with fertility and family planning, but I would like to note that compulsory birth control does have some dangers with regard to sexual bias. There are excellent arguments, based on considerations of liberty and freedom, against such compulsion in the first place. But the possible effect of such a measure on female mortality adds another dimension to the debate. Chinese success in slowing the birth rate is often cited in discussions about the need for forceful family planning in the Third World. It is true that the Chinese birth rate of



AFRICAN-AMERICANS who live in inner-city environments similar to the one portrayed in this photograph have less favorable chances for survival than do the citizens of Kerala. This discrepancy highlights the failure of U.S. policies to make equitable arrangements for public education, health care, nutrition and social peace.

21 per 1,000 compares very favorably with India's 30 per 1,000 (and the average of 38 per 1,000 seen in low-income countries other than China and India). Yet Kerala's birth rate of 20 per 1,000 is comparable to China's of 21 per 1,000—without any compulsory birth control policy and without the problem of female infant mortality.

Considerable demographic evidence indicates that declines in birth rates quite often follow declines in death rates. This pattern relates to a decreasing urgency to have many children to ensure survivors. It also reflects the interdependence between birth control and death control: providing people with access to contraception can be effectively combined with the delivery of medical care. As the death rate has fall-

en in Kerala, so has the birth rate: from 44 per 1,000 between 1951 and 1961 to 20 per 1,000 between 1988 and 1990.

Mortality data provide a gauge of economic deprivation that goes well beyond the conventional focus on income and financial means. The assessment of economic achievement in terms of life and death can draw attention to pressing questions of political economy. This perspective can help in providing a fuller understanding of famine, health care and sexual inequality, as well as poverty and racial inequality, even in wealthy nations such as the U.S. The need to widen the scope of conventional economics to include the economics of life and death is no less acute in the U.S. than it is in famine-stricken sub-Saharan Africa.

FURTHER READING

POVERTY AND FAMINES: AN ESSAY ON ENTITLEMENT AND DEPRIVATION. Amartya Sen. Oxford University Press, 1981.
ROUTES TO LOW MORTALITY IN POOR COUNTRIES. John C. Caldwell in *Population and Development Review*, Vol. 12, No. 2, pages 171-220; June 1986.
HUNGER AND PUBLIC ACTION. Jean Drèze and Amartya Sen. Oxford University Press, 1989.
THE EFFECT OF KNOWN RISK FACTORS ON THE EXCESS MORTALITY OF BLACK ADULTS

IN THE UNITED STATES. Mac W. Otten, Jr., Steven M. Teutsch, David F. Williamson and James S. Marks in *Journal of the American Medical Association*, Vol. 263, No. 6, pages 845-850; February 9, 1990.
INEQUALITY REEXAMINED. Amartya Sen. Harvard University Press, 1992.
HUMAN DEVELOPMENT IN POOR COUNTRIES: ON THE ROLE OF PRIVATE INCOMES AND PUBLIC SERVICES. Sudhir Anand and Martin Ravallion in *Journal of Economic Perspectives*, Vol. 7, No. 1, pages 133-150; Winter 1993.

The Core-Mantle Boundary

This interactive zone may be the most dynamic part of the planet, directly affecting the earth's rotation and magnetic field

by Raymond Jeanloz and Thorne Lay

About 2,900 kilometers away—less than three days' drive, if that were possible—lies the most dramatic structure of the earth. Largely ignored in past research, the remote region between the lowermost mantle and the upper core is proving to be crucial in understanding the chemical and thermal evolution of the planet. No longer regarded as simply a contact delineating the liquid-iron outer core from the rocky mantle, the core-mantle region may actually be the most geologically active zone of the earth. Its features seem to have changed immensely during the earth's history, and its physical properties now evident vary from place to place near the bottom surface of the mantle. In fact, the physical changes across the interface between the core and mantle are more pronounced than are those across the planetary surface separating air and rock.

The strong heterogeneity of the core-mantle boundary region is thought to influence many global-scale geologic processes [see "The Earth's Mantle," by D. P. McKenzie; *SCIENTIFIC AMERICAN*, September 1983]. The dynamics of the zone affect the slight wobbling of the earth's axis of rotation and characteristics of the geomagnetic field. Variations in the core-mantle region also modu-

late the convection in the earth's mantle, which is responsible for the movement of continents and tectonic plates.

The first hint that something unusual was going on at the depth where the core and mantle meet came in the mid-1930s. Vibrations generated by earthquakes provided the clue. Throughout most of the mantle, the speed of seismic waves increases as a function of depth. Furthermore, lateral variations in seismic-wave velocity are only minor. One can interpret these characteristics as meaning that the earth gets "simpler" with respect to depth, that is, the composition and structure of the planet become more uniform. In contrast, the great diversity of geologic structures and rocks observed underfoot reveal the surface to be the most complicated region.

Yet the velocity behavior of seismic waves holds only to a certain point. At the lowermost few hundred kilometers of the mantle, just before the core begins, the average speed of seismic waves does not increase appreciably, and more meaningful changes in velocity appear from region to region [see *illustration on pages 50 and 51*]. The effect is subtle, amounting to only a few percent difference. Yet by geologic standards, these few percent represent enormous variations in structure or temperature, or both. Early workers recognized the significance of the changes from the simple behavior in the overlying lower mantle and consequently named this region, which was deduced to be about 200 to 400 kilometers thick, the D'' layer.

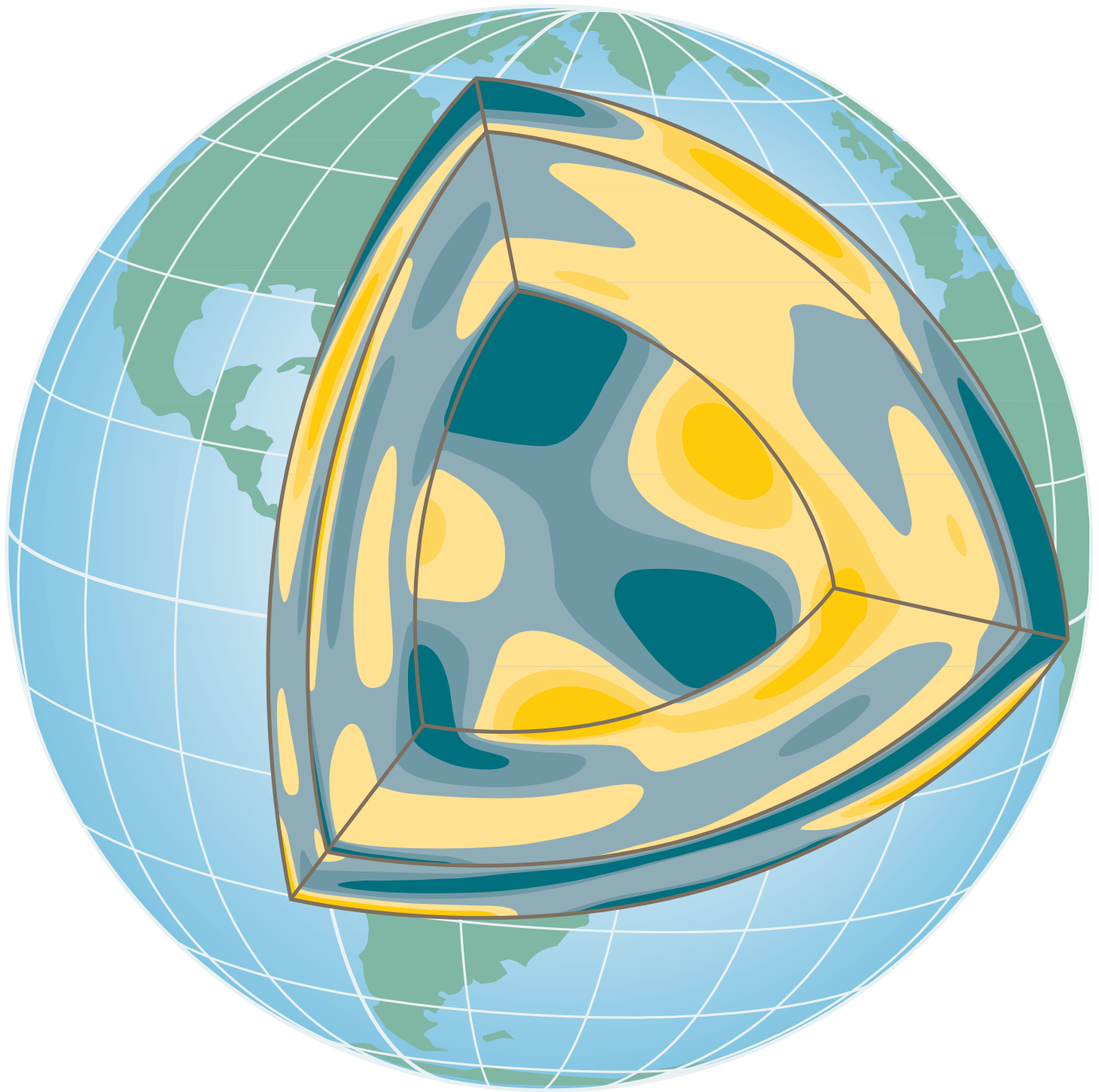
The origin of the layer's name (pronounced "dee double prime") is more historic than poetic. Early geologists had labeled the parts of the deep earth with letters of the alphabet, rather than as crust, mantle and core. This form of identification, however, meant that any intervening layer subsequently discovered had to incorporate a "prime" symbol to distinguish it. Although other layers were eventually renamed, the D'' nomenclature has endured.

Investigators proposed numerous interpretations to account for the seismic properties of the D'' layer. Unfortunately, there were too many possible explanations and too little information to permit a definitive characterization of the layer. Better descriptions of the D'' layer had to wait until the technological breakthroughs of the 1980s. Then, using arrays of recording instruments deployed around the world, seismologists could for the first time collect and process enough data to derive three-dimensional images of the earth's interior [see "Seismic Tomography," by Don L. Anderson and Adam M. Dziewonski; *SCIENTIFIC AMERICAN*, October 1984]. The seismometers used primarily operate in the range between about one and 0.0003 hertz, or cycles per second. (These acoustic frequencies are far below the range of human hearing, which extends from about 20 to 20,000 hertz.) Seismic tomography is often compared to computed tomographic scans used in medicine. But because it relies on sound waves, seismic tomography is more akin to the ultrasonic imaging done during pregnancy. The main drawback is its resolution: images of features smaller than 2,000 kilometers tend to be smeared out.

Nevertheless, seismic tomography helped to quantify the properties of the D'' layer. It showed that the region differs drastically from the overlying mantle. The fact that the velocity of seismic waves is affected over continent-size areas shows that large-scale structures dominate D''. Still, seismic tomography could not explain the causes of this variability in physical properties. Could large, chemically distinct structures exist at the bottom of the mantle, just as continents mark the seismic heterogeneity of the earth's surface? Or are the heterogeneities simply large-scale temperature differences at the base of the mantle?

To answer these questions, one of us (Lay) began in the early 1980s to implement a new method to explore the core-

RAYMOND JEANLOZ and THORNE LAY study the physics of the deep earth. Jeanloz, professor of geology and geophysics at the University of California, Berkeley, received his Ph.D. in 1979 from the California Institute of Technology. A MacArthur Fellow, Jeanloz also studies the internal evolution of other terrestrial planets and the formation of new types of glass that have novel properties. Lay is professor of earth sciences at the University of California, Santa Cruz, where he is also director of the Institute of Tectonics. His specialty is the study of earthquakes and the structure of the earth's interior. A recipient of the American Geophysical Union's 1991 Macelwane Medal, Lay earned his Ph.D. in 1983 from Caltech.



SEISMIC-WAVE VELOCITIES differ throughout the earth's interior, as depicted in this image generated by seismic tomography. In some regions the waves move more quickly than is average for that depth (*blues*); in others the waves are slower

(*yellows*). Such variations can suggest differences in composition. Much of the complexity of the core-mantle boundary (the exposed surface of the outer core) is not evident, because the heterogeneities are too small to be resolved.

mantle boundary. The idea was to use computer calculations to analyze all the characteristics of the observed seismic wave front, not just the wave velocity, as in the case of tomography. Such waveform analysis is a powerful approach because the technique can resolve structures as small as a few tens of kilometers across instead of those 2,000 kilo-

meters or more in size. The disadvantage is that one can look only at limited parts of the core-mantle boundary. There are not enough earthquakes or other sources of seismic energy to obtain a global picture at such a high level of detail.

The waveform studies suggest that neighboring regions within the D'' lay-

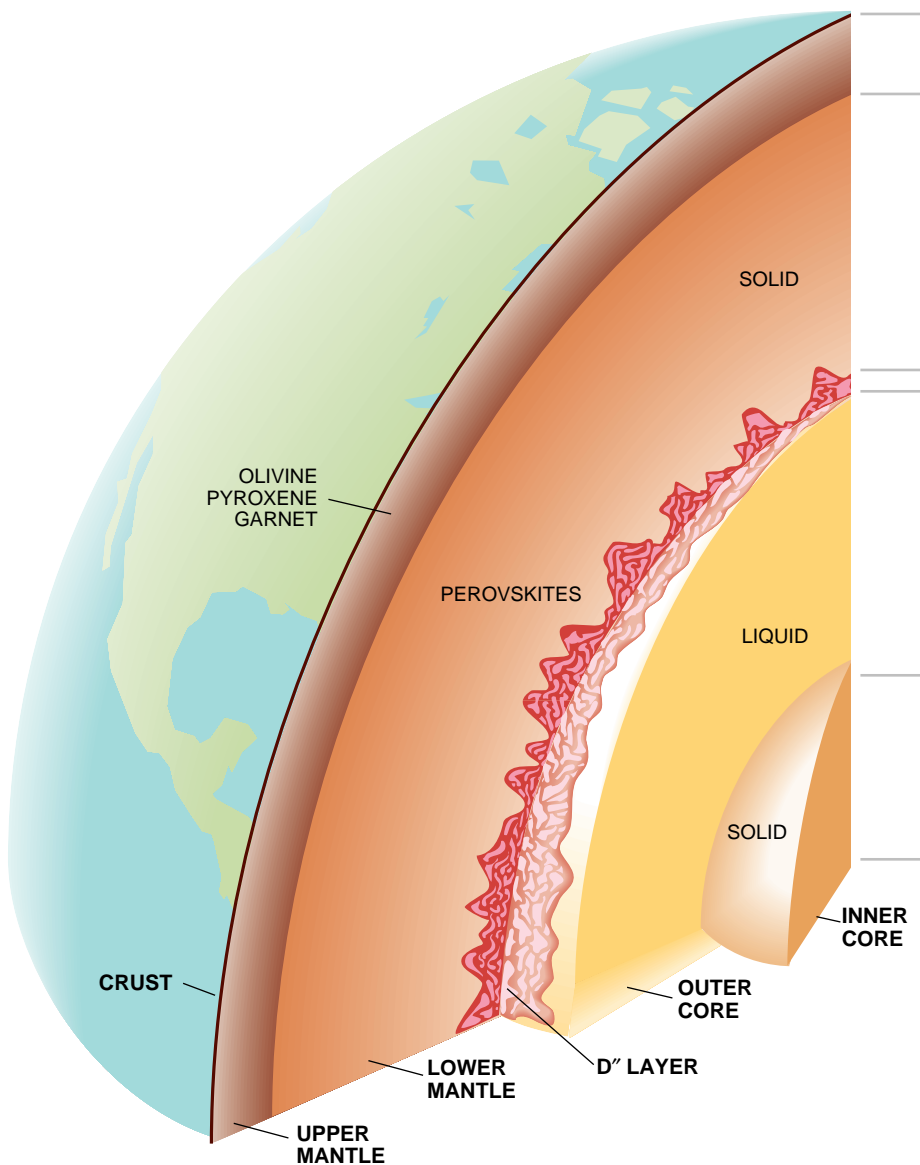
er can be more distinct than had once been thought. For example, several research groups studying the core-mantle boundary below northern Siberia found that acoustic velocities vary so radically over short distances that closely spaced seismometers systematically recorded different waveforms. The finding can best be explained by assuming that the

heterogeneity in seismic velocities is large in magnitude and occurs over distances smaller than can be resolved, that is, within a few tens of kilometers. Waveform studies can also map the differences in thickness of the D' layer. In many places the top of the D' layer causes an abrupt increase in wave velocity, a process that reflects seismic energy. The reflections have revealed that the thickness of the D' layer varies dramatically. The layer can be so thin as to be undetectable, or it can span as many as 300 kilometers.

Stanley M. Flatté's group at the University of California at Santa Cruz helped to confirm the great variability of the D' layer. During the mid- to late 1980s, he and his colleagues began to apply new methods of wave analysis to the signals obtained from seismic waves that have been scattered in the deep mantle. Their method relies on a statistical description of how waves propagate through a strongly scattering substance. Such material would be analogous to fog or clouds. Flatté's approach is to observe how the wave front from an earthquake changes shape after traveling through the D' region. An earthquake initially sends out a smooth, spherically expanding wave. But as that wave is refracted and scattered by variations in seismic features, such as the strong heterogeneities near the core-mantle boundary, the front no longer remains smooth. It becomes rippled, or corrugated [see illustration on page 53].

The trick in measuring the degree of wave-front corrugation is a dense array of seismometers. Taking observations from one such collection located in Norway, Flatté has shown that the D' region appears quite murky to seismic waves. It must contain heterogeneous features as small as 10 kilometers in length. The seismological observations thus indicate that the D' region is a heterogeneous layer that laterally varies in thickness.

In contrast to the murkiness of the D' layer, the core-mantle boundary (on which the D' layer rests) appears smooth and sharp. Last year John E. Vidale and Harley Benz of the U.S. Geological Survey beautifully demonstrated the abruptness of the interface. They used a vast number of seismic recording stations that had been deployed across the western U.S. The array of seismometers generally monitors regional earthquake activity, but Vidale and Benz have employed it to find seismic waves that have bounced off the core-mantle boundary. Remarkably, seismic waves arrived coherently across more than 900 stations in the array. This coherence implies that the core-mantle boundary



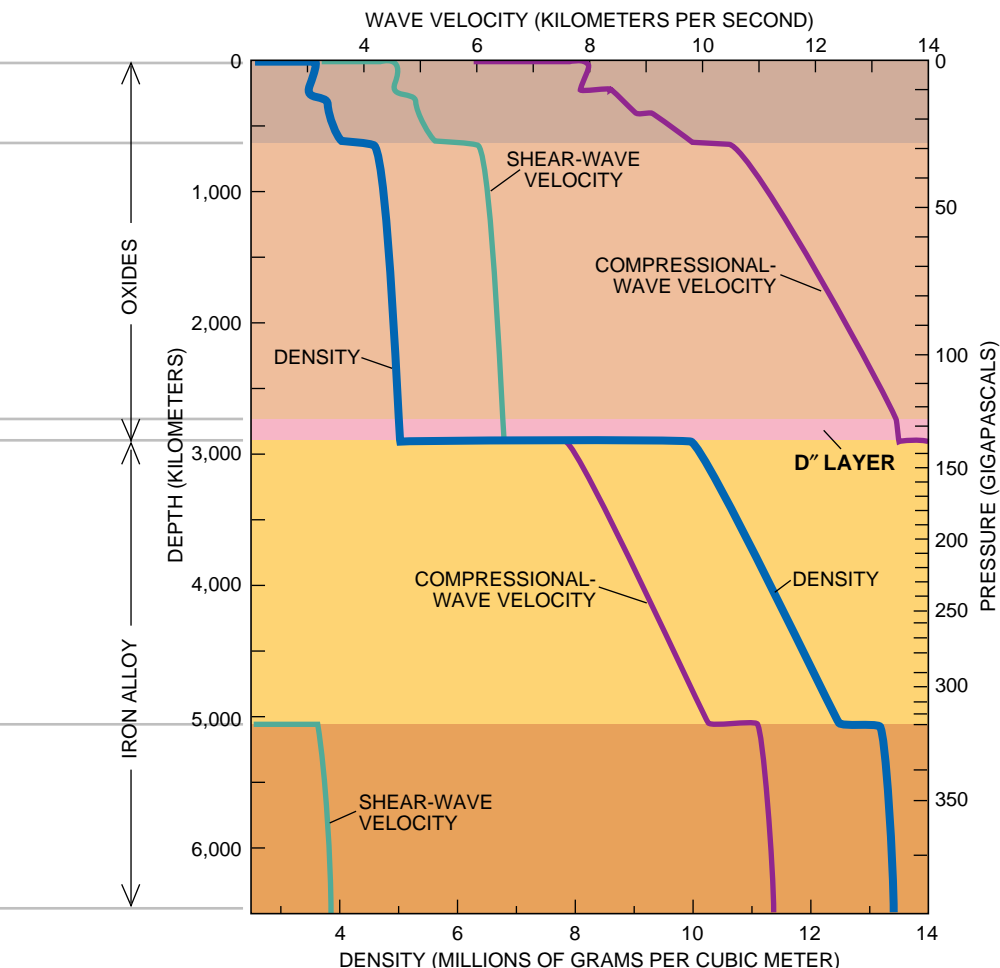
represents a sharp transition from the mantle to the core, at least for the area measured. The sudden transition reflects as much as 50 percent of the seismic waves and transmits the remainder. Analyses of the reflected and transmitted waves show that the boundary varies in depth by no more than a few kilometers.

Seismic-wave studies have done much to elucidate the D' layer and the core-mantle boundary. But the inaccessibility of the regions has prevented geophysicists from understanding completely how such complicated structures came about.

If seismic studies cannot thoroughly breach the remoteness of the deep earth, why not bring the core and mantle to the surface? That is precisely the approach taken by many research-

ers, including one of us (Jeanloz). Specifically, we sought to duplicate the high pressure and temperature existing in the deep mantle and core. A breakthrough in engineering made such a feat possible: investigators had learned to compress minuscule samples between the points of two diamonds and to heat the specimen using a high-powered laser beam [see "The Diamond-Anvil High-Pressure Cell," by A. Jayaraman; *SCIENTIFIC AMERICAN*, April 1984]. By 1986 the diamond cells could generate pressures greater than those at the center of the earth.

Diamond's hardness is not the only reason for using the substance as an anvil. The utility of diamond also lies in its transparency. A laser beam can be focused directly through the diamond to heat the sample to thousands of degrees Celsius. Moreover, one can ob-



CROSS SECTION OF EARTH shows the planet's primary regions (*opposite page*). The crust and mantle consist of oxide crystals such as olivine, pyroxene and garnet in the upper mantle and silicate perovskite in the lower mantle. The core is an iron alloy, liquid in the outer part and solid in the center. The layers correspond to the observed variations in density and velocity of seismic waves as they travel through the earth (*above*). Both density and wave velocity increase as a function of depth except at the D' layer. Note that seismic energy can propagate as shear waves (waves that oscillate at right angles to the direction of motion) and as compressional waves (waves that move back and forth in the travel direction). Because liquids do not have rigidity, shear waves cannot propagate in the outer core. Shear-wave motions reappear in the inner core because a fraction of the compressional waves transforms into shear waves at the liquid-solid interface.

serve the specimen while it is at super-high pressures and temperatures. One determines the temperature of the sample by measuring the thermal radiation the sample emits through the diamond. In this way, one can quantify how "red hot" or "white hot" the material has become; astronomers infer the surface temperatures of stars by color in the same manner. Using the laser-heated diamond cell, we can simulate the appropriate temperatures and pressures at the core-mantle boundary. We wanted to see what would happen when we placed matter that constitutes the outer core in contact with minerals of the lowermost mantle.

Of course, we needed to know what

materials make up the mantle and core before squeezing them together. To determine the mantle constituents, Elise Knittle, working with Jeanloz, followed up on research by groups at the Australian National University, the Carnegie Institution of Washington and elsewhere. We relied on prior experimental work, on theoretical models and on the fact that the pressure in the lower mantle exceeds 20 gigapascals (20,000 atmospheres).

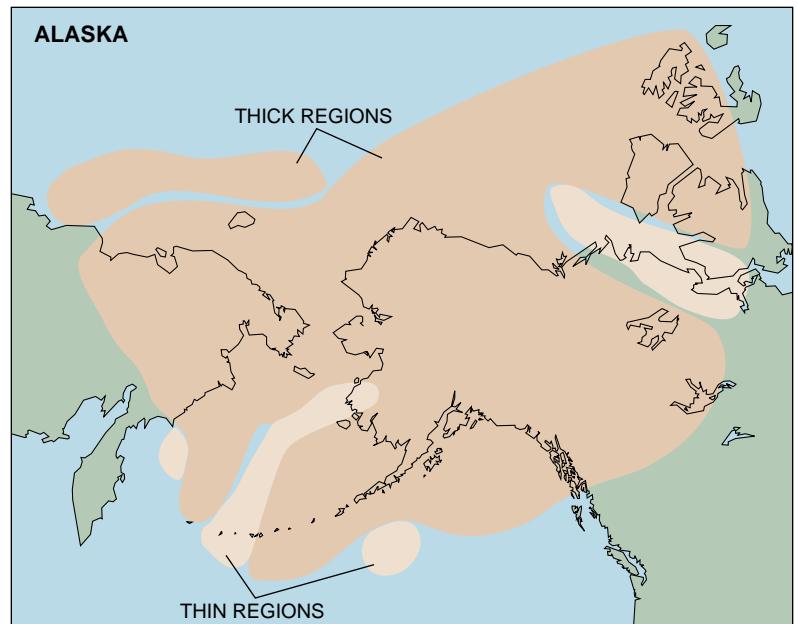
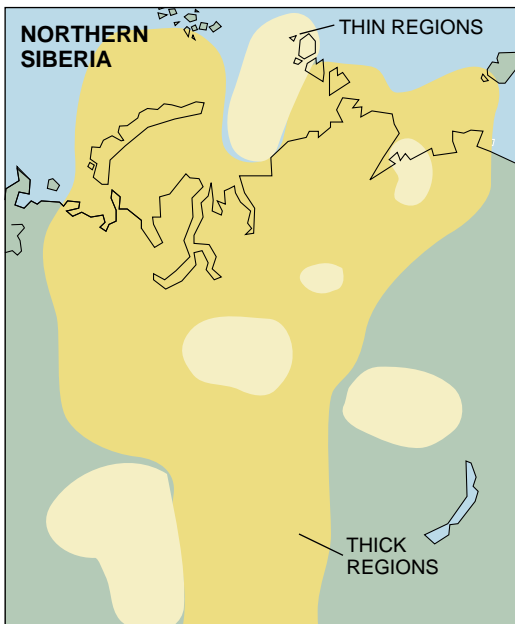
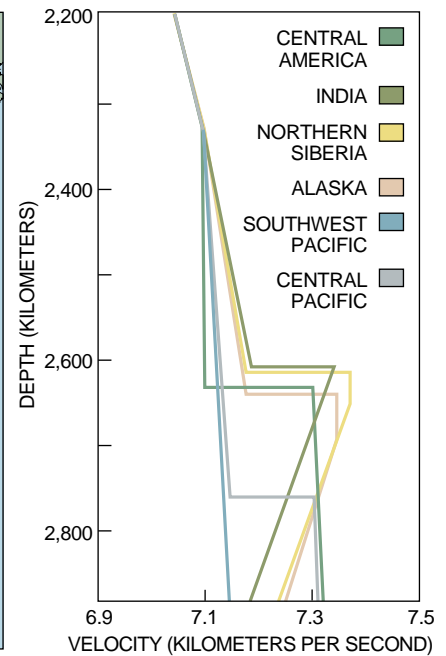
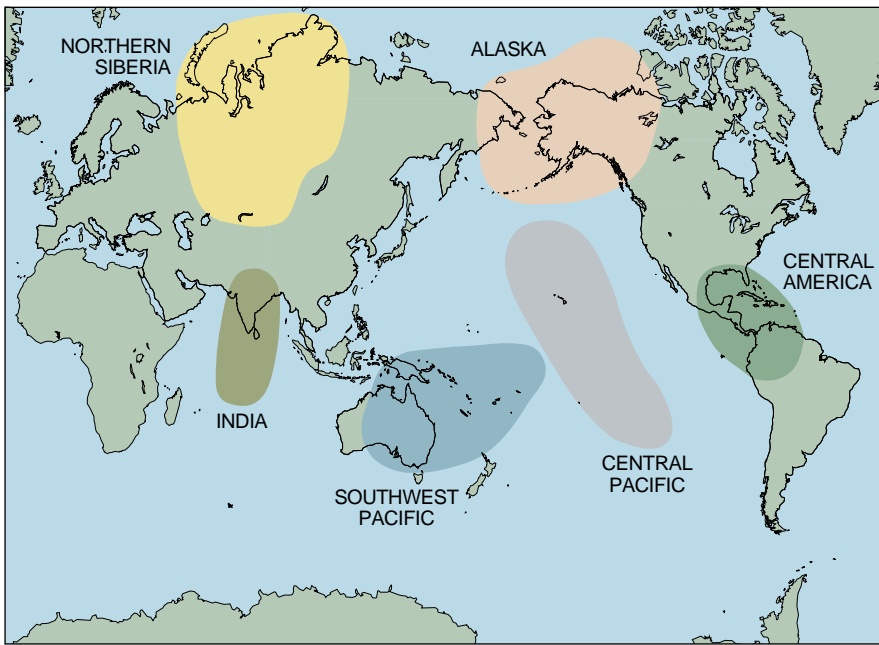
From that information, we deduced that a single high-pressure mineral phase must dominate the lowermost mantle. This mineral is a dense form of iron magnesium silicate, or (Mg,Fe) SiO₃, a robust and chemically simple

compound that can be formed only under pressures above 20 gigapascals. Because it has the same crystalline structure as the mineral perovskite (CaTiO₃), it is consequently called magnesium silicate perovskite. The lower mantle rock probably also contains minor amounts of magnesiowüstite—a combination of magnesium oxide (MgO) and wüstite (FeO). This composition is quite unlike the nature of rocks at or near the earth's surface. Such rocks are composed of many different, complex minerals that react chemically and transform into new minerals under modest changes of pressure or temperature. The deduced chemical simplicity of the deep mantle accords well with the data derived from seismic waves, which show it to be relatively devoid of structure (except for the D' layer). This consistency gives us confidence that we are examining the appropriate minerals in our laboratory simulations.

Determining the constituent of the core was more straightforward. Seismological studies done more than 50 years ago enabled geophysicists to infer its structure. The core consists of a molten substance surrounding a solid center. The fluid is acknowledged to be a metal—specifically, an alloy of iron. In fact, the churning of the molten iron generates the earth's magnetic field.

Having established the compounds involved, Knittle carried out a series of experiments in which liquid iron was put in contact with crystalline silicate perovskite at high pressures. She found that the perovskite reacts vigorously with liquid iron, even if these substances touch for just a few seconds. The nature of the chemical reaction is quite interesting and unexpected. The products are a mixture of electrically insulating oxide minerals—magnesium silicate perovskite and stishovite (SiO₂)—and metallic alloys—iron silicide (FeSi) plus wüstite. Wüstite had not been known to be able to form a metallic alloy at any temperature or pressure. Qualitatively speaking, wüstite can react this way because its oxygen atom at high pressures takes on the chemical attributes normally ascribed to its neighbor in the periodic table, sulfur. Metallic sulfides such as iron disulfide (pyrite, or fool's gold) are of course well known.

The experiments also showed that liquid iron begins to react with mantle substances at pressures of 20 to 30 gigapascals. Such pressures are far less than those at the core-mantle boundary (136 gigapascals). Therefore, the reactions have probably persisted since the earliest history of the planet—that



SHEAR-WAVE VELOCITY in the D' layer changes across the earth, as indicated by the six regions (colored areas, top left) that have been most intensely studied. The corresponding velocity distribution as a function of depth (top right) shows that each region exhibits a discontinuity at the D' layer. The unique-

ness of each velocity signature implies that D' varies over the entire globe. The expanded maps (bottom) for areas below northern Siberia and Alaska summarize the heterogeneity of D', showing the intermingling of thick regions (dark patches) with parts so thin as to be seismically invisible (light patches).

is, when the earth was developing and the core might have been forming at pressures below 136 gigapascals. Such chemical reactions are likely to have significantly altered the core-mantle system. A considerable amount of oxygen has probably been drawn into, or alloyed with, the core metal over geologic history. In essence, the lower mantle rock has been and still is slowly dissolving into the liquid metal of the outer core. Berni J. Alder of Lawrence Livermore National Laboratory made this suggestion more than 25 years ago. Our

experiments substantiate his conjecture. Indeed, one of the remarkable consequences of this hypothesis is that it offers a simple explanation for why the properties of the core are nearly but not exactly those of iron at the equivalent pressure and temperature. Most notably, the density of the outer core is about 10 percent lower than that of pure iron [see "The Earth's Core," by Raymond Jeanloz; SCIENTIFIC AMERICAN, September 1983]. But as indicated by Alder's hypothesis and our diamond-cell experiments, the core cannot

be completely iron. A purely iron core would have become tainted by reaction with the overlying rock over geologic time. Quite plausibly, the core was never pure iron. Instead it probably contained some nickel, sulfur and other minor constituents. Iron-rich meteorites provide the basis for this hypothesis. Such meteorites, considered partial remnants of the materials from which the earth formed, harbor many similar contaminants. Like pure iron, these iron-rich alloys can react chemically with rocky compounds at high pressures and tem-

peratures, forming an alloy with oxygen.

According to our experiments, the dense liquid of the outer core must seep into the rock, probably by capillary action. The molten metal would penetrate along the boundaries between the mineral grains at the bottom of the mantle. Estimates of the capillary forces involved suggest that the core liquid could move upward some tens to hundreds of meters above the core-mantle boundary. The reaction between core liquid and mantle rock probably takes place in less than a million years—instantaneously, in geologic terms.

The liquid, however, does not necessarily always have to move upward and to work against gravity. The interface between the mantle and core is not likely to be perfectly flat. Metallic liquid would permeate laterally and downward into the mantle rock from regions where the core-mantle boundary is elevated. Measurements from geodetic and seismological studies indicate that the topography of the core-mantle boundary deviates from absolute flatness by hundreds of meters to a few kilometers. Therefore, the zone of permeation and direct chemical reaction between the core liquid and mantle rock is no more than hundreds to at most thousands of meters thick. The size estimate explains why studies of seismic waves do not currently detect signs of reaction at the core-mantle boundary. The thickness of the reaction zone is less than typical seismic wavelengths. In addition, no more than a modest fraction of the reaction zone consists of liquid at any given moment. Thus, the presence of a small amount of liquid would not noticeably alter the velocity of seismic waves in the lowermost mantle.

How do these chemical reactions at the core-mantle boundary account for the observed characteristics of the D'' layer? The answer lies in a complex and indirect process resulting from forces that act on the core-mantle interface. The forces come from the thermal energy of the underlying core, which heats the rock at the base of the mantle. As a result, the heated part of the mantle moves upward over a period

of tens to hundreds of millions of years—far longer than the reaction between the core and mantle, which takes place in less than one million years. The convection must disrupt the reaction zone at the core-mantle boundary, entraining it upward and exposing fresh mantle rock to the corrosive liquid of the core. The convection is the same force that causes the tectonic plates to move at the earth's surface.

Mantle convection does not entrain liquids very far; any liquid metal that might be present in the boundary probably flows out, spongelike, through porous rock before moving upward. On the other hand, the iron-rich crystalline products from the reaction zone, such as wüstite, are readily incorporated into the mantle flow. The slow convection of the mantle pulls up the crystalline alloy a modest distance before the den-

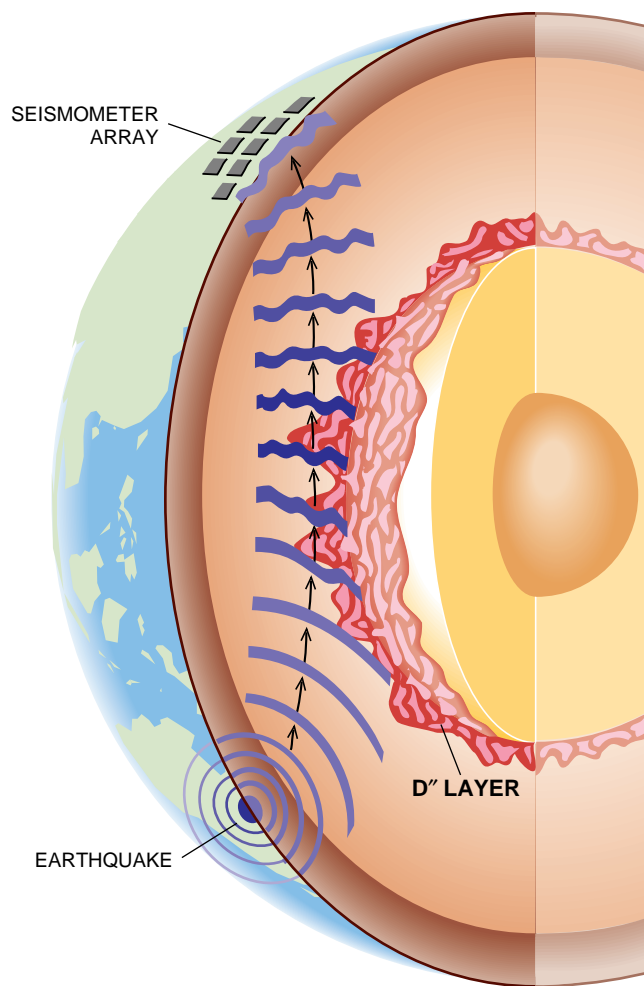
sity of the metallic solids causes them to sink back toward the bottom. These solids essentially resemble the dregs of spice that remain at the bottom of a pot of mulled wine.

As a result, the alloy-rich substances would tend to pile up on the bottom of the mantle, especially near regions of upwelling, much as snowdrifts form in a blizzard. The upward dispersal abets infiltration of material from the core and builds a thicker zone of intermixing; the intermixing of reaction products and unreacted mantle causes the seismic heterogeneity. In contrast, downwelling regions would disperse the dregs and thus tend to thin the D'' layer and to depress the core-mantle boundary. Modeling by Louise Kellogg of the University of California at Davis and Norman H. Sleep of Stanford University and others suggests that the metallic alloys in

local regions of the reaction zone may be swept upward several hundred kilometers into the mantle. The process would require tens of millions of years.

The buildup of the alloy-rich drifts at the bottom of the mantle solves an important mystery. Specifically, the drifts would explain the variation in thickness of the D'' layer observed by seismologists. Moreover, calculations indicate that the height of the alloy drift swept up in the mantle is comparable to the thickest parts of D''. Given the billions of years for progressive accumulation of the metallic dregs, it is plausible that much of the complexity and many of the variations in thickness of D'' result from the way mantle flow modulates the alloy-rich reaction layer. The flow may have also caught in its wake other dense mantle material or products from the core. We suspect that reaction dregs can collect, albeit to a lesser extent, on the inner side of the core-mantle boundary. A thinner version of the D'' layer probably exists there, just inside the liquid outer core.

In view of the intense dynamics taking place 2,900 kilometers below the earth's surface, it should not be surprising that the forces in the core-mantle system might be making their presence felt throughout the earth as a whole. Indeed, workers have



DISTORTION OF SEISMIC WAVES enables researchers to analyze the heterogeneous characteristics of the D'' layer. Waves emanating from an earthquake are smooth. When they pass through the D'' region, their wave fronts become rippled, or corrugated. The corrugation is measured by a dense array of seismometers located on another part of the earth. One such array, in Norway, was originally constructed to monitor seismic waves generated by underground nuclear tests.

found tantalizing evidence that suggests that the core-mantle zone strongly influences two features observable at the surface. They are the wobbling in the earth's rotation, known as nutations, and the geomagnetic field.

Bruce A. Buffett, working with Irwin I. Shapiro at Harvard University, concluded that the core-mantle boundary affects the earth's nutations. He did so after making highly accurate calculations of the wobbling. The workers measured the wobbling using very long baseline interferometry. Radio astronomers often rely on this technique to make highly precise measurements of stellar objects. Various tidal forces had been thought to be solely responsible for the earth's nutations. Such mechanisms include the friction generated as the solid surface of the earth rubs against the atmosphere and oceans as well as the gravitational interactions with the sun and the moon. Buffett discovered, however, a component of the nutations that could not be explained by tidal forces. Motivated by the diamond-cell results, he considered the possibility that a thin reaction zone at the core-mantle boundary might offer an explanation for the anomalous nutation component.

He showed that such a reaction layer can easily account for the nutation signal if the layer contains electrically conducting material, as inferred from experiments. The magnetic-field lines emanating from the core would induce small electric currents to flow in the conducting mixture. These small cur-

rents in turn produce their own magnetic fields. The small magnetic fields interact with the main geomagnetic-field lines, much as poles of a magnet can either attract or repel. In essence, the core and mantle behave as two magnets that push against each other. This coupling affects the nutations. The baseline interferometry data are nicely explained if one invokes a heterogeneous reaction zone that contains metal and is a few hundred meters thick.

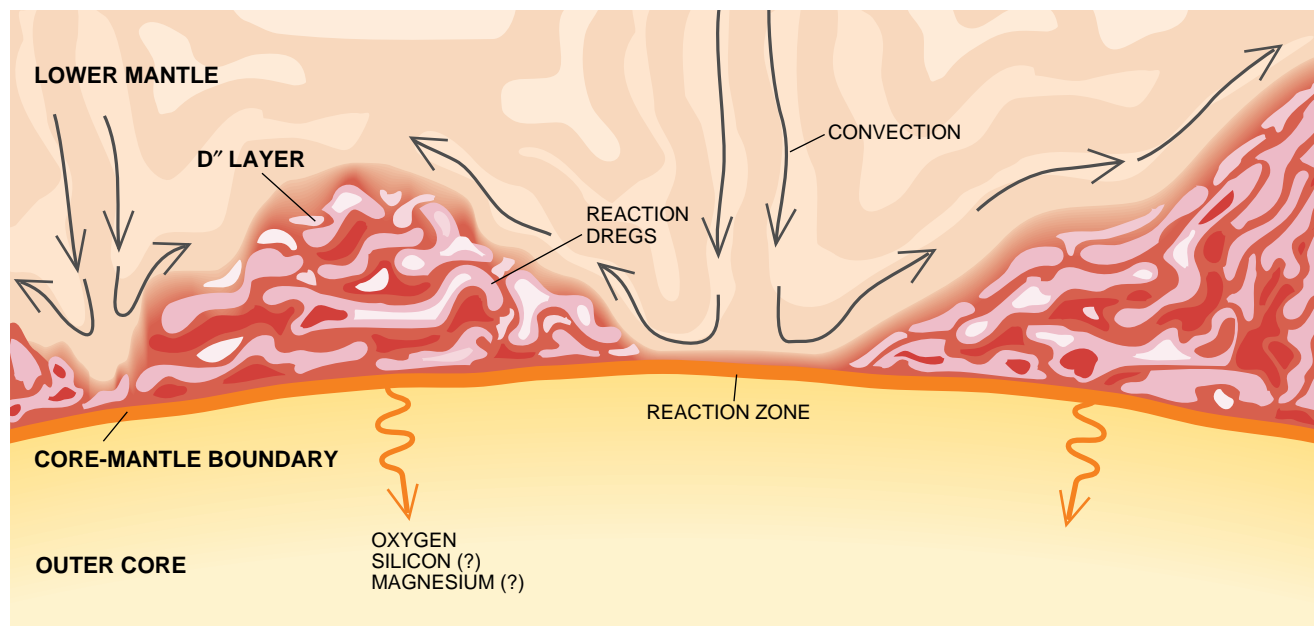
Indeed, our experiments predicted just such a configuration for the reaction zone. The products of the reaction at the bottom of the mantle are expected to consist of a few tens of percent of electrically conducting alloys, such as iron silicide and wüstite. A zone consisting of only 15 to 20 percent alloy would be sufficient to account for the nutations. Thus, our conclusion that the reaction zone would be hundreds of meters thick and would fluctuate in thickness and conductivity along the core-mantle boundary accords well with Buffett's hypothesis.

The second observable surface effect that the core-mantle region influences is the earth's magnetic field. The origin of the main geomagnetic field is well understood, at least in general terms [see "The Evolution of the Earth's Magnetic Field," by Jeremy Bloxham and David Gubbins; *SCIENTIFIC AMERICAN*, December 1989]. A dynamo effect, rather than conventional magnetism of the iron in the core, pro-

duces the geomagnetic field. (Iron is no longer magnetic at either the pressures or the temperatures existing in the core.) The churning of the liquid-metal outer core essentially acts as an electric current moving through wire. Like the wire, the core then generates a magnetic field around itself.

Convection powers the motion of the molten outer core. The hot liquid from deep inside rises toward the cooler top of the core. The movement transfers heat upward and causes a convective flow. Cooler liquid from near the core-mantle boundary sinks downward and thus also helps to power the convection. Additional sources of convection, such as internal separation of solids and liquid in the outer core, are possible. In this way, the mechanical energy of convection—fluid flow in the outer core—is converted to magnetic energy.

The principles that govern this process are called magnetohydrodynamics—a combination of hydrodynamics, or the physics of fluid flow, and electromagnetism. The mathematical equations behind the process, however, are so complicated that no one has been able to solve them in complete generality. As a result, the solutions obtained are based on physically plausible but greatly simplified assumptions. The solutions obtained from these assumptions do not necessarily explain the small but observable details of the earth's magnetic field, such as the slight ripples in the field intensity. Perhaps the discrepancy results from one of the tra-



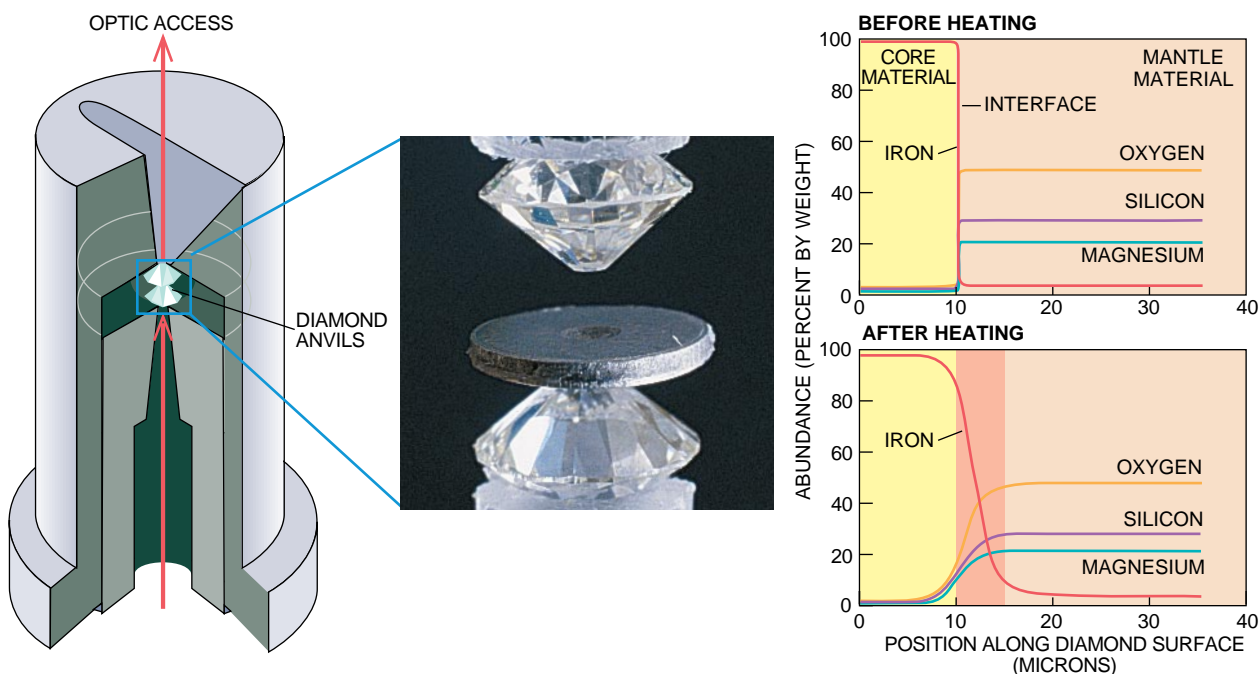
D'' LAYER forms as a result of chemical reactions between the core and mantle. In essence, the mantle rock partly dissolves in the liquid iron of the outer core, producing metal-rich "dregs" that are deposited on the core-mantle boundary. Con-

vection in the mantle tends to disperse the products under downwelling regions and to build up material at upwellings. A thin layer enriched in oxygen and possibly silicon and magnesium may exist on the inner side of the core-mantle interface.

The Diamond-Anvil High-Pressure Cell

This device (*left*) can duplicate the pressures and temperatures of the deep earth. The material to be squeezed and heated is placed in a metal-foil gasket between the tips of two diamond anvils (*photograph*). Turning a thumbscrew (*not shown*) brings the anvils together, compressing the sample. A laser beam can be focused through the diamond to heat the sample. Compositional profiles (*right*) show the abundance of iron, oxygen, silicon

and magnesium (elements at the core-mantle boundary) before and after heating. The amounts have been plotted against the element's position on the surface of one of the diamonds, as measured from an edge. After heating, the interface region broadens, spanning between about 10 to 15 microns. The broadening indicates that the elements have reacted. The reaction produces a mixture of metallic alloys (FeSi and FeO) and insulating oxides (MgSiO₃ and SiO₂).



ditional simplifications used in the calculation: that the metallic core is surrounded by an electrically insulating region, corresponding to the mantle. Geophysicists are now recognizing that the lowermost mantle is not completely insulating but consists of a heterogeneous mixture of metallic alloys and insulating silicates.

Motivated by this information, Friedrich H. Busse of Bayreuth University in Germany recently reexamined the magnetohydrodynamic equations. He discovered an entirely new class of mathematical solutions to the dynamo problem that result directly from the variations in electrical conductivity in the lowermost mantle. The solutions depend on two major factors. One is that the geomagnetic-field lines are essentially "frozen" into the liquid metal of the outer core. So, locked into place, the field lines move only with the convective flow of the liquid outer core. The second factor is that metallic regions embedded within the D'' layer interfere with the horizontal movement of magnetic-field

lines emanating from the core. The D'' layer can then deflect or pile together the field lines from the core. Both factors would, according to Busse's calculations, create local magnetic fields at the bottom of the mantle. The fields would explain several complexities of the geomagnetic field, including the observed ripples in field strength.

The electromagnetic characteristics of the core-mantle boundary may also affect the reversals of the earth's magnetic field [see "Ancient Magnetic Reversals: Clues to the Geodynamo," by Kenneth A. Hoffman; *SCIENTIFIC AMERICAN*, May 1988]. During reversals, which occur every few 100,000 years, the magnetic poles seem to follow a preferred trajectory. Such preference seems especially evident for the most recent reversals in the earth's history. S. Keith Runcorn of Imperial College in London and of the University of Alaska has postulated several mechanisms by which the electrical variations of the D'' layer might influence the path of the magnetic poles.

In a sense, then, the dynamics be-

tween the core and mantle extend beyond the earth, stretching well into space via the geomagnetic field. We now recognize the planetary importance of the core-mantle interface, and improved technology is certain to clarify how this remote region shapes the evolution of the earth.

FURTHER READING

STRUCTURE OF THE CORE-MANTLE TRANSITION ZONE: A CHEMICAL AND THERMAL BOUNDARY LAYER. Thorne Lay in *EOS: Transactions, American Geophysical Union*, Vol. 70, No. 4, pages 49-59; January 24, 1989.

THE NATURE OF THE EARTH'S CORE. R. Jeanloz in *Annual Review of Earth and Planetary Sciences*, Vol. 18, pages 357-386; 1990.

EARTH'S CORE-MANTLE BOUNDARY: RESULTS OF EXPERIMENTS AT HIGH PRESSURES AND TEMPERATURES. E. Knittle and R. Jeanloz in *Science*, Vol. 251, pages 1438-1443; March 22, 1991.

DEEP INTERIOR OF THE EARTH. John A. Jacobs. Chapman & Hall, 1992.

How Cells Respond to Stress

During emergencies, cells produce stress proteins that repair damage. Inquiry into how they work offers promise for coping with infection, autoimmune disease and even cancer

by William J. Welch

Immediately after a sudden increase in temperature, all cells—from the simplest bacterium to the most highly differentiated neuron—increase production of a certain class of molecules that buffer them from harm. When biologists first observed that phenomenon 30 years ago, they called it the heat-shock response. Subsequent studies revealed that the same response takes place when cells are subjected to a wide variety of other environmental assaults, including toxic metals, alcohols and many metabolic poisons. It occurs in traumatized cells growing in culture, in the tissues of feverish children and in the organs of heart-attack victims and cancer patients receiving chemotherapy. Because so many different stimuli elicit the same cellular defense mechanism, researchers now commonly refer to it as the stress response and to the expressed molecules as stress proteins.

In their pursuit of the structure and function of the stress proteins, biologists have learned that they are far more than just defensive molecules. Throughout the life of a cell, many of these proteins participate in essential metabolic processes, including the pathways by which all other cellular proteins are synthesized and assembled. Some stress proteins appear to orchestrate the activities of molecules that

regulate cell growth and differentiation.

The understanding of stress proteins is still incomplete. Nevertheless, investigators are already beginning to find new ways to put the stress response to good use. It already shows great potential for pollution monitoring and better toxicologic testing. The promise of medical applications for fighting infection, cancer and immunologic disorders is perhaps more distant, but it is clearly on the horizon.

Such uses were far from the minds of the investigators who first discovered the stress response; as happens so often in science, it was serendipitous. In the early 1960s biologists studying the genetic basis of animal development were focusing much of their attention on the fruit fly *Drosophila melanogaster*. *Drosophila* is a convenient organism in which to study the maturation of an embryo into an adult, in part because it has an unusual genetic feature. Cells in its salivary glands carry four chromosomes in which the normal amount of DNA has been duplicated thousands of times; all the copies align beside one another. These so-called polytene chromosomes are so large that they can be seen through a light microscope. During each stage of the developmental process, distinct regions along the polytene chromosomes puff out, or enlarge. Each puff is the result of a specific change in gene expression.

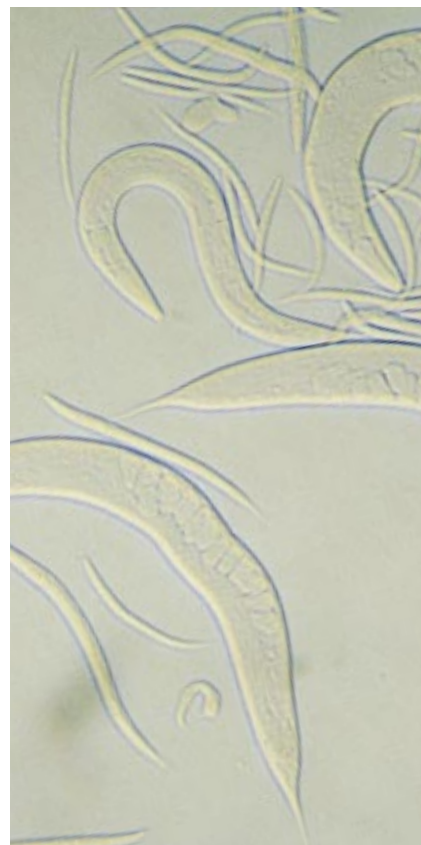
During the course of his studies, F. M. Ritossa of the International Laboratory of Genetics and Biophysics in Naples saw that a new pattern of chromosomal puffing followed the exposure of the

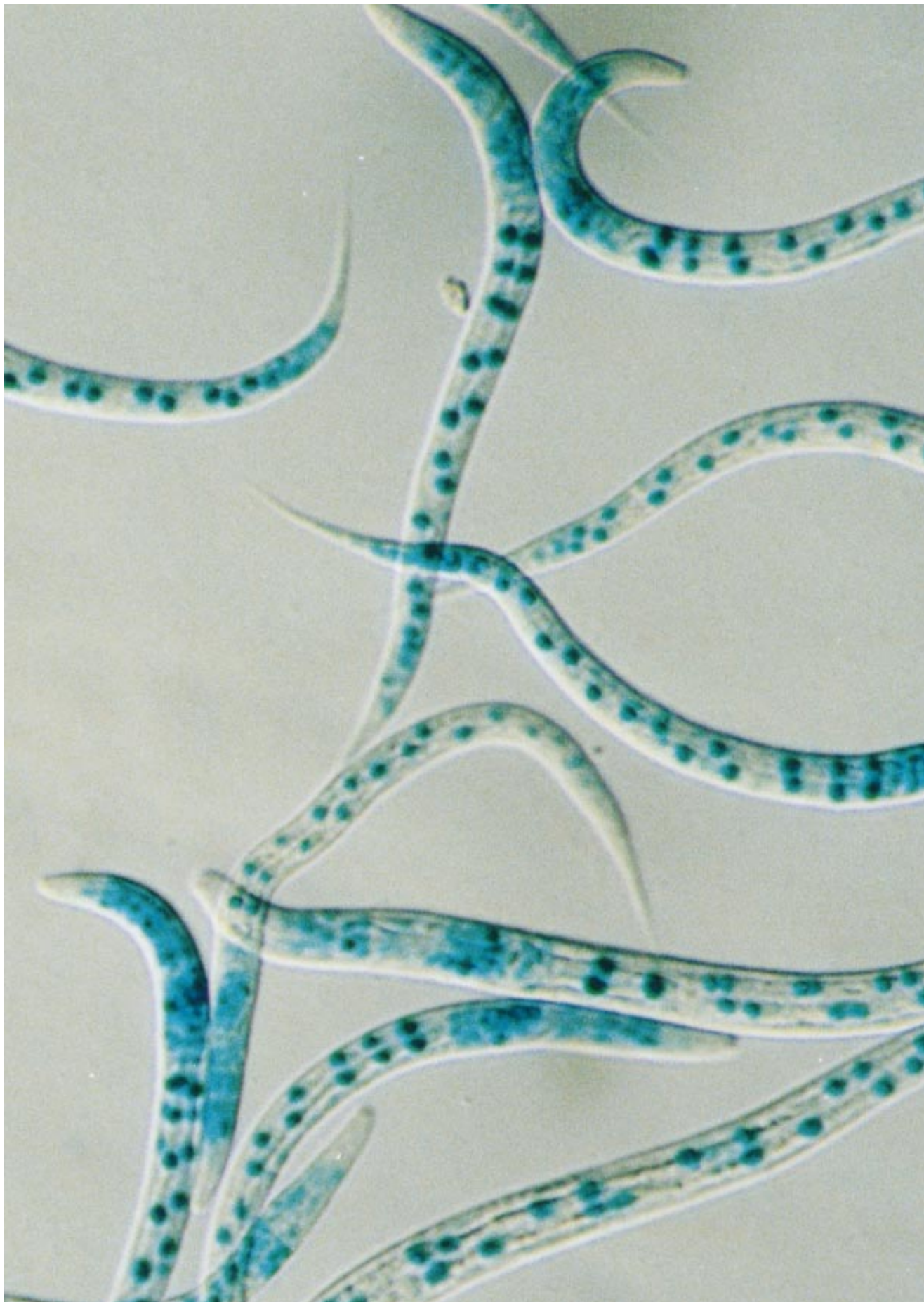
isolated salivary glands to temperatures slightly above those optimal for the fly's normal growth and development. The puffing pattern appeared within a minute or two after the temperature rise, and the puffs continued to increase in size for as long as 30 to 40 minutes. Over the next decade, other investigators built on Ritossa's findings.

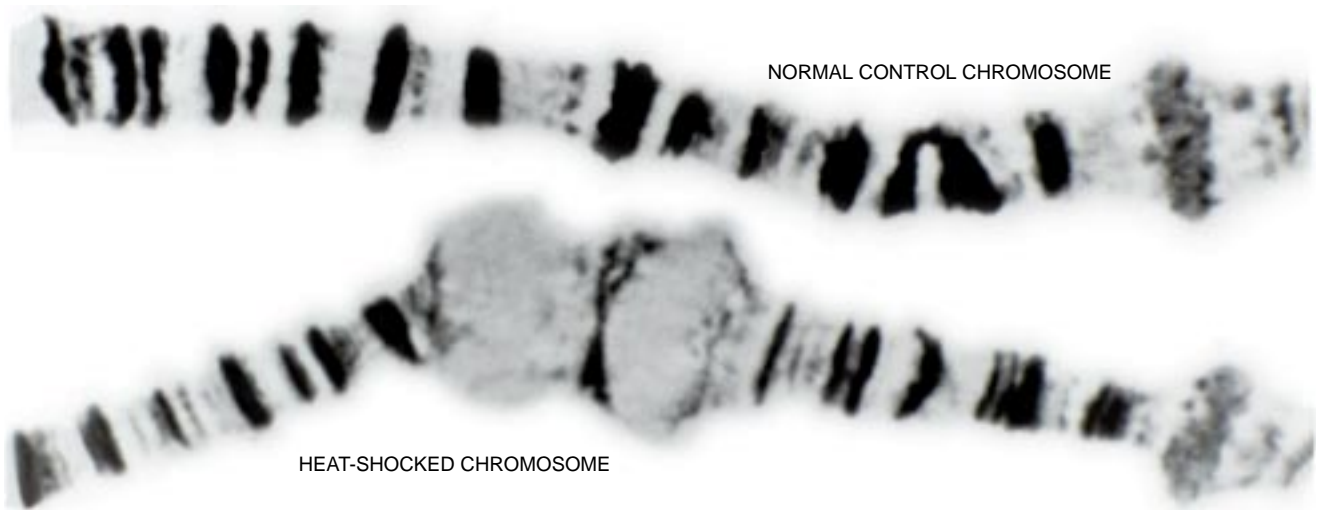
In 1974 Alfred Tissières, a visiting scientist from the University of Geneva, and Herschel K. Mitchell of the California Institute of Technology demonstrated that the heat-induced chromosomal puffing was accompanied by the high-level expression of a unique set of "heat shock" proteins. Those new chromosomal puffs represented sites in the DNA where specific messenger RNA

WILLIAM J. WELCH has spent more than a decade characterizing the stress response in mammalian cells and investigating its role in human disease. He is associate professor at the Lung Biology Center of the University of California, San Francisco. After completing his undergraduate studies in biology and chemistry in 1976 at the University of California, Santa Cruz, Welch went on to graduate work in chemistry at the Salk Institute for Biological Studies and the University of California, San Diego. The latter institution awarded him a Ph.D. in 1980. Welch is also a consultant to Stressgen Biotechnologies in Victoria, British Columbia.

GENETICALLY ENGINEERED WORMS are normally clear (*right*) but can be made to turn blue (*opposite page*) when they are subjected to toxins, excess heat or other environmental assaults. The color is caused by the activity of a reporter gene linked to the expression of genes for stress proteins that help the organisms survive harsh conditions.







PUFFS in the polytene chromosomes of the fruit fly *Drosophila melanogaster* (left) indicate local gene activity. As the fly passes through developmental stages, the puffing pattern changes. Abnormally high temperatures also stimulate certain puffs to form, as shown above. These puffs reflect the expression of genes for heat-shock proteins belonging to the hsp 70 molecular family.

molecules were made; these messenger RNAs carried the genetic information for synthesizing the individual heat-shock proteins.

By the end of the 1970s evidence was accumulating that the heat-shock response was a general property of all cells. Following a sudden increase in temperature, bacteria, yeast, plants and animal cells grown in culture all increased their expression of proteins that were similar in size to the *Drosophila* heat-shock proteins. Moreover, investigators were finding that cells produced one or more heat-shock proteins whenever they were exposed to heavy metals, alcohols and various other metabolic poisons.

Because so many different toxic stimuli brought on similar changes in gene expression, researchers started referring to the heat-shock response more generally as the stress response and to the accompanying products as stress proteins. They began to suspect that this universal response to adverse changes in the environment represented a basic cellular defense mechanism. The stress proteins, which seemed to be expressed only in times of trouble, were presumably part of that response.

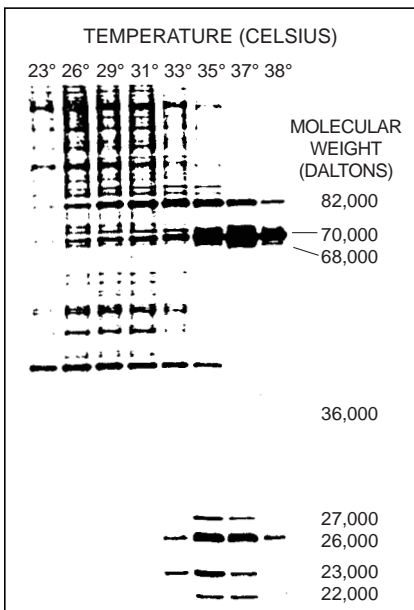
Mounting evidence during the next few years confirmed that stress proteins did play an active role in cellular defense. Researchers were able to identify and isolate the genes that encoded the individual stress proteins. Mutations in those genes produced interesting cellular abnormalities. For example, bacteria carrying mutations in the genes encoding several of the stress proteins exhibited defects in DNA and RNA syn-

thesis, lost their ability to undergo normal cell division and appeared unable to degrade proteins properly. Such mutants were also incapable of growth at high temperatures.

Cell biologists soon discovered that, as in bacteria, the stress response played an important role in the ability of animal cells to withstand brief exposures to high temperatures. Animal cells given a mild heat shock—one sufficient to increase the levels of the stress proteins—were better protected against a second heat treatment that would otherwise have been lethal. Moreover, those thermotolerant cells were also less susceptible to other toxic agents. Investigators became convinced that the stress response somehow protected cells against varied environmental insults.

As scientists continued to isolate and characterize the genes encoding the stress proteins from different organisms, two unexpected results emerged. First, many of the genes that encoded the stress proteins were remarkably similar in all organisms. Elizabeth A. Craig and her colleagues at the University of Wisconsin reported that the genes for heat-shock protein (hsp) 70, the most highly induced stress protein, were more than 50 percent identical in bacteria, yeast and *Drosophila*. Apparently, the stress proteins had been conserved throughout evolution and likely served a similar and important function in all organisms.

The second unexpected finding was that many stress proteins were also expressed in normal and unstressed cells, not only in traumatized ones. Consequently, researchers subdivided the



HEAT-SHOCK PROTEIN LEVELS rise in cells as the temperature increases. In these electrophoretic gels, each horizontal band is a protein found in the cells of *Drosophila*. As the temperature rises, the cells stop making most proteins and produce far more of the heat-shock proteins. The most prevalent of these belong to the hsp 70 family, which has molecular weights of around 70,000 daltons.

stress proteins into two groups: those constitutively expressed under normal growth conditions and those induced only in cells experiencing stress.

Investigators were still perplexed as to how so many seemingly different toxic stimuli always led to the increased expression of the same group of proteins. In 1980 Lawrence E. Hightower, working at the University of Connecticut, provided a possible answer. He noticed that many of the agents that induced the stress response were protein denaturants—that is, they caused proteins to lose their shapes. A protein consists of long chains of amino acids folded into a precise conformation. Any disturbance of the folded conformation can lead to the protein's loss of biological function.

Hightower therefore suggested that the accumulation of denatured or abnormally folded proteins in a cell initiated a stress response. The stress proteins, he reasoned, might somehow facilitate the identification and removal of denatured proteins from the traumatized cell. Within a few years Richard Voellmy of the University of Miami and Alfred L. Goldberg of Harvard University tested and confirmed Hightower's proposal. In a landmark study, they showed that injecting denatured proteins into living cells was sufficient to induce a stress response.

Thereafter, several laboratories set out to purify and characterize the biochemical properties of the stress proteins. The most highly inducible heat-shock protein, hsp 70, was the focus of much of this work. Using molecular probes, researchers learned that after a heat shock, much hsp 70 accumulated inside a nuclear structure called the nucleolus. The nucleolus manufactures ribosomes, the organelles that synthesize proteins. That location for hsp 70 was intriguing: previous work had demonstrated that after heat shock, cells stopped making ribosomes. Indeed, their nucleolus became awash in denatured ribosomal particles. Hugh R. B. Pelham of the Medical Research Council's Laboratory of Molecular Biology in Cambridge, England, therefore suggested that hsp 70 might somehow recognize denatured intracellular proteins and restore them to their correctly folded, biologically active shape.

In 1986 Pelham and his colleague Sean Munro succeeded in isolating several genes, all of which encoded proteins related to hsp 70. They noticed that one form of hsp 70 was identical to immunoglobulin binding protein (BiP). Other researchers had shown that BiP was involved in the preparation of immunoglobulins, or antibodies, as well as other proteins for secretion. BiP bound

to newly synthesized proteins as they were being folded or assembled into their mature form. If the proteins failed to fold or assemble properly, they remained bound to BiP and were eventually degraded. In addition, under conditions in which abnormally folded proteins accumulated, the cell synthesized more BiP.

Taken together, those observations indicated that BiP helped to orchestrate the early events associated with protein secretion. BiP seemed to act as a molecular overseer of quality control, allowing properly folded proteins to enter the secretory pathway but holding back those unable to fold correctly.

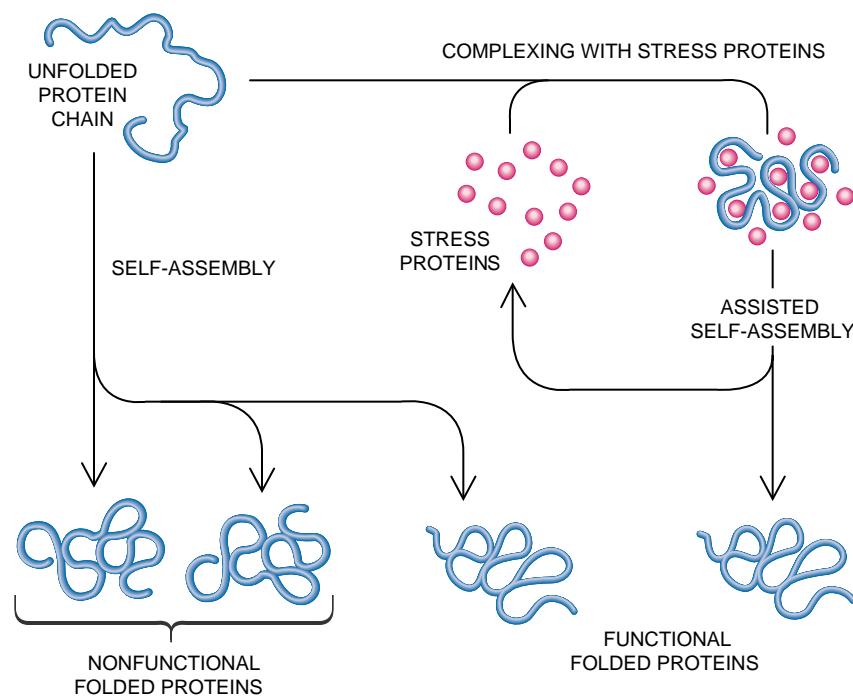
As more genes encoding proteins similar to hsp 70 and BiP came to light, it became evident that there was an entire family of hsp 70-related proteins. All of them shared certain properties, including an avid affinity for adenosine triphosphate (ATP), the molecule that serves as the universal, intracellular fuel. With only one exception, all these related proteins were present in cells growing under normal conditions (they were constitutive), yet in cells experiencing metabolic stress, they were synthesized at much higher levels. Moreover, all of them mediated the maturation of other cellular proteins, much as BiP did. For example, the cytoplasmic

forms of hsp 70 interacted with many other proteins that were being synthesized by ribosomes.

In healthy or unstressed cells the interaction of the hsp 70 family member with immature proteins was transient and ATP-dependent. Under conditions of metabolic stress, however, in which newly synthesized proteins experienced problems maturing normally, the proteins remained stably bound to an hsp 70 escort.

The idea that members of the hsp 70 family participated in the early steps of protein maturation paralleled the results emerging from studies of a different family of stress proteins. Pioneering work by Costa Georgopoulos of the University of Utah and others had shown that mutations in the genes for two related stress proteins, groEL and groES, render bacteria unable to support the growth of small viruses that depend on the cellular machinery provided by their hosts. In the absence of functional groEL or groES, many viral proteins fail to assemble properly.

Proteins similar to the bacterial groEL and groES stress proteins were eventually found in plant, yeast and animal cells. Those proteins, which are known as hsp 10 and hsp 60, have been seen only in mitochondria and chloroplasts. Recent evidence suggests that more forms probably appear in other intracellular compartments.



PROTEIN FOLDING occurs spontaneously because of thermodynamic constraints imposed by the protein's sequence of hydrophilic and hydrophobic amino acids. Although proteins can fold themselves into biologically functional configurations (self-assembly), errors in folding can occasionally occur. Stress proteins seem to help ensure that cellular proteins fold themselves rapidly and with high fidelity.

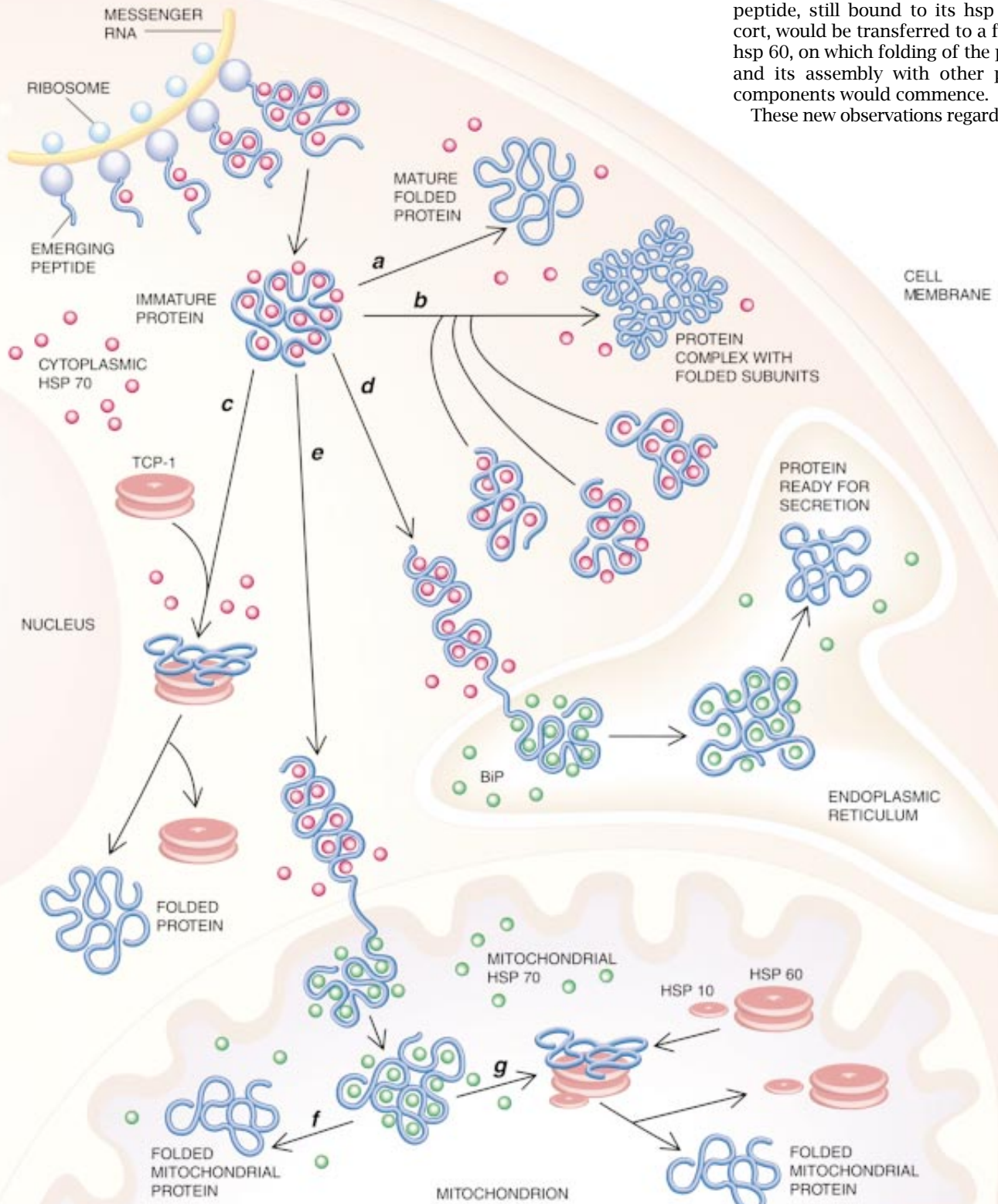
Biochemical studies have provided compelling evidence that hsp 10 and hsp 60 are essential to protein folding and assembly. The hsp 60 molecule consists of two seven-membered rings stacked one atop the other. This large structure appears to serve as a "work-

bench" onto which unfolded proteins bind and acquire their final three-dimensional structure. According to current thought, the folding process is extremely dynamic and involves a series of binding and release events. Each event requires energy, which is provided by the enzymatic splitting of ATP, and the participation of the small hsp 10 molecules. Through multiple rounds of

binding and release, the protein undergoes conformational changes that take it to a stable, properly folded state.

Investigators suspect that both the hsp 60 and the hsp 70 families work together to facilitate protein maturation. As a new polypeptide emerges from a ribosome, it is likely to become bound to a form of hsp 70 in the cytoplasm or inside an organelle. Such an interaction may prevent the growing polypeptide chain from folding prematurely. Once its synthesis is complete, the new polypeptide, still bound to its hsp 70 escort, would be transferred to a form of hsp 60, on which folding of the protein and its assembly with other protein components would commence.

These new observations regarding the



properties of hsp 70 and hsp 60 have forced scientists to reconsider previous models of protein folding. Work done in the 1950s and 1960s had established that a denatured protein could spontaneously refold after the denaturing agent was removed. This work led to the concept of protein self-assembly, for which Christian B. Anfinsen received a Nobel Prize in Chemistry in 1972. According to that model, the process of folding was dictated solely by the sequence of amino acids in the polypeptide. Hydrophobic amino acids (those that are not water soluble) would position themselves inside the coiling molecule, while hydrophilic amino acids (those that are water soluble) would move to the surface of the protein to ensure their exposure to the aqueous cellular environment. Folding would thus be driven entirely by thermodynamic constraints.

The principle of self-assembly is still regarded as the primary force that drives proteins into their final conformation. Now, however, many investigators suspect that protein folding requires the activity of other cellular components, including the members of the hsp 60 and hsp 70 families of stress proteins.

Accordingly, R. John Ellis of the University of Warwick and other scientists have begun to refer to hsp 60, hsp 70 and other stress proteins as "molecular chaperones." Although the molecules do not convey information for the folding or assembly of proteins, they do ensure that those processes occur quickly and with high fidelity. They expedite self-assembly by reducing the possibility that a maturing protein will head down an inappropriate folding pathway.

Having established a role for some stress proteins as molecular chaperones in healthy and unstressed cells, investigators have turned their attention to determining why those proteins are expressed at higher levels in times of stress. One clue is the conditions that increase the expression of the stress proteins. Temperatures that are suffi-

cient to activate the stress response may eventually denature some proteins inside cells. Heat-denatured proteins, like newly synthesized and unfolded proteins, would therefore represent targets to which hsp 70 and hsp 60 can bind. Over time, as more thermally denatured proteins become bound to hsp 60 and hsp 70, the levels of available molecular chaperones drop and begin to limit the ability of the cell to produce new proteins. The cell somehow senses this reduction and responds by increasing the synthesis of new stress proteins that serve as molecular chaperones.

Researchers suspect that a rise in the expression of stress proteins may also be a requirement for the ability of cells to recover from a metabolic insult. If heat or other metabolic insults irreversibly denature many cellular proteins, the cell will have to replace them. Raising the levels of those stress proteins that act as molecular chaperones will help facilitate the synthesis and assembly of new proteins. In addition, higher levels of stress proteins may prevent the thermal denaturation of other cellular proteins.

The repair and synthesis of proteins are vital jobs in themselves. Nevertheless, stress proteins also serve a pivotal role in the regulation of other systems of proteins and cellular responses. Another family of stress proteins, epitomized by one called hsp 90, is particularly noteworthy in this regard.

Initial interest in hsp 90 was fueled by reports of its association with some cancer-causing viruses. In the late 1970s and early 1980s cancer biologists were focusing considerable attention on the mechanism by which certain viruses infect cells and cause them to become malignant. In the case of Rous sarcoma virus, investigators had pinpointed a viral gene that was responsible for the development of malignant properties. The enzyme it produced, pp60src, acted on other proteins that probably regulated cellular growth. Three laborato-

ries independently reported that after its synthesis in the cytoplasm, pp60src rapidly associates with two proteins: one called p50 and the other hsp 90.

When pp60src is in the cytoplasm and is linked to its two escorts, it is enzymatically inactive. As the trio of molecules moves to the plasma membrane, the hsp 90 and the p50 fall away and allow the pp60src to deposit itself in the membrane and become active. Similar interactions between hsp 90, p50 and cancer-causing enzymes encoded by several other tumor viruses have been discovered. When bound to hsp 90 and p50, these viral enzymes seem incapable of acting on the cellular targets necessary for the development of the malignant state.

Some studies have also linked hsp 90 to another important class of molecules in mammalian cells, the steroid hormone receptors. Steroid hormones mediate several vital biological processes in animals. For example, the glucocorticoid steroids help to suppress inflammation. Other steroid hormones play important roles in sexual differentiation and development. When a steroid receptor binds to its specific hormone, the receptor becomes capable of interacting with DNA and either activating or repressing the expression of certain genes.

A crucial question concerned how steroid receptors were kept inactive inside a cell. The answer became clear following the characterization of both the active and inactive forms of the progesterone receptor. In the absence of hormone the receptor associates with several cellular proteins, among them hsp 90, which maintain it in an inactive state. After binding to progesterone, the receptor is released from the hsp 90 and experiences a series of events that allows it to bind with DNA. As with the viral enzymes, hsp 90 seems to regulate the biological activity of steroid hormone receptors.

Scientists are beginning to realize practical applications for the stress response. Medicine is one area that stands to benefit. When an individual suffers a heart attack or stroke, the delivery of blood to the heart or brain is temporarily compromised, a condition referred to as ischemia. While deprived of oxygen, the affected organ cannot maintain its normal levels of ATP, which causes essential metabolic processes to falter. When blood flow is restored, the ischemic organ is rapidly reoxygenated—yet that too can be harmful. Often the rapid reexposure to oxygen generates highly reactive molecular species, known as free radicals, that can do further damage.

In animal studies, researchers have

SEVERAL PATHWAYS for folding and distributing proteins inside cells are managed by stress proteins. In many cases, different stress proteins seem to work in tandem. The cytoplasmic form of hsp 70 binds to proteins being produced by the ribosomes to prevent their premature folding. The hsp 70 may dissociate from the protein and allow it to fold itself into its functional shape (a) or to associate with other proteins and thereby form larger, multimeric complexes (b). In some cases, proteins are passed from hsp 70 to another stress protein, TCP-1, before final folding and assembly occur (c). If the protein is destined for secretion, it may be carried to the endoplasmic reticulum and given to BiP or another related stress protein that directs its final folding (d). Other proteins are transferred to mitochondria or other organelles (e). Inside the mitochondrion, another specialized form of hsp 70 sometimes assists the protein in its final folding (f), but in many cases the protein is passed on to a complex of hsp 60 and hsp 10 (g). The hsp 60 molecule seems to serve as a "workbench" on which the mitochondrial protein folds.

observed the induction of stress responses in both the heart and brain after brief episodes of ischemia and reperfusion. The magnitude of the resulting stress response appears to correlate directly with the relative severity of the damage. Clinicians are therefore beginning to examine the utility of using changes in stress protein levels as markers for tissue and organ injury.

Cells that produce high levels of stress proteins appear better able to survive the ischemic damage than cells that do not. Consequently, raising the levels of stress proteins, perhaps by pharmacological means, may provide additional protection to injured tissues and organs. Such a therapeutic approach might reduce the tissue damage from ischemia incurred during surgery or help to safeguard isolated organs used for transplantation, which often suffer from ischemia and reperfusion injury.

One exciting development concerns the role of the stress response in immunology and infectious diseases. Tuberculosis, malaria, leprosy, schistosomiasis and other diseases that affect millions of people every year are a consequence of infection by bacteria or parasitic microorganisms. Immunologists have found that the stress proteins made by these organisms are often the major antigens, or protein targets, that the immune system uses to

recognize and destroy the invaders. The human immune system may be constantly on the lookout for alien forms of stress proteins. The stress proteins of various pathogens, when produced in the laboratory by recombinant-DNA techniques, may therefore have potential as vaccines for preventing microbial infections. In addition, because they are so immunogenic, microbial stress proteins are being considered as adjuvants. Linked to viral proteins, they could enhance immune responses against viral infections.

Immunologists have also discovered a possible connection between stress proteins and autoimmune diseases. Most autoimmune diseases arise when the immune system turns against antigens in healthy tissues. In some of these diseases, including rheumatoid arthritis, ankylosing spondylitis and systemic lupus erythematosus, antibodies against the patient's own stress proteins are sometimes observed. If those observations are confirmed on a large number of patients, they may prove helpful in the diagnosis and perhaps the treatment of autoimmune disorders.

Because microbial stress proteins are so similar in structure to human stress proteins, the immune system may constantly be obliged to discern minor differences between the stress proteins of the body and those of invading microor-

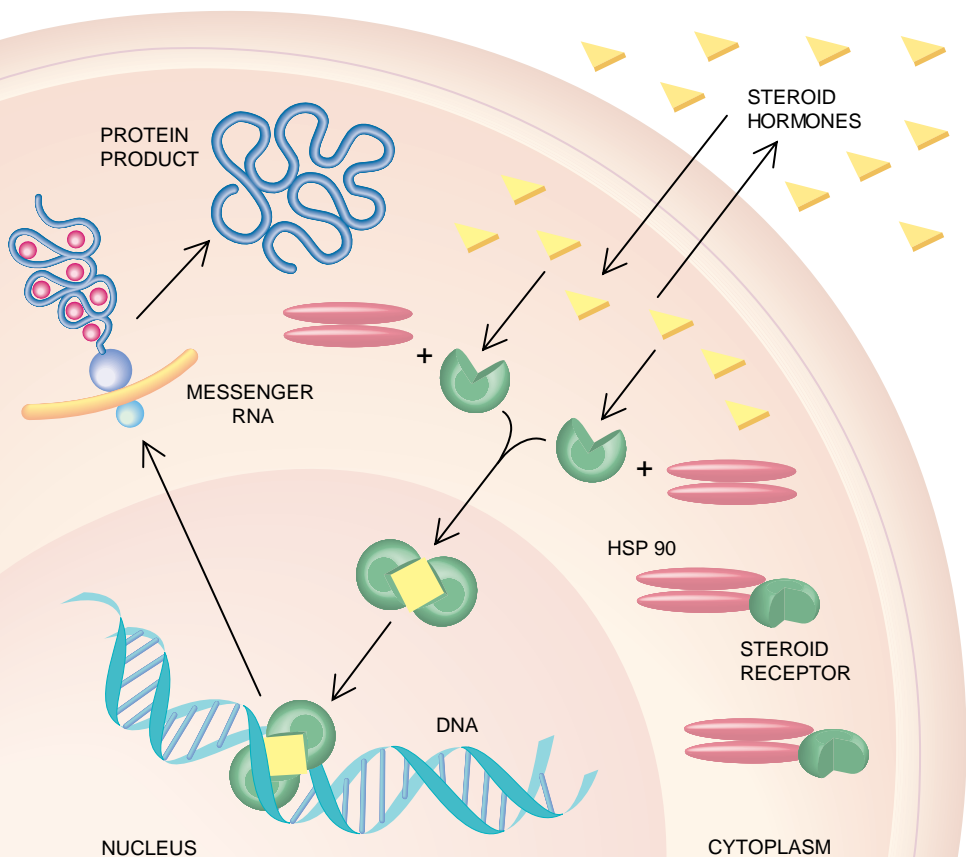
ganisms. The possibility that the stress proteins are uniquely positioned at the interface between tolerance to an infectious organism and autoimmunity is an intriguing idea that continues to spark debate among researchers.

The presence of antibodies against microbial stress proteins may prove useful in diagnostics. For example, the bacterium *Chlamydia trachomatis* causes a number of diseases, including trachoma, probably the world's leading cause of preventable blindness, and pelvic inflammatory disease, a major cause of infertility in women. Infection with chlamydia generally triggers the production of antibodies against chlamydial antigens, some of which are stress proteins. Often that immune response is effective and eventually eliminates the pathogen. Yet in some individuals, particularly those who have had repeated or chronic chlamydial infections, the immune response is overly aggressive and causes injury and scarring in the surrounding tissues.

Richard S. Stephens and his colleagues at the University of California at San Francisco have observed that more than 30 percent of women with pelvic inflammatory disease and more than 80 percent of women who have had ectopic pregnancies possess abnormally high levels of antibodies against the chlamydial groEL stress protein. Measurements of antibodies against chlamydial stress proteins may prove useful for identifying women at high risk for ectopic pregnancies or infertility.

The link between stress proteins, the immune response and autoimmune diseases becomes even more intriguing in light of other recent discoveries. Some members of the hsp 70 family of stress proteins are remarkably similar in structure and function to the histocompatibility antigens. The latter proteins participate in the very early stages of immune responses by presenting foreign antigens to cells of the immune system.

Researchers have wondered how any one histocompatibility protein could bind to a diverse array of different antigenic peptides. Recently Don C. Wiley and his colleagues at Harvard University helped to resolve that issue by deter-



RESPONSES TO STEROID HORMONES are controlled in part by hsp 90. This stress protein helps to maintain steroid receptors in their inactive form. When hormones are present, they bind to the receptor, and the hsp 90 is released. The activated receptor complex can then interact with DNA and initiate the expression of genes for certain proteins.

mining the three-dimensional structure of the class I histocompatibility proteins. A pocket or groove on the class I molecule, they found, is able to bind to different antigenic peptides. Simultaneously, James E. Rothman, who was then at Princeton University, reported that members of the hsp 70 family of stress proteins were also capable of binding to short peptides. That property of hsp 70 is consistent with its role in binding to some parts of unfolded or newly made polypeptide chains.

Computer models revealed that hsp 70 probably has a peptide-binding site analogous to that of the class I histocompatibility proteins. The apparent resemblance between the two classes of proteins appears even more intriguing because several of the genes that encode hsp 70 are located very near the genes for the histocompatibility proteins. Taken together, all the observations continue to support the idea that stress proteins are integral components of the immune system.

The ability to manipulate the stress response may also prove important in developing new approaches to treating cancer. Tumors often appear to be more thermally sensitive than normal tissues. Elevating the temperature of tissues to eradicate tumors is one idea that is still at the experimental stage. Nevertheless, in early tests, the use of site-directed hyperthermia, alone or in conjunction with radiation or other conventional therapies, has brought about the regression of certain types of tumors.

The stress response is not necessarily the physician's ally in the treatment of cancer—it may also be one of the obstacles. Because stress proteins afford cells added protection, anticancer therapies that induce a stress response may make a tumor more resistant to subsequent treatments. Still, researchers may yet discover ways to inhibit the ability of a tumor to mount a stress response and thereby render it defenseless against a particular therapy.

Scientists are also beginning to explore the potential use of the stress response in toxicology. Changes in the levels of the stress proteins, particularly those produced only in traumatized cells, may prove useful for assessing the toxicity of drugs, cosmetics, food additives and other products. Such work is only at a preliminary stage of development, but several application strategies are already showing signs of success.

Employing recombinant-DNA technologies, researchers have constructed cultured lines of "stress reporter" cells that might be used to screen for

Some Conditions That Induce the Expression of Stress Proteins

ENVIRONMENTAL STRESSORS

- HEAT SHOCK
- TRANSITION HEAVY METALS
- INHIBITORS OF ENERGY METABOLISM
- AMINO ACID ANALOGUES
- CHEMOTHERAPEUTIC AGENTS

STATES OF DISEASE

- VIRAL INFECTION
- FEVER
- INFLAMMATION
- ISCHEMIA
- HYPERTROPHY
- OXIDANT INJURY
- MALIGNANCY

NORMAL CELLULAR INFLUENCES

- CYCLE OF CELL DIVISION
- GROWTH FACTORS
- DEVELOPMENT AND DIFFERENTIATION

biological hazards. In such cells the DNA sequences that control the activity of the stress protein genes are linked to a reporter gene that encodes an enzyme, such as β -galactosidase. When these cells experience metabolic stress and produce more stress proteins, they also make the reporter enzyme, which can be detected easily by various assays. The amount of β -galactosidase expressed in a cell can be measured by adding a chemical substrate. If the reporter enzyme is present, the cell turns blue, and the intensity of the color is directly proportional to the concentration of the enzyme in the cell.

Using such reporter cells, investigators can easily determine the extent of the stress response induced by chemical agents or treatments. If such assays prove reliable, they could ultimately reduce or even replace the use of animals in toxicology testing.

An extension of the technique could also be used to monitor the dangers of environmental pollutants, many of which evoke stress responses. Toward that end, scientists have begun developing transgenic stress reporter organisms. Eve G. Stringham and E. Peter M. Candido of the University of British Columbia, along with Stressgen Biotechnologies in Victoria, have created trans-

genic worms in which a reporter gene for β -galactosidase is under the control of the promoter for a heat-shock protein. When these transgenic worms are exposed to various pollutants, they express the reporter enzyme and turn blue. Candido's laboratory is currently determining whether those stress reporter worms might be useful for monitoring a wide variety of pollutants.

Voellmy and Nicole Bournias-Vardibas, then at City of Hope National Medical Center in Duarte, Calif., have used a similar approach to create a line of transgenic stress reporter fruit flies. The fruit flies turn blue when exposed to teratogens, agents that cause abnormal fetal development. Significantly, that bioassay is responsive to many of the teratogens that are known to cause birth defects in humans. The door appears open for the development of other stress reporter organisms that could prove useful in toxicological and environmental testing.

More than 30 years ago heat-shock and stress responses seemed like mere molecular curiosities in fruit flies. Today they are at the heart of an active and vital area of research. Studies of the structure and function of stress proteins have brought new insights into essential cellular processes, including the pathways of protein maturation. Scientists are also learning how to apply their understanding of the stress response to solve problems in the medical and environmental sciences. I suspect we have only begun to realize all the implications of this age-old response by which cells cope with stress.

FURTHER READING

- THE INDUCTION OF GENE ACTIVITY IN *DROSOPHILA* BY HEAT SHOCK. M. Ashburner and J. J. Bonner in *Cell*, Vol. 17, No. 2, pages 241-254; June 1979.
- STRESS PROTEINS IN BIOLOGY AND MEDICINE. Richard I. Morimoto, Alfred Tissières and Costa Georgopoulos. Cold Spring Harbor Laboratory Press, 1990.
- MOLECULAR CHAPERONES. R. John Ellis and S. M. Van der Vies in *Annual Reviews of Biochemistry*, Vol. 60, pages 321-347; 1991.
- SUCCESSIVE ACTION OF DNAK, DNAJ AND GROEL ALONG THE PATHWAY OF CHAPERONE-MEDIATED PROTEIN FOLDING. Thomas Langer, Chi Lu, Harrison Echols, John Flanagan, Manajit K. Hayer and F. Ulrich Hartl in *Nature*, Vol. 356, No. 6371, pages 683-689; April 23, 1992.
- MAMMALIAN STRESS RESPONSE: CELL PHYSIOLOGY, STRUCTURE/FUNCTION OF STRESS PROTEINS, AND IMPLICATIONS FOR MEDICINE AND DISEASE. William J. Welch in *Physiological Reviews*, Vol. 72, pages 1063-1081; October 1992.

Intelligent Gels

*Soft aggregations of long-chain molecules
can shrink or swell in response to stimuli.
They may form the basis of a new kind of machine*

by Yoshihito Osada and Simon B. Ross-Murphy

Industrial products are generally made of metal, ceramic or plastic. These substances are by nature tough, hard, dry and easy to work with. Most engineers avoid "wet" components, such as liquids and gels. Liquids are entirely unable to maintain their shapes; gels are weak and tend to fail under small loads. Indeed, gels may be chemically unstable, and their properties suffer if they are allowed to dry. At present, they are used only in a few specialized applications, such as foods, water absorbents and soft contact lenses.

Yet a growing number of workers, taking inspiration from nature, have begun to see opportunities in these materials. Biological systems consist mostly of soft and wet substances. Indeed, many creatures live entirely without a rigid frame. The sea cucumber, for example, is essentially a water-swollen gel containing primitive organs; nevertheless, it can feed, reproduce and even defend itself from attackers. The sea cucumber responds rapidly to touch by stiffening its usually flexible body, and if it is further mishandled, it can cause part of its body wall to turn into a viscous fluid mass that prevents it from being grasped firmly.

To create such biomimetic systems, researchers employ a class of materials

known as polymer gels. In academic and industrial laboratories, gels have been made that can change both their size and shape, thereby converting chemical energy directly into mechanical work. Such materials could be used wherever power for more conventional devices is limited or difficult to obtain: underwater, in space or in the human body.

The gel-based machines built thus far are clearly a mere shadow of the sensing and self-regulating systems observed in the sea cucumber, much less the intricate choreography of muscles, tendons and other organs found in higher life-forms. But we believe that in the not too distant future, we will find a way to build "soft" machines that can respond in an intelligent fashion to their environments. Furthermore, seekers on that path have already invented gels that have immediate uses in chemistry, mechanical engineering and medicine.

As the chemist Dorothy Jordan Lloyd noted nearly 70 years ago, gels are "easier to recognize than to define." In general, it is safe to say, they consist of two components: a liquid and a network of long polymer molecules that hold the liquid in place and so give the gel what solidity it has.

To understand how a gel forms, consider a solution containing molecules of a typical synthetic polymer, such as polystyrene. Each molecule consists of about 10,000 linked units of styrene monomer, with a total weight of perhaps one million atomic mass units. Stretched out, it would be about three microns long. The actual space occupied by the molecule, however, is much smaller than this length suggests because the chain is flexible. Statistically speaking, there are many more states in which the chain is coiled or bunched than states in which it is fully stretched out. Therefore, the average distance between one end and the other is about one tenth of the stretched-out length.

In a dilute solution, each polymer coil is for the most part independent of

any other, but as the concentration of polymers increases, the spaces occupied by the coils begin to overlap. Because friction between polymer molecules hinders rapid flow, the solution starts to become thick and viscous. At still higher concentrations, the coils intertwine and entangle with one another like strands of spaghetti. This kind of system is viscoelastic: it has the properties both of a viscous solution and of an elastic solid. Metaphorically speaking, if one were to grasp one of the strands and pull slowly, it would flow out from the mass of other entangled chains. If, on the other hand, one were to pull sharply on a single strand, the entire mass would respond as a unit.

The definitions of "slow" and "rapid" here depend on the so-called relaxation time of the dissolved molecular chains. This time is, in effect, a measure of how long the molecule takes to return to a coiled state after being stretched. It is determined mostly by the mass of the molecule in question. Each molecule has a range of relaxation times, and real solutions contain molecules of different lengths; consequently, they have a distribution of relaxation times. These solutions gradually behave more and more like solids as the speed with which they are pushed or tugged increases. The length and stiffness of the polymer chains also affect the onset of viscoelastic behavior. Both factors increase the effective volume occupied by a single chain and, consequently, the likelihood that chains will interfere.

Under certain circumstances, polymer chains in solution do not merely entwine; permanent bonds form between them, creating large, branched chains. Eventually at least one of these dissolved megamolecules completely spans the container in which it sits. This is the transition from a viscoelastic solution to a gel. The gelled sample no longer flows like a polymer solution but rather has the properties of a solid. For example, if one side of the sample is perturbed mechanically, the distur-

YOSHIHITO OSADA and SIMON B. ROSS-MURPHY study the behavior and molecular structure of gels. Osada is professor of polymer science at Hokkaido University in Sapporo, Japan, where he leads a group that is developing gel-based mechanical devices. He received his doctorate from Moscow State University (Lomonosov) in 1970 and taught on the chemistry faculty of Ibaraki University from 1973 to 1991. Ross-Murphy, now a professor at King's College in London, worked in industry for 10 years before returning to academia in 1989. He also chairs the working party on polymer networks of the International Union of Pure and Applied Chemistry.

bance propagates directly along the molecular chains to the other side.

In some gels, the bonds that hold the molecular network together are classic covalent links in which two atoms share a pair of electrons, but in others they are more subtle. These cross-links include van der Waals forces, which attract adjacent atoms, hydrophobic interactions, which bring together water-shy parts of molecules, and hydrogen bonding, in which two molecules are held together by a hydrogen atom.

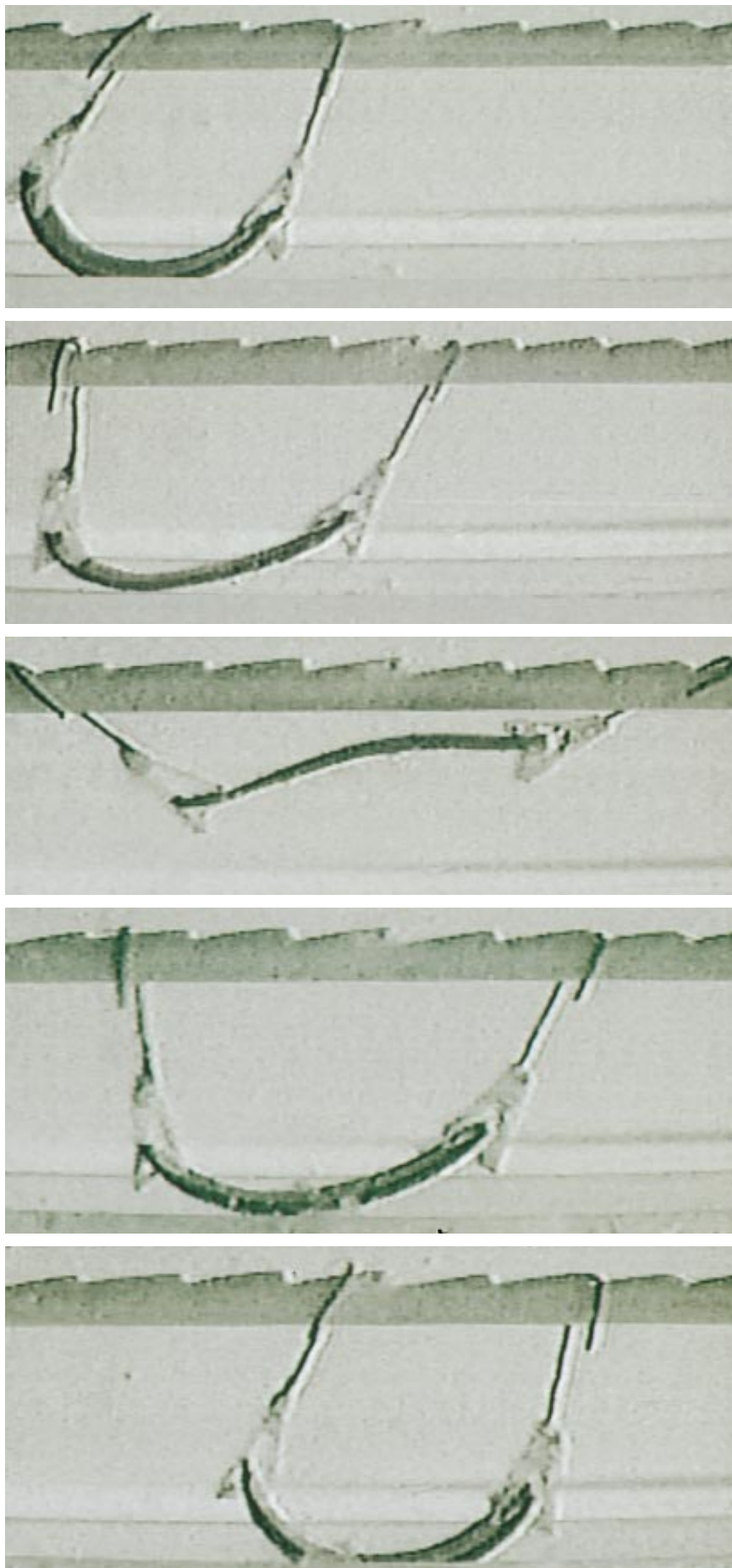
Sometimes these links are not localized, as are covalent bonds, but instead form extended "junction zones." On an even more complex level, biological polymers are held together by interactions between large-scale molecular structures. For example, gelatin desserts—those brightly colored, sugary cubes of bone-derived gelatin—get their wobbly mechanical strength from interwound triple helices of protein.

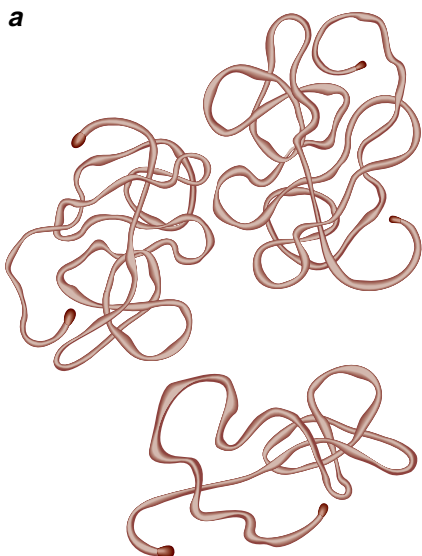
The tangled network of polymer strands is one aspect of a gel's essential nature. Another, equally important in principle and vastly predominant in volume, is the solvent that pervades the network and gives it bulk. The amount of solvent that any particular gel contains depends on a complex interaction between the elasticity of the polymer network and the affinity of the network's atoms for those of the solvent.

If a single polymer molecule is placed in a solvent, the molecule will either spread out or bunch up. If the polymer segments and the solvent atoms are attracted to one another, the average coil dimensions tend to increase in order to maximize the number of interactions between polymer segments and solvent molecules. The polymer swells, and the solvent is "thermodynamically good." Conversely, in a poor solvent the polymer segments will tend to avoid solvent molecules, and the polymer will appear to shrink.

If a gel is placed in a good solvent, its dimensions will increase until the polymer network has stretched enough so the resulting elastic force counteracts the inflow of solvent molecules. The elastic force depends on the degree of

GEL LOOPER, an inchwormlike device that moves by repeatedly curling and straightening itself, was developed by Osada. Surfactant molecules in the liquid surrounding the looper collect on the gel's top surface under the influence of an electric field, causing the gel to shrink. When the polarity of the electric field is changed, the surfactant goes back into the solution.





LONG-CHAIN MOLECULES cause a solution to become viscous (a) because they interfere with one another as the solution flows. As their concentration increases, the molecules become entangled, yielding viscoelastic behavior that partakes of both solid and liquid traits (b). If the intertwined molecules bond with one another, the result is a gel (c, d).

cross-linking (which limits the maximum extension of the polymer chains); thus, a strongly cross-linked gel will swell less than a weakly cross-linked one.

If the polymer from which a gel is made contains charged groups (molecules that readily accept or give up electrons) along its backbone, then additional effects may come into play. The first of these is the so-called polyelectrolyte effect. In pure water, a polymer containing charged groups will tend to expand its dimensions in order to minimize the repulsion between them. If a simple electrolyte, such as common salt, is dissolved in the water, the ions of opposite charge to those carried by the polymer can neutralize its charge. Thus, as the ionic strength is increased, the polymer returns to its coiled shape. A polyelectrolyte gel will swell enormously in pure water or in low-ionic-strength electrolyte solutions. At high electrolyte concentrations, it will shrink. This effect can be amplified by adjusting the “cocktail” of ionic species to maximize the shielding.

Most gels swell or shrink in fairly strict proportion to the thermodynamic quality of the solvent pervading them. Some, however, undergo a sudden change in dimensions in response to a relatively small change in solvent quality. For instance, if a partially charged polyacrylamide gel is immersed in a mixture of ethanol and water, it will shrink slightly as ethanol is added to the solution, until a point is reached at which the addition of a tiny increment of ethanol causes the gel to shrink to a few percent of its former volume. This phenomenon is analogous to the behav-

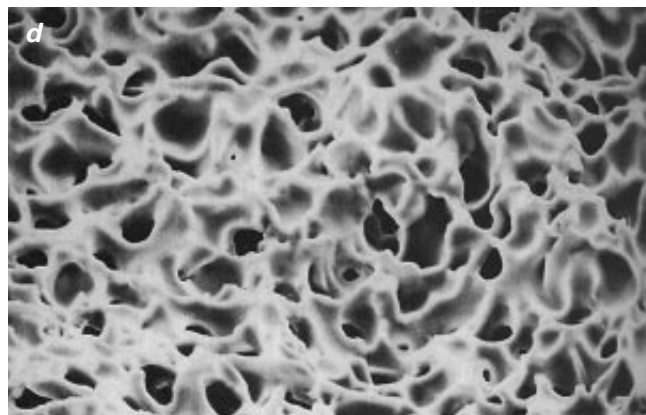
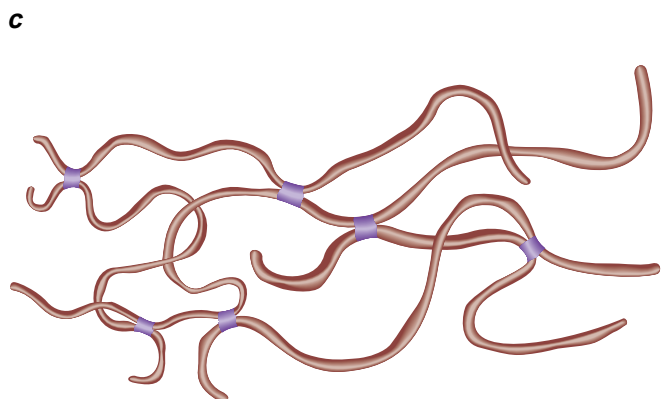
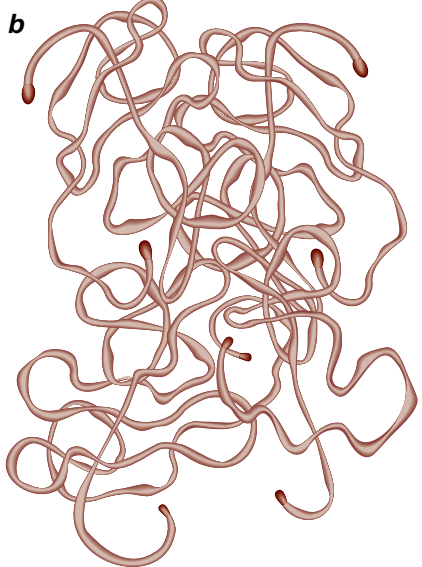
ior of a “critical fluid” (such as high-pressure carbon dioxide) that is close to the liquid-vapor phase transition. In that circumstance, a very small change in temperature or pressure can convert the liquid to its vapor phase or a vapor to liquid [see “Gels,” by Toyochi Tanaka; SCIENTIFIC AMERICAN, January 1981].

The polyacrylamide gel’s sudden shrinkage depends on a subtly snowballing interaction between the network’s affinity for solvent molecules and the elastic forces holding it together. The gel shrinks slowly at first, but as it expels more solvent, its molecular chains interact more strongly with one another; this interaction tends to expel more solvent, enhancing the interaction further until the network has shrunk into a tightly bunched state. If more water is added, thereby reducing the ethanol concentration, the same feedback loop plays itself out in reverse, and the gel regains its former size.

Depending on the precise structure of the gel, this simple story may be complicated by a number of factors—in some cases, the gel acts as a semipermeable membrane, and the interaction between solvent ions and charged sites on the gel dominates the transition between shrinking and swelling. In addition, the sharpness of the transition also depends on the stiffness of the gel’s polymer chains. Flexible molecules generally make for continuous transitions, as do those that are very stiff.

During the past decade, researchers around the world have developed new gels that swell or shrink in response to many different stimuli—temperature, pH or electric fields—depending on the chemical composition of the gel and solvent.

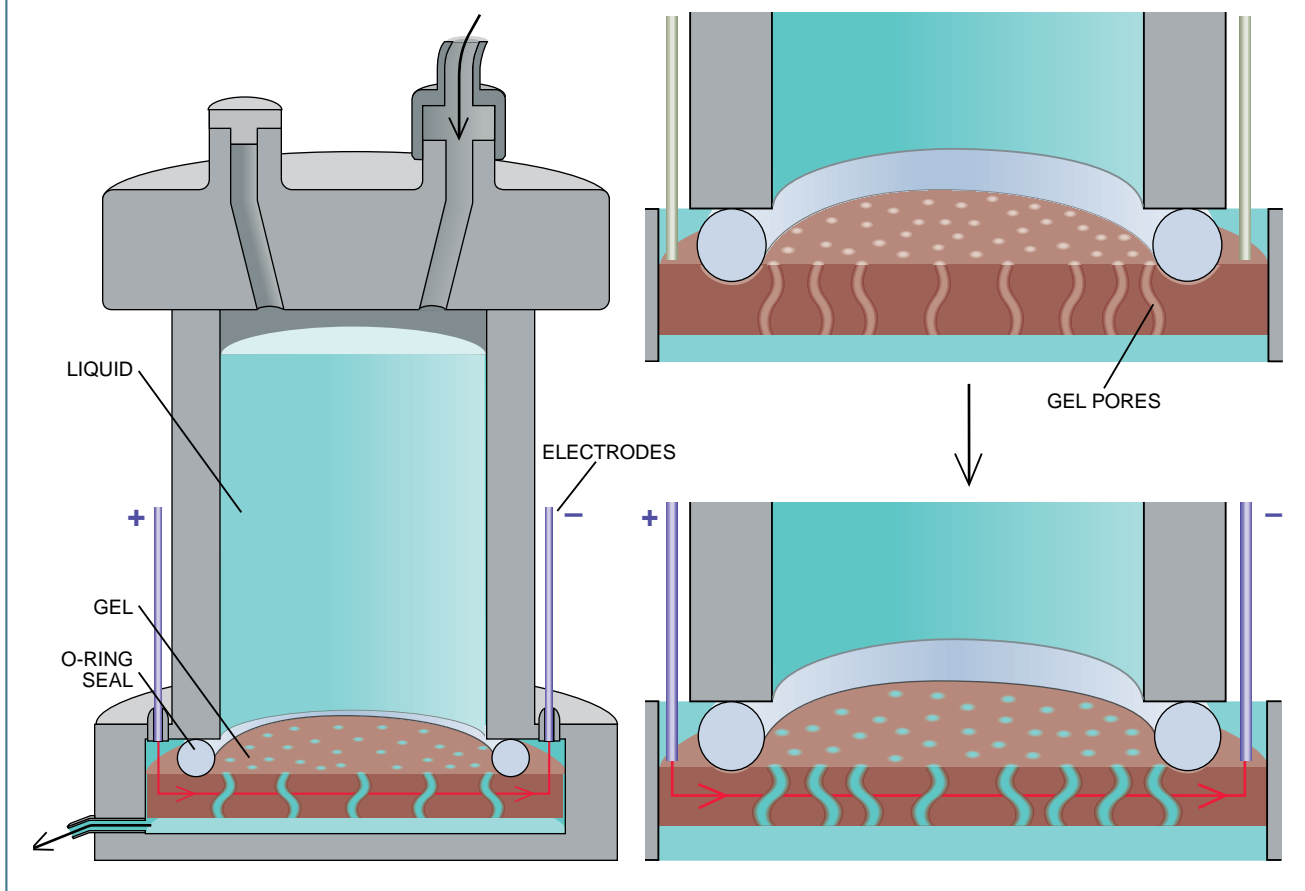
A gel derived from poly(*N*-isopropylacrylamide) can shrink to 30 percent of its original volume when heated above a critical temperature. A similar heat-sensitive gel, made by cross-linking an aqueous solution of poly(methyl



Gel Valves

A gel that expands or contracts in response to stimuli can control the flow of a liquid. In this prototype (left), a gel membrane is fixed to the bottom rim of a container. Under normal conditions, the gel is swollen and impervious to liquid (top right). When an electric current

passes through the gel, however, its larger internal pores open up (bottom right), allowing liquid to flow through. If the gel is very homogeneous, its pore size can be controlled so precisely that it passes small molecules while blocking large ones.



vinyl ether), undergoes rapid, reversible swelling and shrinking at 37 degrees Celsius. The gel fibrils shrink from 400 microns across at 20 degrees C to 200 microns at 40 degrees C.

An electric field of a mere half volt per centimeter will induce a similar shrinkage in a polyacrylamide gel immersed in acetone and water. Gel particles shrink or swell at a rate that depends on the current flowing through them and on the square of their size. In theory, an electric field of five volts per millimeter could shrink a collection of gel particles that are one micron in diameter to 4 percent of their original volume within one millisecond. This rapid response could make gels suitable for use as "muscle" for robots or other mechanical devices—or even for human prostheses.

The first widespread use of intelligent gels will probably not be as a replacement for muscles but rather as a highly

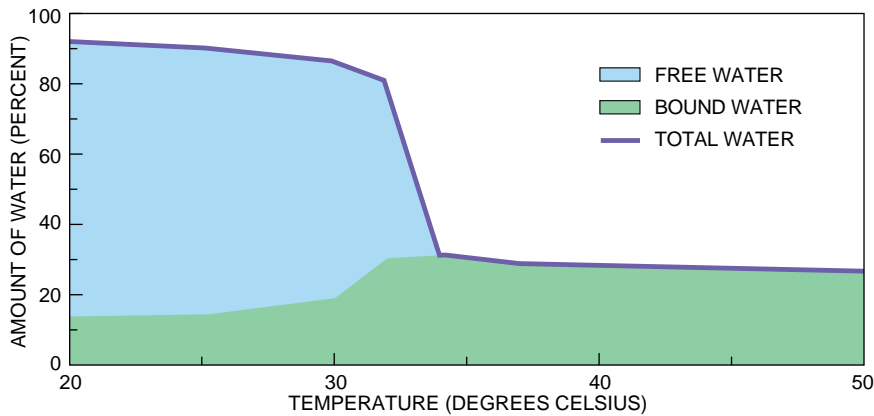
sophisticated descendant of the mustard plaster and other liniments. Delivering medication to organs that need it in required doses at any given moment has long been a crucial problem in medicine. In recent years, pharmaceutical manufacturers have employed semipermeable membranes to release drugs at a constant rate. Devices based on intelligent gels could improve these systems: the gel could sense conditions inside the body and vary the delivery rate to maintain an appropriate level of the drug in the bloodstream.

Ronald A. Siegel and his colleagues at the University of California at San Francisco have developed a simple, gel-based system for protecting acid-sensitive medications from the hostile environment of the stomach. Their gel shrinks when exposed to low pH, but in the more alkaline environment of the intestines it expands and becomes permeable, allowing the encapsulated drug

to diffuse under the proper conditions.

Researchers have developed gels that release drugs or biomolecules in response to variations in electric field. Insulin diffuses out of a gel made of weakly cross-linked polyelectrolyte gels when electric current is turned on, but the flow ceases immediately when it is turned off. A gel could be the basis of an implantable insulin pump with no moving parts. Indeed, a team from the University of Trondheim in Norway and the Veterans Administration Islet Transplant Center in Los Angeles has short-circuited the pump mechanism by enclosing insulin-producing cells in an alginate gel. They hope to begin human clinical trials later this year.

Gels that expand and contract under electrical control can also serve as general-purpose "chemical valves." The gel is made in the shape of a porous membrane, and its edges are fixed in place. When the gel contracts, the pores in



SOURCE: Liang C. Dong and Allan S. Hoffman, University of Washington

PHASE TRANSITION causes a thermosensitive polymer gel to collapse as its temperature increases. The graph above shows water content as a function of temperature. When the gel is cold, the swollen state is thermodynamically stable, but as the gel warms, interactions among the molecular chains stabilize the shrunken state instead.

the membrane perforce expand, permitting liquids and dissolved molecules to flow through the membrane. When the gel expands, the pores shrink, and flow stops. By maintaining the current at intermediate levels, researchers can control the precise dimensions of the pores and so determine what molecules can pass through. Controllable membranes are now being used to separate solvent mixtures containing molecules of different sizes.

Gels can also be used to recover large molecules from dilute aqueous solutions, a reaction that could be useful in manufacturing. Edward L. Cuss-

ler, Jr., and his colleagues at the University of Minnesota have found that a swollen gel preferentially absorbs water while excluding substances dissolved in it. They simply immerse the gel and let it soak up liquid, leaving the desired product behind. The gel can then be collapsed to expel most of the liquid and returned to the swollen state to continue the separation process.

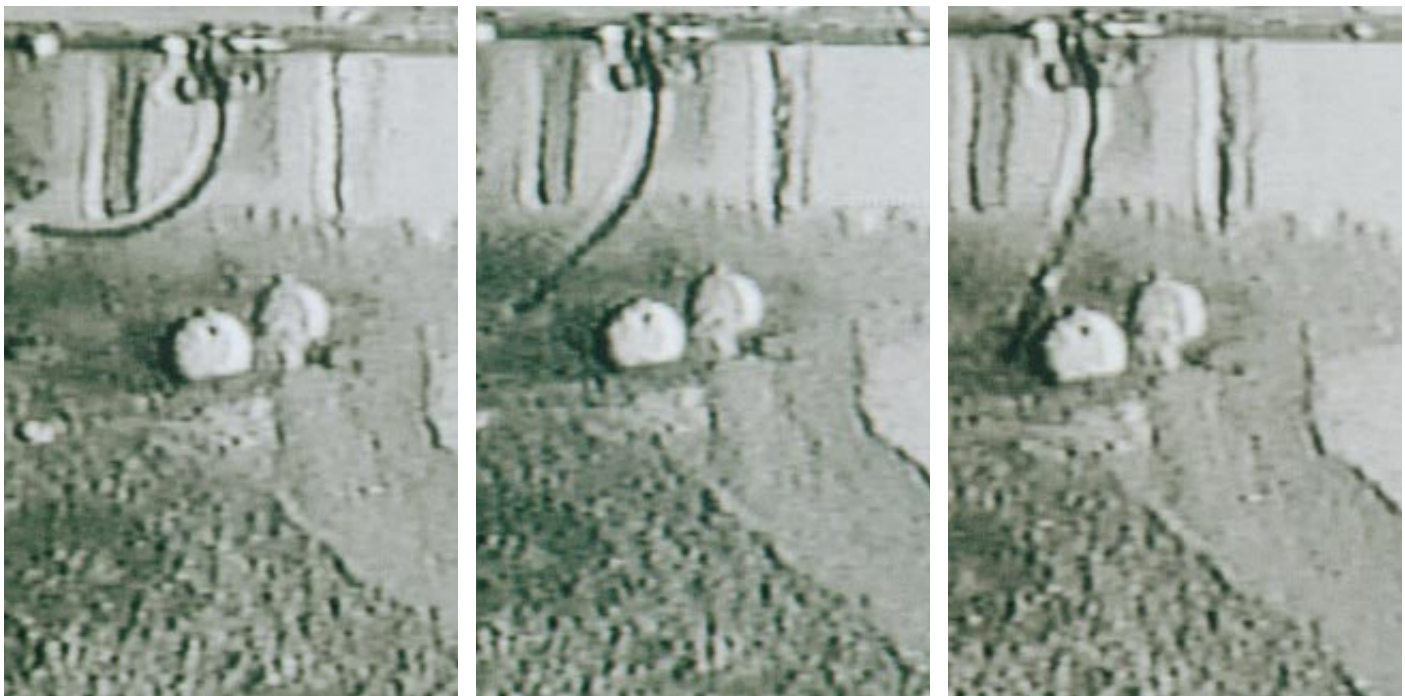
Although the swelling and shrinking of sessile gels has thus far proved most useful, investigators have been captivated for more than four decades by the idea of using gels to produce motive power. Aharon Katchalsky of the Weiz-

mann Institute of Science in Rehovot, Israel, and Werner Kuhn of the University of Basel built the first "chemomechanical" systems in 1950, altering the pH of their gels' acid environment to drive expansion and contraction [see "Muscle as a Machine," by A. Katchalsky and S. Lifson; SCIENTIFIC AMERICAN, March 1954].

Recently one of us (Osada) and his collaborators developed a new chemomechanical system that we call a "gel looper." We demonstrated an early version of the looper in December 1991 at the Second Polymer Gel Symposium and Robo-bug Fest in Tsukuba, Japan; eight other Japanese research organizations also exhibited chemomechanical moving devices.

The looper consists of a strip of gel that moves, inchwormlike, under the influence of an alternating electric field, along a supporting rod. The gel hangs from its rod by means of metal hooks while immersed in an aqueous solution containing surfactant molecules (essentially a sophisticated version of soapy water). Parallel electrodes above and below the looper control its motion.

When a voltage is applied across the electrodes, the surfactant molecules, which are positively charged, migrate toward the negative electrode. On their way they encounter the negatively charged surface of the gel and attach themselves to it, causing the gel to contract. The surfactant molecules alight preferentially on the side of the gel



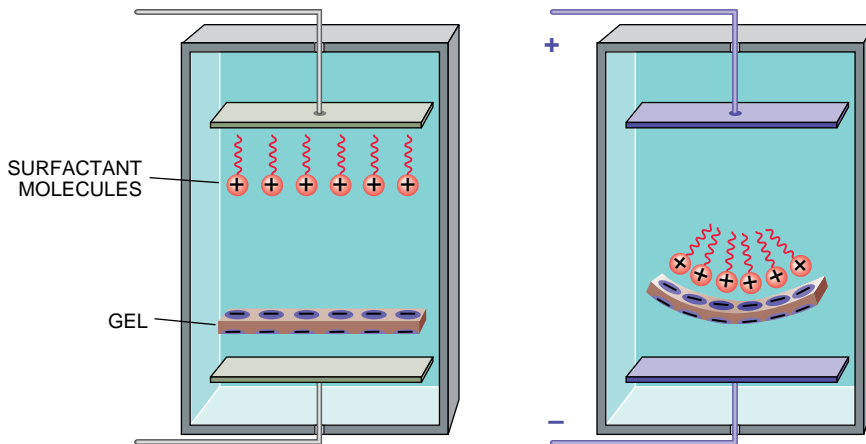
GEL GOLF demonstrates the ability of an intelligent gel to act on its surroundings. A strip of gel made of the same material

as the gel looper curls first one way and then the other under the influence of an electric field. During this transition, it

that faces the positive electrode; consequently, the gel contracts mainly on that side and curls up. When the polarity of the electric field is reversed, the surfactant molecules are released, and the gel straightens out [see illustration at right]. A sawtooth pattern on the top of the supporting rod ensures that the looper moves forward with each cycle of bending and straightening.

Simple though it is, the gel looper exhibits the essential characteristics that set "soft" chemomechanical systems apart from mechanical devices made of more rigid materials. In contrast to conventional motors and pumps, gels are gentle and flexible, and their movement is more reminiscent of muscle than is that of metallic machines. This pliant motion is usually seen only in biological systems such as the wings of birds, which reshape themselves continuously to maximize lift.

Because gels are soft, they can manipulate delicate materials without damaging them. Even more important, however, gels are soft with respect to their environments. Machines made of metal or silicon operate as closed systems. They do not adapt to changes in their operating conditions unless a separate sensor system or a human operator is at the controls. Gels, in contrast, are thermodynamically "open": they exchange chemicals with the solvent surrounding them and alter their molecular state in the process of accomplishing work. If free energy is added to the



ASYMMETRIC CONTRACTION drives both the gel looper and the gel "golf club" [see illustration below]. The curling of the gel is caused by surfactant molecules that drift through the surrounding solvent under the influence of an electric field. When the field is turned on, the positively charged heads of the molecules migrate to negatively charged sites on the gel. They exclude water molecules and cause one side of the gel to shrink. When the field is reversed, the surfactant molecules diffuse back out into the solution, and the gel uncurls.

solvent in the form of new chemicals, chemomechanical systems recover their original state without further intervention. We believe it will eventually be possible to make use of these properties to create self-sensing and self-regulating machines that respond intelligently to changes in their surroundings. Although soft machines will probably never replace hard ones, "wetware" may soon take its place next to hardware and software in the designer's lexicon.

FURTHER READING

GELS. Toyochi Tanaka in *Scientific American*, Vol. 244, No. 1, pages 124-138; January 1981.

PHYSICAL NETWORKS: POLYMERS AND GELS. Edited by W. Burchard and S. B. Ross-Murphy. Elsevier Science Publishers, 1990.

POLYMER GELS: FUNDAMENTALS AND BIOMEDICAL APPLICATIONS. D. DeRossi, K. Kajiwara, Y. Osada and A. Yamauchi. Plenum Press, 1991.



strikes a golf ball, propelling it down a slope. Although the "club" material is sturdy enough to strike the ball directly, it

must be submerged in liquid and so in actual use would probably be encased in a protective container.

The Power of Maps

The authoritative appearance of modern maps belies their inherent biases. To use maps intelligently, the viewer must understand their subjective limitations

by Denis Wood



The objectivity of modern maps of the world is so taken for granted that they serve as powerful metaphors for other sciences, on occasion even for scientific objectivity itself. The canonical history of Western cartography reinforces that assumption of objectivity. The history tells of a gradual progression from crude Medieval views of the world to depictions exhibiting contemporary standards of

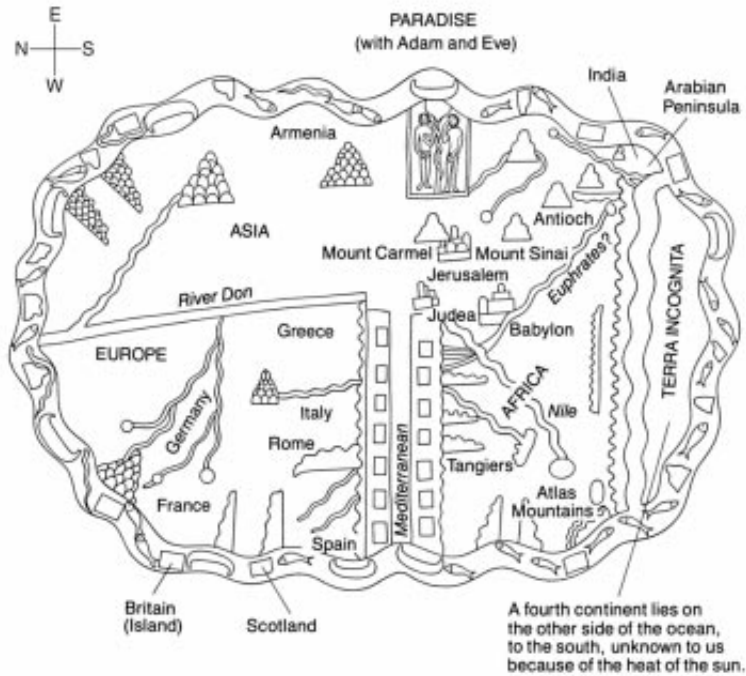
IMAGE OF EARTH was compiled by Tom Van Sant from 37 million satellite-derived pixels. Van Sant manipulated the pixels to remove clouds and to emphasize certain natural features. NASA's photograph from space (*right*) shows the earth with its customary veil of clouds.



precision. In actuality, all maps incorporate assumptions and conventions of the society and the individuals who create them. Such biases seem blatantly obvious when one looks at ancient maps but usually become transparent when one examines maps from modern times. Only by being aware of the subjective omissions and distortions inherent in maps can a user make intelligent sense of the information they contain.

The putative history of cartography typically begins in earnest at the time of the Egyptian and Babylonian mapmakers. The scene quickly shifts to ancient Greek and Roman contributions, followed by an acknowledgment of those of the Arabs during the Middle Ages. Mapmaking in Medieval Europe has been long regarded as the nadir of the craft. From the 15th century forward, cartography smoothly advanced, culminating in present maps that benefit from sophisticated optics, satellite imaging and digital processing.

Given this widely accepted history, it may be surprising to learn that few objects that can be unambiguously identified as maps remain from its early years. Several pieces of textual evidence of Greek mapmaking exist, but no actual maps. Except for Medieval copies of Roman itineraries, no Roman world maps are known, despite the elaborate instructions for producing them in Ptolemy's *Geography*. Indeed, historians know of no maps of the world that definitely predate the Middle Ages. There are few enough of the Medieval maps that are usually taken as the baseline from which to measure the heights to which cartography has risen.



EARLY WORLD MAPS clearly reflect the assumptions of their creators. The Beatus map (*opposite page, with modern explanation above it*) portrayed the 10th-century Christian world-view and thus includes Paradise in the east at the top. By the 15th century Ptolemaic maps (*top right*) had moved toward a depiction of the world that would better serve the interests of the nascent commerce centered on Europe. Nineteenth-century maps, such as one from Cary's atlas (*middle right*), embodied these commercial and political concerns. Oriental mapmakers, on the other hand, placed China at the center of the world, as in the example from 19th-century Korea (*bottom right*).

These Medieval world maps, or *mappaemundi*, come in several varieties. The Beatus maps, drawn to accompany the *Commentary on the Apocalypse of Saint John* by the abbot Beatus of Liebana, are illustrative. Dating from the 10th century and later, they can be traced back to a lost prototype of the eighth century. The Beatus maps are rectangular, but like other *mappaemundi* they are orientated so that east—where Paradise can be found in a square vignette—sits at the top. The three continents peopled by the sons of Noah are arranged as in most *mappaemundi*: Europe appears in the lower left, Africa in the lower right, Asia above. The Beatus maps also include a fourth continent, a *terra incognita*, required by the maps' evangelistic context (the apostles were to preach the gospel in the "four corners of the earth").

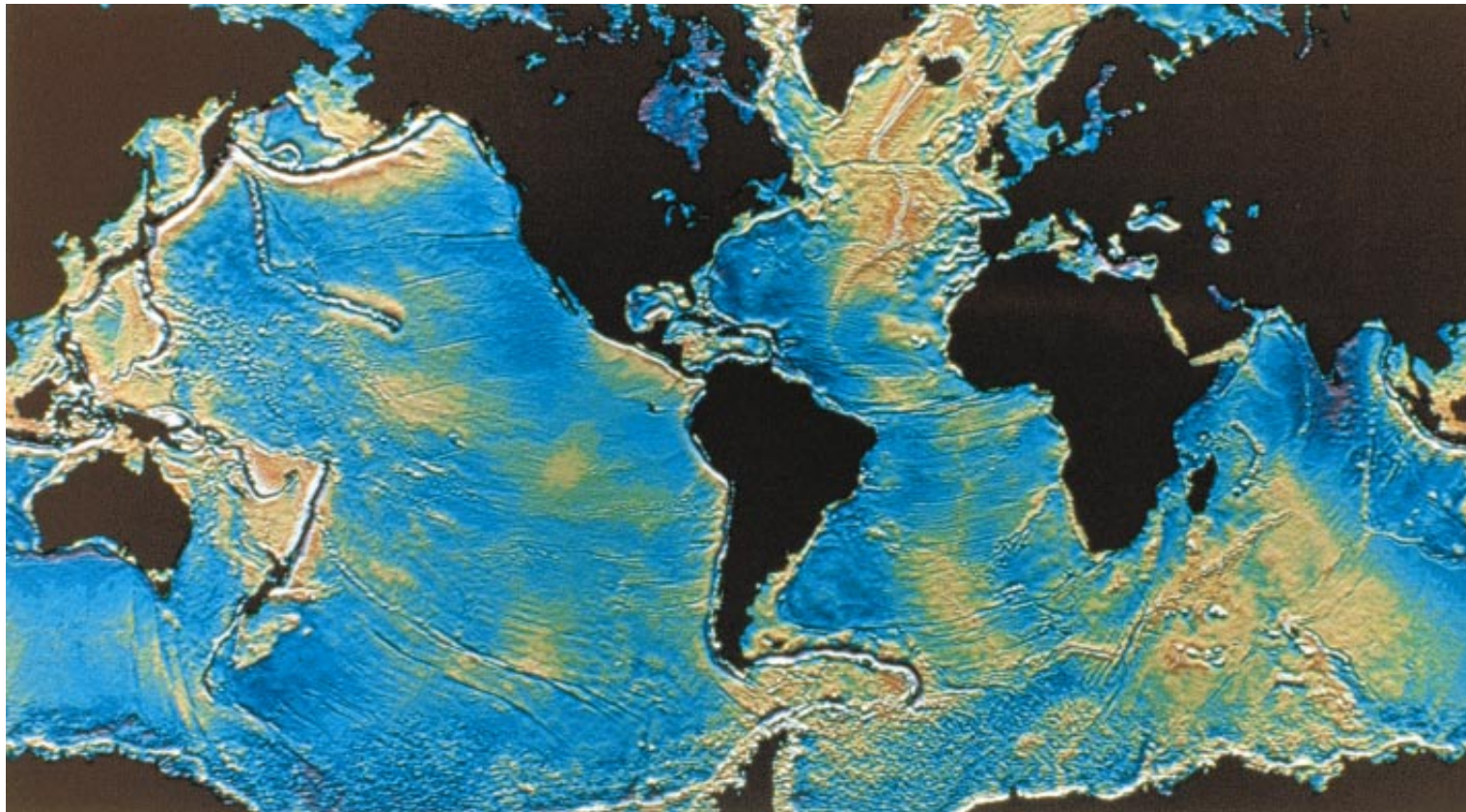
First and foremost, the Beatus maps are visions of the earth as a stage for the Christian history of the world; physical geographic accuracy is a secondary concern. Compared to a Rand McNally atlas, the Beatus maps seem quaintly erroneous, but it is meaningless to declare that more recent maps convey a "truer" sense of the world. Given their spiritual function, the Beatus maps are thoroughly accurate.

The rediscovery of Ptolemy's texts on the making of maps in the wake of the Crusades led to the rise of a more modern-looking style of mapmaking during the 14th century. Mapmakers, following Ptolemy's instructions and using his data, produced maps that displayed north at the top and that showed locations fixed in a graticule of latitude and longitude. Although they bear a superficial resemblance to those of our own time, the Ptolemaic maps remained deeply marked by the conventions of Medieval *mappaemundi*. They continued to show the *terra incognita* to the south, made extensive use of decorative pictures (as of the Twelve Winds) and employed traditional chromatic conventions (for instance, the Red Sea might be painted red). Yet the Ptolemaic maps did herald a shift away from a commitment to interpreting the world through the lens of the Bible, in favor of more practical concerns better able to serve the nascent world commerce focused on Europe.

Eighteenth- and 19th-century maps embodied the commercial and political interests of European nation-states. In the increasingly common atlases of the 19th century, a distinctly Eurocentric world appeared. The borders, markings, illustrations and notations on these maps graphically express the European states' political, commercial and scientific interests; colonial possessions were prominently displayed. These maps build on the Ptolemaic tradition and establish a new set of conventions. North is at the top, zero degrees longitude runs through Greenwich, England, and the maps are centered on Western Europe, North America or the North Atlantic. The resulting configuration has become so familiar that few people notice just how arbitrary it is.

As mapmaking transformed itself into the scientific art of cartography, people found it more and more difficult to accept that maps are windows onto the societies that shaped them as much as they are windows onto the world itself. In the West, the dominant positivist epistemology and wide-





spread belief in continuous material progress encouraged historians and laypeople alike to castigate non-Western maps as primitive and to denigrate earlier maps as products of a barbaric, but superseded, past. By logical extension, the present maps must be the best, most accurate and most objective ones ever produced. That perception has been powerfully reinforced by the correspondence, indeed the confusion, between maps and images from earth-orbiting satellites.

Consider, for instance, the “GeoSphere” map created by Tom Van Sant of GeoSphere Project in Santa Monica, Calif., with technical assistance from Lloyd Van Warren of the Jet Propulsion Laboratory in Pasadena, Calif., which completely blurs the line between cartography and satellite imagery. Van Sant and Van Warren crafted the map by sifting through millions of pixels that had been transmitted from *TIROS-N* satellites operated by the National Oceanic and Atmospheric Administration. The researchers rejected images where the presence of cloud cover obscured the ground, leaving an exposed earth whose continental outlines have been defined by an impartial, electronic eye.

The GeoSphere map embodies its own ideological commitments every bit as much as do the *mappaemundi*, the Ptolemaic maps and the atlases of the 19th century. Like many of his predecessors, Van Sant opted to run the equator across

DENIS WOOD is a geographer who studies relations between humans and their environment, especially as manifested in maps, drawings and the behavior of children. He earned a doctorate in geography from Clark University and is now professor of design at the School of Design at North Carolina State University. He was co-curator, with Lucy Fellowes, of “The Power of Maps” exhibition at the Cooper-Hewitt, National Museum of Design, Smithsonian Institution, in New York City.

the heart of his map, to put the Atlantic in its middle and to orient it north up. Furthermore, Van Sant openly acknowledges that he filtered and modified the satellite data in a number of deliberately subjective ways.

The exclusion of clouds itself omits a distinctive dominant feature of the earth as seen from space. In places where cloud-free images were unavailable, the researchers artificially subtracted the clouds, pixel by pixel. For low and moderate latitudes, the map’s creators selected images showing the most pronounced summer vegetation; for high latitudes and altitudes, they chose images that highlighted winter snowfall. River systems have been thickened to make them visible, and false color was applied—ironically, to make the vegetative cover appear more real. All these decisions serve a useful purpose in that they emphasize certain aspects of the earth, thereby making the Van Sant map more useful and easier to read than it would be otherwise. But one should keep in mind that the absence of clouds, the extent of the vegetative cover, the visibility of the rivers and all the colors seen on the map are expressions of the mapmakers’ vision, not attributes intrinsic to the earth itself.

The lighting and coloration of the Van Sant map emphasize the natural aspects of the earth (vegetation, desert, snow and ice) while omitting the landmarks of human society. W. T. Sullivan of the University of Washington created a nearly antithetical map that focuses instead on the impact of humans. Sullivan’s map, which was published by the Hansen Planetarium, also makes use of cloud-free images transmitted by *TIROS-N* satellites, but in this case, the images were taken on the earth’s night side. Simply by switching to a nighttime perspective, Sullivan eliminated the outlines of the ocean and continents, leaving only lights of cities and fires to serve as ghostly proxies of the sites of human activity.

A third map, also derived from satellite data, presents another scientifically accurate yet completely different view of

MODERN MAPS produce their own distortions to emphasize their purpose. In William F. Haxby's "Gravity Fields of the World's Oceans" (*opposite page*), the continents fade to insignificance. Conservation International's "Human Disturbance of Ecosystems" (*top right*) distorts the shape of the continents to preserve their relative sizes. W. T. Sullivan's "Earth at Night" (*middle right*), made from satellite images taken on the earth's night side, highlights humanity's presence on the planet. A fanciful map, "McArthur's Universal Corrective Map" (*bottom right*), places south at top, to reveal an Australian view of the world.

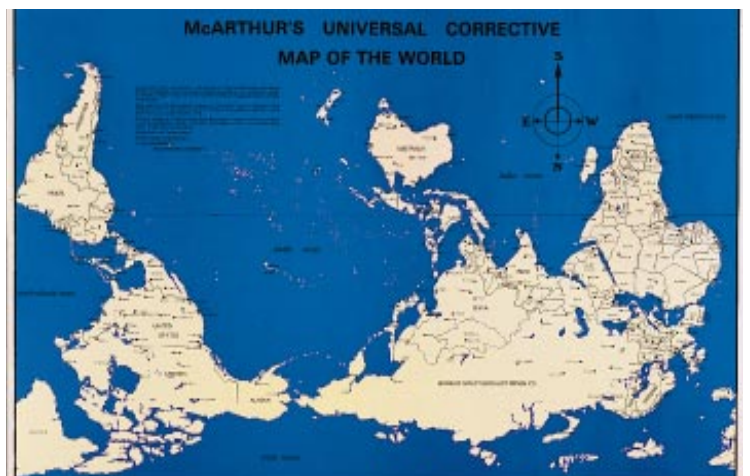
the world. Made by William F. Haxby, while at Lamont Doherty Geological Observatory of Columbia University, the map illustrates anomalies in the gravity field (that is, slight variations greater than and less than the mean gravitational tug) of the ocean floor, which he finds by means of radar altimeter readings of the sea surface made by the satellite *SEASAT*. Haxby ignores the physical surface of the earth entirely. The gravity anomaly map seems to show the topography of the seafloor, an illusion heightened by the way Haxby has lighted the ridges and the valleys as if by a late afternoon sun. That kind of depiction makes the data more comprehensible, but it also holds the potential to mislead the careless viewer.

Haxby's map clearly proclaims its subjective character: ocean gravity anomalies are plotted in vibrant false colors, and continents are blacked out, rendered unimportant. The fact that the Haxby map explicitly excludes information about certain parts of the earth enhances its maplike qualities. One must be more careful in interpreting maps that seem to show what the earth would actually look like to an outside viewer. Those maps are often described as "portraits" or "views," which bolsters that false impression.

Every flat map is subjective in that it cannot help distorting the sizes and shapes of the earth's features. Mapmakers deal with that limitation in many different ways. The Van Sant map displays relative accuracy of shape at the expense of representing relative size, an approach appropriate for its goal. Conservation International, in contrast, has opted for an equal-area projection that preserves the relative sizes of the continents at the cost of distorting their shapes. Conservation International is committed to saving tropical rain forests, so it has selected a projection that does not exaggerate the size of Europe and North America at the expense of tropical Africa, Asia and South America as do so many common projections. Like the Van Sant map, the Conservation International map uses coloration to enhance certain natural aspects of the world. For its purposes, however, Conservation International uses obviously false colors that focus attention specifically on threatened regions of rain forest.

Each of the modern maps discussed here makes its own claim to special accuracy, and yet the results could hardly look more dissimilar. This is the contradiction maps present: a claim to represent objectively a world they can only subjectively present, a claim made to win acceptance for a view of the world whose utility lies precisely in its partiality.

Given that the usefulness of maps derives from their bias and subjectivity, these are qualities to be highlighted and celebrated. Maps need to be explicit about the choices they represent among potential sets of data and the ways of presenting them. Titles should make clear the nature of the limitations and distortion—the advantages and emphases—inherent in the map. And viewers should be better educated about what maps can and cannot do. The future of cartography lies in transcending the dichotomy between the utility of the subjective and the authority of the objective. Beyond lie maps that will be ever more useful because they will be more open about their real relation to the world.



FURTHER READING

- THE HISTORY OF CARTOGRAPHY, Vol. 1: CARTOGRAPHY IN PRE-HISTORIC, ANCIENT, AND MEDIEVAL EUROPE AND THE MEDITERRANEAN. Edited by J. B. Harley and David Woodward. University of Chicago Press, 1987.
- MAPPING THE NEXT MILLENNIUM: THE DISCOVERY OF NEW GEOGRAPHIES. Stephen S. Hall. Random House, 1992.
- THE POWER OF MAPS. Denis Wood. Guilford Press, 1992.

The Neurobiology of Fear

Researchers are beginning to tease apart the neurochemical processes that give rise to different fears in monkeys. The results may lead to new ways to treat anxiety in humans

by Ned H. Kalin

Over the years, most people acquire a repertoire of skills for coping with a range of frightening situations. They will attempt to placate a vexed teacher or boss and will shout and run when chased by a mugger. Some individuals, though, become overwhelmed in circumstances others would consider only minimally stressful: fear of ridicule might cause them to shake uncontrollably when called on to speak in a group, or terror of strangers might lead them to hide at home, unable to work or shop for groceries. Why do certain people fall prey to excessive fear?

At the University of Wisconsin at Madison, my colleague Steven E. Shelton and I are addressing this problem by identifying specific brain processes that regulate fear and its associated behaviors. Despite the availability of noninvasive imaging techniques, such information is still extremely difficult to obtain in humans. Hence, we have turned our attention to another primate, the rhesus monkey (*Macaca mulatta*). These animals undergo many of the same physiological and psychological developmental stages as humans do, but in a more compressed time span. As we gain more insight into the nature and operation of neural circuits that

modulate fear in monkeys, it should be possible to pinpoint the brain processes that cause inordinate anxiety in people and to devise new therapies to counteract that anxiety.

Effective interventions would be particularly beneficial if they were applied at an early age. Growing evidence suggests overly fearful youngsters are at high risk for later emotional distress. Jerome Kagan and his colleagues at Harvard University have shown, for example, that a child who is profoundly shy at the age of two years is more likely than a less inhibited child to suffer from anxiety and depression later in life.

This is not to say these ailments are inevitable. But it is easy to see how excessive fear could contribute to a lifetime of emotional struggle. Consider a child who is deeply afraid of other children and is therefore taunted by them at school. That youngster might begin to feel unlikable and, in turn, to withdraw further. With time the growing child could become mired in a vicious circle leading to isolation, low self-esteem, underachievement and the anxiety and depression noted by Kagan.

There are indications that unusually fearful children might also be prone to physical illness. Many youngsters who become severely inhibited in unfamiliar situations chronically overproduce stress hormones, including the adrenal product cortisol. In times of threat, these hormones are critical. They ensure that muscles have the energy needed for "fight or flight." But some evidence indicates long-term elevations of stress hormones may contribute to gastric ulcers and cardiovascular disease.

Further, through unknown mechanisms, fearful children and their families are more likely than others to suffer from allergic disorders. Finally, in rodents and nonhuman primates, persistent elevation of cortisol has been shown to increase the vulnerability of neurons in the hippocampus to damage by other substances; this brain region is involved in memory, motivation

and emotion. Human neurons probably are affected in a similar manner, although direct evidence is awaited.

When we began our studies about 10 years ago, Shelton and I knew we would first have to find cues that elicit fear and identify behaviors that reflect different types of anxiety. With such information in hand, we could proceed to determine the age at which monkeys begin to match defensive behaviors selectively to specific cues. By also determining the parts of the brain that reach maturity during the same time span, we could gain clues to the regions that underlie the regulation of fear and fear-related behavior.

The experiments were carried out at the Wisconsin Regional Primate Research Center and the Harlow Primate Laboratory, both at the University of Wisconsin. We discerned varied behaviors by exposing monkeys between six and 12 months old to three related situations. In the alone condition, an animal was separated from its mother and left by itself in a cage for 10 minutes. In the no-eye-contact condition, a person stood motionless outside the cage and avoided looking at the solitary infant. In the stare condition, a person was again present and motionless but, assuming a neutral expression, peered directly at the animal. These conditions are no more frightening than those primates encounter frequently in the wild or those human infants encounter every time they are left at a day-care center.

In the alone condition, most monkeys became very active and emitted frequent "coo" calls. These fairly melodious sounds are made with pursed lips.

RHESUS MONKEY REGISTERS ALARM (right) as another monkey (left) approaches her baby. The mother's fear is evident in her "threat" face: the open mouth and piercing stare serve to intimidate would-be attackers.

NED H. KALIN, a clinician and researcher, is professor of psychiatry and psychology and chairman of the department of psychiatry at the University of Wisconsin-Madison Medical School. He is also a scientist at the Wisconsin Regional Primate Research Center and the Harlow Primate Laboratory at the university. He earned his B.S. degree in 1972 from Pennsylvania State University and his M.D. in 1976 from Jefferson Medical College in Philadelphia. Before joining his current departments, he completed a residency program in psychiatry at Wisconsin and a postdoctoral fellowship in clinical neuropharmacology at the National Institute of Mental Health.

They start at a low pitch, become higher and then fall. More than 30 years ago Harry F. Harlow, then at Wisconsin, deduced that when an infant monkey is separated from its mother, its primary goal is affiliative; that is, it yearns to regain the closeness and sense of security provided by nearness to the parent. Moving about and cooing help to draw the mother's attention.

By contrast, in the more frightening no-eye-contact situation, the monkeys reduced their activity greatly and sometimes "froze," remaining completely still for prolonged periods. When an infant spots a possible predator, its goal shifts

from attracting the mother to becoming inconspicuous. Inhibiting motion and freezing—common responses in many species—reduce the likelihood of attack.

If the infant perceives that it has been detected, its aim shifts again, to warding off an attack. And so the stare condition evoked a third set of responses. The monkeys made several hostile gestures, among them "barking" (forcing air from the abdomen through the vocal cords to emit a growllike sound), staring back, producing so-called threat faces [see *illustration below*], baring their teeth and shaking the cage. Sometimes the animals mixed the threatening displays

with submissive ones, such as fear grimaces, which look something like wary grins, or grinding of the teeth. In this condition, too, cooing increased over the amount heard when the animals were alone. (As will be seen, we have recently come to think the cooing displayed in the stare condition may serve a somewhat different function than it does in the alone situation.)

Monkeys, by the way, are not unique in becoming aroused by stares and in using them reciprocally to intimidate predators. Animals as diverse as crabs, lizards and birds all perceive staring as a threat. Some fishes and insects have



THREE EXPERIMENTAL CONDITIONS elicit distinct fear-related behaviors in rhesus monkeys older than about two months. When isolated in a cage (*left*), youngsters become quite active and emit “coo” sounds to attract their mothers. If a human appears but avoids eye contact (*center*), the monkeys try to evade discovery, such as by staying completely still (freezing) or hiding behind their food bin. If the intruder stares at the animals (*right*), they become aggressive.

evolved protective spots that resemble eyes; these spots either avert attacks completely or redirect them to nonvital parts of the body. In India, fieldworkers wear face masks behind their heads to discourage tigers from pouncing at their backs. Studies of humans show that we, too, are sensitive to direct gazes: brain activity increases when we are stared at, and people who are anxious or depressed tend to avoid direct eye contact.

Having identified three constellations of defensive behaviors, we set about determining when infant monkeys first begin to apply them effectively. Several lines of work led us to surmise that the ability to make such choices emerges sometime around an infant's two-month birthday. For instance, rhesus mothers generally permit children to venture off with their peers at that time, presumably because the adults are now confident that the infants can protect themselves reasonably well. We also knew that by about 10 weeks of age infant monkeys respond with different emotions to specific expressions on the faces of other monkeys—a sign that at least some of the innate wiring or learned skills needed to discriminate threatening cues are in place.

To establish the critical period of development, we examined four groups of monkeys ranging in age from a few days to 12 weeks old. We separated the babies from their mothers and let them acclimate to an unfamiliar cage. Then we exposed them to the alone, no-eye-contact and stare conditions. All sessions were videotaped for analysis.

We found that infants in the youngest group (newborns to two-week-olds) engaged in defensive behaviors. But they lacked some motor coordination and seemed to act randomly, as if they were oblivious to the presence or gaze of the human intruder. Babies in our two intermediate-age groups had good motor control, but their actions seemed unrelated to the test condition. This finding meant motor control was not

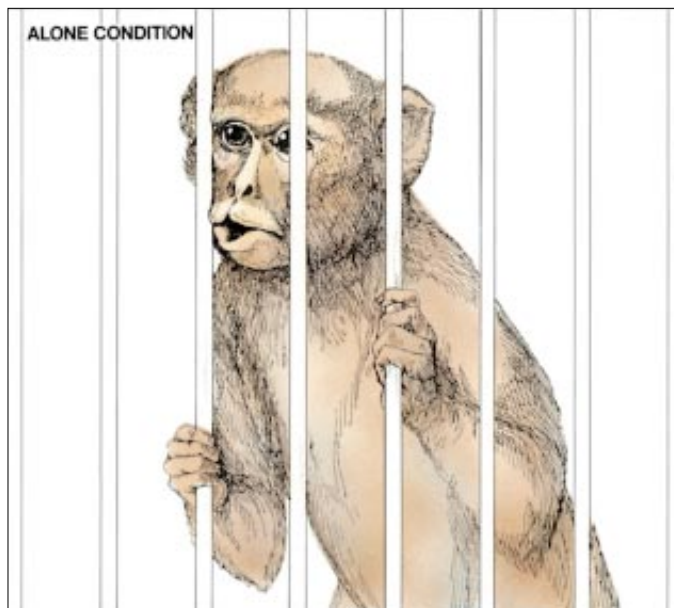
the prime determinant of selective responding.

Only animals in our oldest group (nine- to 12-week-olds) conducted themselves differently in each situation, and their reactions were both appropriate and identical to those of mature monkeys. Nine to 12 weeks, then, is the critical age for the appearance of a monkey's ability to adaptively modulate its defensive activity to meet changing demands.

Studies by other workers, who primarily examined rodents, suggested that three interconnected parts of the brain regulate fearfulness. We suspected that these regions become functionally mature during the nine- to 12-week period and thus give rise to the selective reactivity we observed. One of these regions is the prefrontal cortex, which takes up much of the outer and side areas of the cerebral cortex in the frontal lobe [see illustration on page 98]. A cognitive and emotional area, the prefrontal cortex is thought to participate in the interpretation of sensory stimuli and is probably a site where the potential for danger is assessed.

The second region is the amygdala, a part of a primitive area in the brain called the limbic system (which includes the hippocampus). The limbic system in general and the amygdala in particular have been implicated in generating fear.

The final region is the hypothalamus. Located at the base of the brain, it is a constituent of what is called the hypothalamic-pituitary-adrenal system. In response to stress signals from elsewhere in the brain,

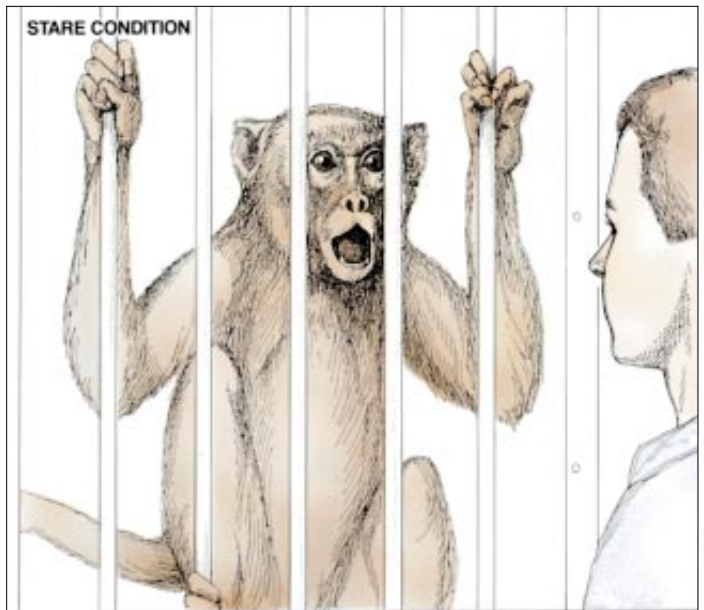
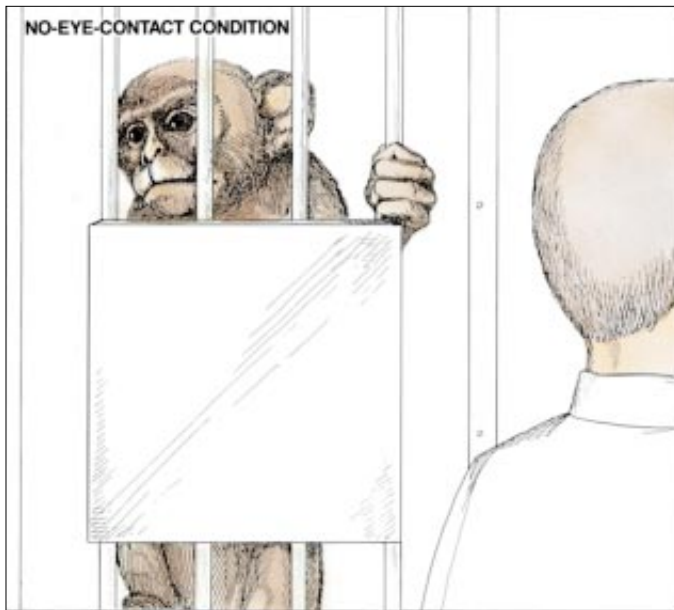


such as the limbic system and other cortical regions, the hypothalamus secretes corticotropin-releasing hormone. This small protein spurs the pituitary gland, located just below the brain, to secrete adrenocorticotropic hormone (ACTH). ACTH, in turn, prods the adrenal gland to release cortisol, which prepares the body to defend itself.

In neuroanatomic data collected in other laboratories, we found support for our suspicion that maturation of these brain regions underlies selective responding in the nine- to 12-week period. For instance, during this time the formation of synapses (contact points between neurons) has been shown to reach its peak in the prefrontal cortex and the limbic system (including the amygdala), as well as in the motor and visual cortices and other sensory areas. Patricia S. Goldman-Rakic of Yale Uni-

TYPICAL BEHAVIORS induced by the alone, no-eye-contact and stare conditions in the laboratory—such as cooing (*left*), freezing (*center*) and hostile display of the teeth (*right*)—are also seen in frightened infants and adults living in the wild. In this case, the setting is Cayo Santiago, an island off the mainland of Puerto Rico.





versity has also established that as the prefrontal cortex matures in rhesus monkeys, the ability to guide behavior based on experience emerges. This skill is necessary if one is to contend successfully with danger.

Maturation of the prefrontal cortex likewise seems important for enabling humans to distinguish among threatening cues. Harry T. Chugani and his co-workers at the University of California at Los Angeles have shown that activity in the prefrontal cortex increases when human offspring are seven to 12 months of age. During this span—which appears to be analogous to the time when monkeys begin to respond selectively to fear—children begin to display marked fear of strangers. They also become adept at what is called social referencing; they regulate their level of fear based on interpreting the expres-

sions they observe on a parent's face.

But what of the hypothalamus, the third brain region we assumed could participate in regulating fear-related behavior? Published research did not tell us much about its development or about the development of the complete hypothalamic-pituitary-adrenal system in monkeys. Our own investigations, however, revealed that the full system matures in parallel with that of the prefrontal cortex and the limbic system.

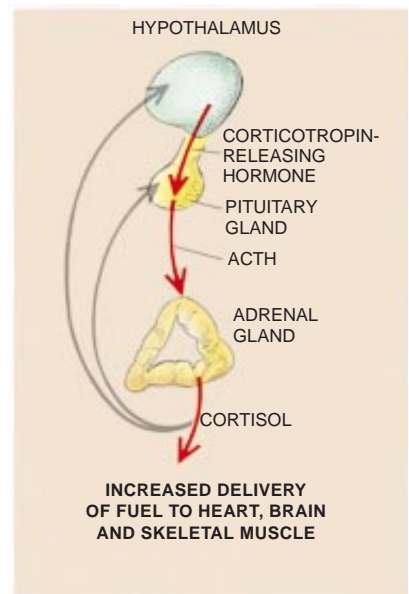
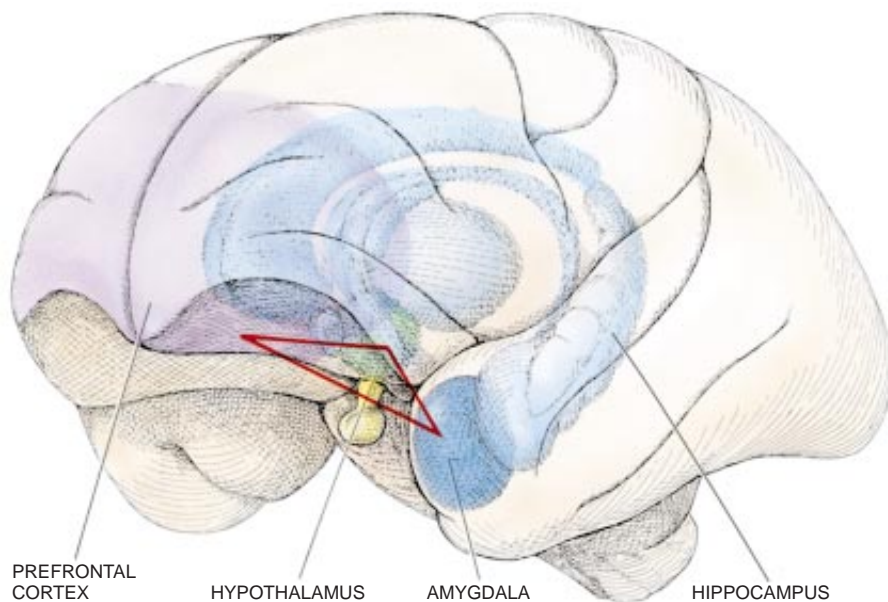
In these studies, we used the pituitary hormone ACTH as a marker of the system's function. We again examined four groups of infants aged a few days to 12 weeks. From each subject, we measured ACTH levels in blood drawn while the youngster was with its mother. This reading provided a baseline. We also measured ACTH levels in

blood samples obtained 20 minutes after the infant was separated from its parent. Hormonal levels rose in all four age groups during separation, but they jumped profoundly only in the oldest (nine- to 12-week-old) monkeys.

The relatively weak response in the younger animals, particularly in those under two weeks old, is consistent with findings in rat pups, whose stress hormone response is also blunted during the first two weeks of life. The development of the rodent and primate stress hormone system may well be delayed during early life to protect young neurons from the potentially damaging effects of cortisol.

Assured that the hypothalamic-pituitary-adrenal system becomes functionally mature by nine to 12 weeks, we pressed the inquiry forward to determine whether levels of cortisol and





THREE BRAIN REGIONS that are interconnected by neural pathways (shown schematically by red lines) are critically important in regulating fear-related behaviors. The prefrontal cortex (purple) participates in assessing danger. The amygdala (dark blue) is a major constituent of the emotion-

producing limbic system (light blue). And the hypothalamus (green), in response to signals from the prefrontal cortex, amygdala and hippocampus, directs the release of hormones (red arrows in box) that support motor responses to perceived threats. (Gray arrows represent inhibitory activity by cortisol.)

ACTH might partly account for individual differences in defensive behavior. We were also curious to know whether the responses of the infants resembled those of their mothers; a correspondence would indicate that further analyses of mothers and their infants could help reveal the relative contributions of inheritance and learning to fearfulness. We mainly examined the propensity for freezing, which we had earlier found was a stable trait in our subjects.

In one set of studies, we measured baseline levels of cortisol in monkeys four months to a year old and then observed how much time the youngsters froze in the no-eye-contact condition. Monkeys that started off with relatively low levels of cortisol froze for shorter periods than did their counterparts with higher cortisol levels—a pattern we also noted in separate studies of adult females. In other studies, we observed that as youngsters pass through their first year of life, they become progressively like their mothers hormonally and behaviorally. By the time infants are about five months old, their stress-induced rises in ACTH levels parallel those of the mothers. And by the time they are a year old, the duration of freezing in the no-eye-contact condition also corresponds to that of the mother.

Strikingly, some of these results echoed those obtained in humans. Extremely inhibited children often have parents who suffer from anxiety. Moreover, Kagan and his colleagues have found that basal cortisol levels are

predictive of such children's reaction to a frightening situation. They measured cortisol concentrations in saliva of youngsters at home (where they are presumably most relaxed) and then observed the children confronting an unfamiliar situation in the laboratory; high basal cortisol levels were associated with greater inhibition in the strange setting.

These similarities between humans and monkeys again imply that monkeys are reasonable models of human emotional reactivity. The link between basal cortisol levels and duration of freezing or inhibition suggests as well that levels of stress hormones influence how appropriately animals and people behave in the face of fear. (This effect may partly be mediated by the hippocampus, where the concentration of cortisol receptors is high.) And the likeness of hormonal and behavioral responses in mothers and infants implies that genetic inheritance might predispose some individuals to extreme fearfulness, although we cannot rule out the contribution of experience.

No one can yet say to what extent the activity of the hypothalamic-pituitary-adrenal system controls, and is controlled by, other brain regions that regulate the choice of defensive behavior. We have, however, begun to identify distinct neurochemical circuits, or systems, in the brain that affect different behaviors. The two systems we have studied most intensely

seemed at first to have quite separate functions. But more recent work implies that the controls on defensive behavior are rather more complicated than the original analyses implied.

We gathered our initial data three years ago by treating six- to 12-month-old monkeys with two different classes of neuroactive chemicals—opiates (morphinelike substances) and benzodiazepines (chemicals that include the anti-anxiety drug diazepam, or Valium). We chose to look at opiates and benzodiazepines because neurons that release or take up those chemicals are abundant in the prefrontal cortex, the amygdala and the hypothalamus. The opiates are known to have natural, or endogenous, counterparts, called endorphins and enkephalins, that serve as neurotransmitters; after the endogenous chemicals are released by certain neurons, they bind to receptor molecules on other nerve cells and thereby increase or decrease nerve cell activity. Receptors for benzodiazepines have been identified, but investigators are still trying to isolate endogenous benzodiazepinelike molecules.

Once again, our subjects were exposed to the alone, no-eye-contact and stare conditions. We delivered the drugs before the infants were separated from their mothers and then recorded the animals' behavior. Morphine decreased the amount of cooing normally displayed in the alone and stare conditions. Conversely, cooing was increased by naloxone, a compound that binds to

opiate receptors but blocks the activity of morphine and endogenous opiates. Yet morphine and naloxone had no influence on the frequency of stare-induced barking and other hostile behaviors, nor did they influence duration of freezing in the no-eye-contact situation. We concluded that opiate-using neural pathways primarily regulate affiliative behaviors (such as those induced by distress over separation from the mother), but those pathways seem to have little power over responses to direct threats.

The benzodiazepine we studied—diazepam—produced a contrary picture. The drug had no impact on cooing, but it markedly reduced freezing, barking and other hostile gestures. Thus, benzodiazepine-using pathways seemed primarily to influence responses to direct threats but to have little power over affiliative behavior.

We still think the opiate and benzodiazepine pathways basically serve these separate functions. Nevertheless, the simple model we initially envisioned grew more interesting as we investigated two additional drugs: a benzodiazepine called alprazolam (Xanax) and a compound called beta-carboline, which binds to benzodiazepine receptors but elevates anxiety and typically produces effects opposite to those of diazepam and its relatives.

When we administered alprazolam in doses that lower anxiety enough to decrease freezing, this substance, like diazepam, minimized hostility in the threatening, stare condition. And beta-carboline enhanced hostility. No surprises here. Yet, unlike diazepam, these drugs modulated cooing, which we had considered to be an affiliative (opiate-controlled), not a threat-related (benzodiazepine-controlled), behavior. Moreover, both these compounds decreased cooing. We cannot explain the similarity of effect, but we have some ideas about why drugs that act on benzodiazepine receptors might influence cooing.

It may be that, contrary to our early view, benzodiazepine pathways can in fact regulate affiliative behavior. We favor a second interpretation, however. Cooing displayed in the stare condition may not solely reflect an affiliative need (a desire for mother's comfort); at times, it may also be an urgent, threat-induced plea for immediate help. One behavior, then, might serve two different functions and be controlled by different neurochemical pathways. (This conclusion was strengthened for me recently, when I tried to photograph a rhesus infant that had become separated from its mother in the wild—where we are now




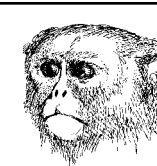

INFANT (left) has strayed a short distance from its mother (*center*) and is producing a rudimentary threat face in an attempt to keep a photographer (the author) at bay. Rhesus monkeys become adept at matching their behavior to the severity and type of a threat when they are between nine and 12 weeks old, probably because certain neuronal pathways in the prefrontal cortex, amygdala and hypothalamus reach functional maturity during this same period.

initiating additional studies. Its persistent, intense coos attracted the mother, along with a pack of protectors. The strategy worked: I retreated rapidly.)

More generally, our chemical studies lead us to suspect that the opiate- and benzodiazepine-sensitive circuits both operate during stress; the relative degree of activity changes with the characteristics of a worrisome situation. As the contribution of each pathway is altered, so too are the behaviors that appear.

Exactly how neurons in the opiate and benzodiazepine pathways function and how they might cooperate are un-

clear. But one plausible scenario goes like this: When a young monkey is separated from its mother, opiate-releasing and, consequently, opiate-sensitive, neurons become inhibited. Such inhibition gives rise to yearning for the mother and a generalized sense of vulnerability. This reduction of activity in opiate-sensitive pathways enables motor systems in the brain to produce cooing. When a potential predator appears, neurons that secrete endogenous benzodiazepines become suppressed to some degree. This change, in turn, leads to elevated anxiety and the appearance of

	 COOING	 FREEZING	 BARKING
MORPHINE (OPIATE)	DECREASES	NO EFFECT	NO EFFECT
NALOXONE (OPIATE BLOCKER)	INCREASES	NO EFFECT	NO EFFECT
DIAZEPAM (BENZODIAZEPINE)	NO EFFECT	DECREASES	DECREASES

EFFECTS ON COOING, FREEZING AND BARKING were evaluated a few years ago for three drugs that act on neurons responsive to opiates (*top two rows*) or to benzodiazepines (*bottom row*). The results implied that opiate-sensitive pathways in the brain control affiliative behaviors (those that restore closeness to the mother, as cooing often does), whereas benzodiazepine-sensitive pathways control responses to immediate threats (such as freezing and barking). Newer evidence generally supports this conclusion but adds some complexity to the picture.



RELAXED MOTHER (left) barely reacts to the presence of the camera-wielding author, whereas a more sensitive mother becomes frightened (*right*), as evinced by her “fear grimace.”



The author hopes explorations of the neural bases for such differences in monkeys will facilitate development of new therapies for excessively anxious human beings.

behaviors and hormonal responses that accompany fear. As the sense of alarm grows, motor areas prepare for fight or flight. The benzodiazepine system may also influence the opiate system, thereby altering cooing during threatening situations.

We are now refining our model of brain function by testing other compounds that bind to opiate and benzodiazepine receptors. We are also examining behavioral responses to substances, such as the neurotransmitter serotonin, that act on other receptors. (Serotonin receptors occur in many brain regions that participate in the expression of fear.) And we are studying the activities of substances that directly control stress hormone production, including corticotropin-releasing hormone, which is found throughout the brain, not solely in the hypothalamus.

In collaboration with Richard J. Davidson, here at Wisconsin, Shelton and I have recently identified at least one brain region where the benzodiazepine system exerts its effects. Davidson had shown that the prefrontal cortex of the right hemisphere is unusually active in extremely inhibited children. We therefore wondered whether we would see the same asymmetry in frightened monkeys and whether drugs that reduced fear-related behavior in the animals would dampen right frontal activity.

This time we used mild restraint as a stress. As we anticipated, neuronal firing rose more in the right frontal cortex than in the left. Moreover, when we delivered diazepam in doses we knew

lowered hostility, the drug returned the restraint-induced electrical activity to normal. In other words, the benzodiazepine system influences defensive behavior at least in part by acting in the right prefrontal cortex.

These findings have therapeutic implications. If human and monkey brains do operate similarly, our data would suggest that benzodiazepines might be most helpful in those adults and children who exhibit elevated electrical activity in the right prefrontal cortex. Because of the potential for side effects, many clinicians are cautious about delivering anti-anxiety medications to children over a long time. But administration of such drugs during critical periods of brain development might prove sufficient to alter the course of later development. It is also conceivable that behavioral training could teach extremely inhibited youngsters to regulate benzodiazepine-sensitive systems without having to be medicated. Alternatively, by screening compounds that are helpful in monkeys, investigators might discover new drugs that are quite safe for children. As the workings of other fear-modulating neurochemical systems in the brain are elucidated, similar strategies could be applied to manage those circuits.

Our discovery of cues that elicit three distinct sets of fear-related behaviors in rhesus monkeys has thus enabled us to gain insight into the development and regulation of defensive strategies in these animals. We propose that the opiate and benzodiazepine pathways in the

prefrontal cortex, the amygdala and the hypothalamus play a major part in determining which strategies are chosen. And we are currently attempting to learn more about the ways in which these and other neural circuits cooperate with one another. We have therefore laid the groundwork for deciphering the relative contributions of various brain systems to inordinate fear in humans. We can envision a time when treatments will be tailored to normalizing the specific signaling pathways that are disrupted in a particular child, thereby sparing that youngster enormous unhappiness later in life.

FURTHER READING

- LOVE IN INFANT MONKEYS. Harry F. Harlow in *Scientific American*, Vol. 200, No. 6, pages 68-74; June 1959.
- THE ETHOLOGY OF PREDATION. Eberhard Curio. Springer-Verlag, 1976.
- STRESS AND COPING IN EARLY DEVELOPMENT. Jerome Kagan in *Stress, Coping, and Development in Children*. Edited by N. Garmezy and M. Rutter. McGraw-Hill, 1983.
- DEFENSIVE BEHAVIORS IN INFANT RHESUS MONKEYS: ENVIRONMENTAL CUES AND NEUROCHEMICAL REGULATION. Ned H. Kalin and Steven E. Shelton in *Science*, Vol. 243, pages 1718-1721; March 31, 1989.
- STRESS IN THE WILD. Robert M. Sapolsky in *Scientific American*, Vol. 262, No. 1, pages 116-123; January 1990.
- DEFENSIVE BEHAVIORS IN INFANT RHESUS MONKEYS: ONTOGENY AND CONTEXT-DEPENDENT SELECTIVE EXPRESSION. N. H. Kalin, S. E. Shelton and L. K. Takahashi in *Child Development*, Vol. 62, No. 5, pages 1175-1183; October 1991.

P.A.M. Dirac and the Beauty of Physics

He preferred the beautiful theory to the fact-butressed ugly one because, as he noted, facts change. He proved his point by predicting the existence of antimatter

by R. Corby Hovis and Helge Kragh

At the University of Moscow, distinguished visiting physicists are asked to leave on a blackboard some statement for posterity. Niels Bohr, father of the quantum theory of the atom, inscribed the motto of his famous principle of complementarity, "Contraria non contradictoria sed complementa sunt" (opposites are not contradictory but complementary). Hideki Yukawa, pioneer of the modern theory of the strong nuclear force, chalked up the phrase "In essence, nature is simple." Paul Adrien Maurice Dirac chose the epigraph "A physical law must possess mathematical beauty."

Exactly 30 years ago Dirac wrote in these pages, "God is a mathematician of a very high order, and He used very advanced mathematics in constructing the universe" [see "The Evolution of the Physicist's Picture of Nature," SCIENTIFIC AMERICAN, May 1963]. Inspired by the views of Albert Einstein and Hermann Weyl, Dirac, more than any other modern physicist, became preoccupied with the concept of "mathematical beau-

ty" as an intrinsic feature of nature and as a methodological guide for its scientific investigation. "A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data," he asserted.

Dirac's focus on the aesthetics and logic of mathematical physics, coupled with his legendary reticence and introversion, made him an enigmatic figure among the great 20th-century scientists [see box on pages 106 and 107]. Sadly, his extreme rationalism also led him into sterile byways after some amazingly successful early years. Between the ages of 23 and 31, Dirac unveiled an original and powerful formulation of quantum mechanics, a quantum theory of the emission and absorption of radiation by atoms (a primitive but important version of quantum electrodynamics), the relativistic wave equation of the electron, the idea of antiparticles and a theory of magnetic monopoles. Yet few of his subsequent contributions had lasting value, and none had the revolutionary character of his earlier work.

Dirac was born in 1902 in Bristol, England, the second of three children in a family that today would be branded as dysfunctional. The bane of the family was its head, Charles Adrien Ladislas Dirac, who had emigrated from Switzerland to England around 1890 and then met and married Florence Hannah Holten, the daughter of a ship's captain. Charles made a living teaching his native language, French, at the Merchant Venturers' Technical College in Bristol, where he was infamous as a rigid disciplinarian. He ran the Dirac household according to the same principles of regimental decorum. By avoiding displays of feeling and equating parental love with discipline, he imprisoned his children in a domestic tyranny that isolated them from so-

cial and cultural life. Unable or unwilling to revolt, Paul sank into the safety of silence and distanced himself from his father. These unhappy years scarred him for life. When Charles Dirac died in 1936, Paul did not grieve. "I feel much freer now," he wrote to his wife.

Fortunately, Paul had a rich interior world to which he could retreat. Early in life he showed an aptitude for mathematics. At age 12, he enrolled in the Merchant Venturers' Technical College. This school, unlike most others at the time, offered not a classical education in Latin and Greek but a modern curriculum in science, modern languages and the practical arts. These studies suited Dirac well, for as he said, he "did not appreciate the value of old cultures." After completing this secondary school program, he entered another institution housed in the same buildings, the Engineering College of the University of Bristol. There he prepared for the career of electrical engineer, not out of real fervor for the work but because he thought it would please his father.

The engineering curriculum gave short shrift to subjects outside applied

R. CORBY HOVIS and HELGE KRAGH have collaborated since 1987 on several projects concerning the history of modern physics. Hovis conducts research on the foundations and history of cosmology and particle physics at the Center for Radiophysics and Space Research of Cornell University. He is currently completing an examination of contemporary developments in gravitation theory. Kragh served as associate professor of history and physics at Cornell from 1987 to 1989, after which he returned to his native Denmark. He has taught physics and chemistry at secondary schools around Copenhagen and is now collaborating with other scholars on a history of technology in Denmark. He has also begun research for a history of cosmology between about 1945 and 1960.

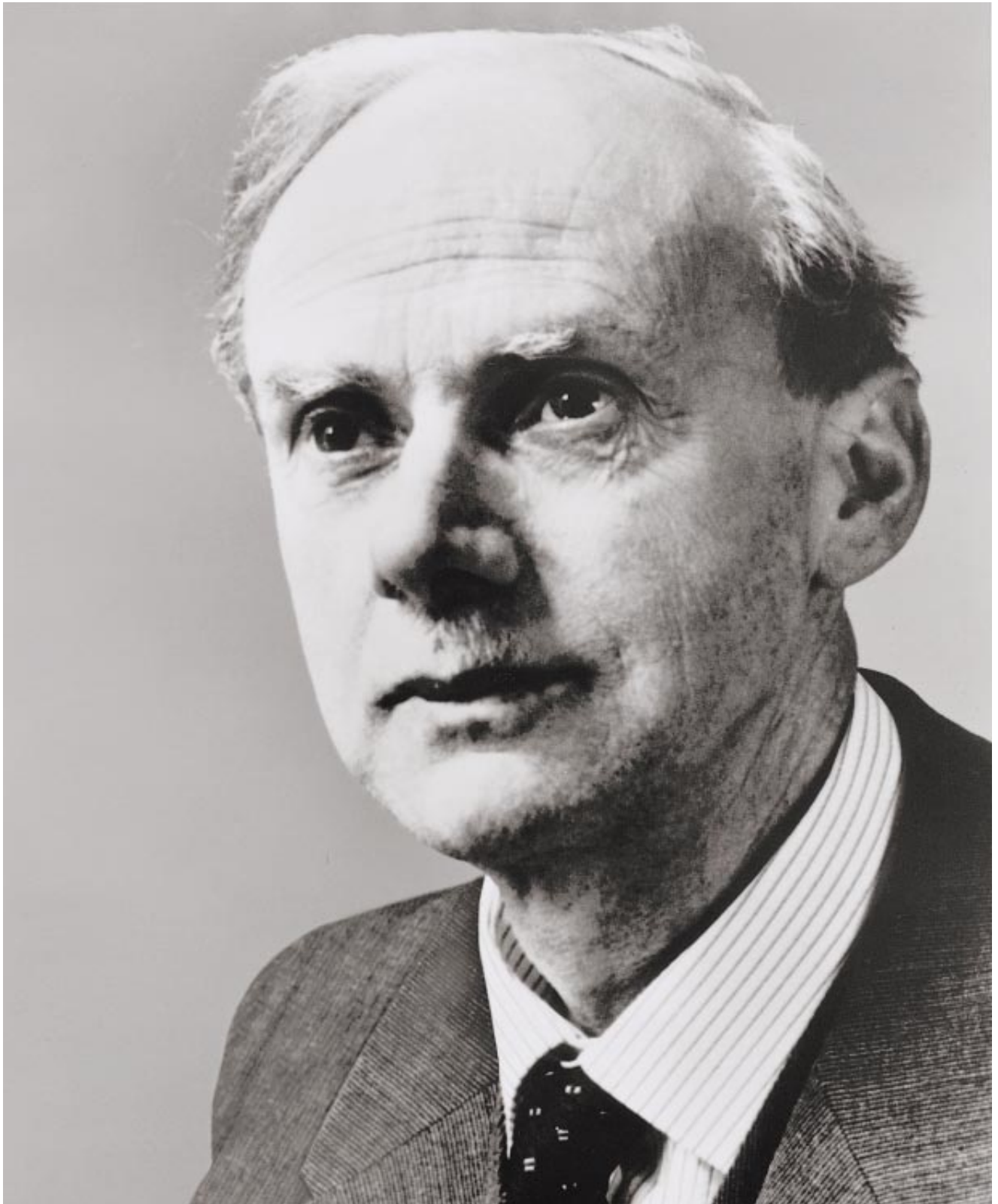
HE WAS TALL, gaunt, awkward, and extremely taciturn," wrote the German physicist and biologist Walter Elsasser. "He had succeeded in throwing everything he had into one dominant interest. He was a man, then, of towering magnitude in one field, but with little interest and competence left for other human activities.... In other words, he was the prototype of the superior mathematical mind; but while in others this had coexisted with a multitude of interests, in Dirac's case everything went into the performance of his great historical mission, the establishment of the new science, quantum mechanics, to which he probably contributed as much as any other man."

physics and mathematics. Despite these omissions, Dirac became fascinated by and soon mastered Einstein's new theories of space, time and gravity—the special and general theories of relativity.

When Dirac graduated with first-class honors in 1921, the postwar economic

depression seemed likely to leave him without a job. He was rescued by a scholarship to study mathematics at Bristol, after which he proceeded, in the fall of 1923, to graduate study in applied mathematics and theoretical physics at the University of Cambridge.

Cambridge was then home to such established scientists as Joseph Larmor, J. J. Thomson, Ernest Rutherford, Arthur Stanley Eddington and James Jeans, as well as to such rising stars as James Chadwick, Patrick Blackett, Ralph Fowler, Edward A. Milne, Douglas R. Hartree



Memorabilia

In 1931, when he was a lecturer and fellow at Cambridge, **Nevill Mott** wrote to his parents: "Dirac is rather like one's idea of Gandhi. We had him to supper here.... It was quite a nice little supper but I am sure he would not have minded if we had only given him porridge. He goes to Copenhagen by the North Sea route because he thinks he ought to cure himself of being sea sick. He is quite incapable of pretending to think anything that he did not really think. In the age of Galileo he would have been a very contented martyr."



Eugene Wigner



J. Robert Oppenheimer

Dirac once attended a luncheon with **Eugene Wigner** and **Michael Polanyi**. There was a lively discussion about science and society, during which Dirac did not say a word. Asked to speak up and give his opinion, he responded, "There are always more people willing to speak, than willing to listen."

A French physicist, who spoke English with great difficulty, once called on Dirac. Dirac listened patiently as the fellow tried to find the right English words to get his point across. Dirac's sister then came into the room and asked Dirac something in French, to which he also replied in fluent French. Naturally, the visitor was indignant and burst out, "Why did you not tell me that you could speak French?" Dirac's terse answer: "You never asked me."

When Dirac passed through Berkeley on his way to Japan in 1934, **J. Robert Oppenheimer** met him and offered him two books to read during the voyage. Dirac politely refused, saying that reading books interferes with thought. Once the Russian physicist **Peter Kapitza** gave Dirac an English translation of Fëdor Dostoevski's *Crime and Punishment*. After some time had passed, Kapitza asked Dirac if he had enjoyed the book. His only comment was: "It is nice, but in one of the chapters the author made a mistake. He describes the Sun as rising twice on the same day." On advice, Dirac also read Leo Tolstói's *War and Peace*; it took him two years.



Peter Kapitza

and Peter Kapitza. Dirac was assigned Fowler as his supervisor, and from him Dirac learned atomic theory and statistical mechanics, subjects he had not previously studied. Of these years, he later recalled: "I confined myself entirely to the scientific work, and continued at it pretty well day after day, except on Sundays when I relaxed and, if the weather was fine, I took a long solitary walk out in the country."

Six months after arriving at the university, he published his first scientific paper; in the next two years he published 10 more. By the time he completed his Ph.D. dissertation in May 1926, he had discovered an original formulation of quantum mechanics and taught the first quantum mechanics course ever offered at a British university. Only 10 years after entering Cambridge, he would receive the Nobel Prize in Physics for his "discovery of new fertile forms of the theory of atoms...and for its applications."

The eight great years in Dirac's life began one day in August 1925, when he received from Fowler the proofs of a forthcoming article by Werner Heisenberg, a young German theorist [see "Heisenberg, Uncertainty and

the Quantum Revolution," by David C. Cassidy; *SCIENTIFIC AMERICAN*, May 1992]. The article laid out the mathematical foundations of a revolutionary theory of atomic phenomena that would soon be known as quantum mechanics. Dirac immediately realized that Heisenberg's work opened up an entirely new way of looking at the world on an ultramicroscopic scale. During the next year, he reformulated Heisenberg's basic insight into an original theory of quantum mechanics that became known as *q*-number algebra, after Dirac's term for an "observable" physical quantity, such as position, momentum or energy.

Although Dirac's work quickly earned him widespread recognition, many of his results were derived contemporaneously by a strong group of theorists working in Germany, including Heisenberg, Max Born, Wolfgang Pauli and Pascual Jordan. Dirac openly competed with them.

Born, Heisenberg and Jordan elaborated Heisenberg's initial scheme in terms of the mathematics of matrices. Then, in the spring of 1926, the Austrian physicist Erwin Schrödinger produced another quantum theory, wave mechanics, which led to the same results as the more abstract theories of

Heisenberg and Dirac and lent itself more readily to computation. Many physicists suspected that the three systems were merely special representations of a more general theory of quantum mechanics.

During a six-month stay at the Institute for Theoretical Physics in Copenhagen, Dirac found the general theory for which so many researchers had hoped—a framework that subsumed all the special schemes and provided definite rules for transforming one scheme into another. Dirac's "transformation theory," together with a similar theory worked out at the same time by Jordan, provided the foundation for all later developments in quantum mechanics.

On December 26, 1927, the English physicist Charles G. Darwin (grandson of the famous naturalist) wrote to Bohr: "I was at Cambridge a few days ago and saw Dirac. He has now got a completely new system of equations for the electron which does the spin right in all cases and seems to be 'the thing.' His equations are first order, not second, differential equations!"

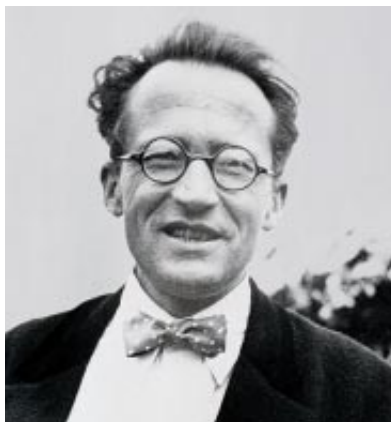
Dirac's equation for the electron was indeed "the thing," for it at once satisfied the requirements of the special theory of relativity and accounted for

Dirac shunned publicity. At first he was inclined not to accept his Nobel Prize. On the day his appointment to the Lucasian chair was announced, he escaped to the zoo to avoid the many congratulations. He refused all honorary degrees—although many were awarded him in his absence and apparently without his acquiescence.

Around 1950 Dirac was assigned to supervise **Dennis Sciama's** graduate studies at Cambridge. One day Sciama enthusiastically entered Dirac's office, saying, "Professor Dirac, I've just thought of a way of relating the formation of stars to cosmological questions. Shall I tell you about it?" Dirac's reply: "No." End of conversation. Dirac did not seem to realize that his brevity and candor could be perceived as impoliteness or impudence.

When Dirac delivered lectures, he strove to present his text with a maximum of lucidity and directness. He considered it illogical to change his carefully chosen phrases just because they had not been understood. More than once, somebody in the audience asked him to repeat a point that had not been understood, meaning that the listener would like a further exposition. In such cases, Dirac would repeat exactly what he had said before, using the very same words.

In 1977 Dirac wrote: "Of all the physicists that I met, I think **Schrödinger** was the one that I felt to be most closely similar to myself. I found myself getting into agreement with Schrödinger more readily than with anyone else. I believe the reason for this is that Schrödinger and I both had a very strong appreciation of mathematical beauty. . . . It was a sort of act of faith with us that any equations which describe fundamental laws of Nature must have great mathematical beauty in them."



Erwin Schrödinger

the experimentally observed "spin" of the electron, which can take either of two values, $+\frac{1}{2}$ or $-\frac{1}{2}$, "up" or "down." Schrödinger's original equation had failed to do this because it was not relativistic, and its relativistic extension, the Klein-Gordon equation, could not account for the spin.

The use of only first derivatives, so impressive to Darwin, was crucial for two reasons. First, Dirac wanted to retain the formal structure of Schrödinger's equation, which contained a first derivative in time. Second, he needed to meet the strictures of relativity, which put space and time on an equal footing. Dirac's difficult reconciliation of the two criteria was at once beautiful and functional: when he applied the new equation to the case of an electron moving in an electromagnetic field, the correct value of the electron's spin came out automatically.

This deduction of a physical property from first principles impressed physicists, who referred to the equation as "a miracle" and "an absolute wonder" and set about to analyze its subtleties. This line of research eventually led to the birth of spinor analysis—a powerful mathematical tool for analyzing problems in virtually all branches of physics—and to the development of relativistic wave equations for particles having spin other than one-half. In another success, when Dirac and others applied his equation to the hydrogen atom, they were able to reproduce exactly the lines observed in its spectrum. Less than a year after publication, the Dirac equation had become what it remains: a cornerstone of modern physics.

Aworshiper of mathematical logic, Dirac was also a master of intuition. These seemingly contradictory intellectual traits were nowhere exhibited more prominently than in his development of his theory of "holes" between 1929 and 1931. This theory illuminated an entire world that had escaped the notice of physicists.

The theory arose from Dirac's realization that his equation pertained not only to familiar, positive-energy electrons but also to electrons having *negative* energy. Such particles would exhibit quite peculiar properties. Furthermore, positive-energy particles would routinely drop down into these negative-energy states, bringing the collapse of the world around us!

In late 1929 Dirac found a way out of the conundrum created by the appar-

ent necessity of negative-energy electrons in nature. He imagined the vacuum to constitute a uniform "sea" of negative-energy states all *filled* by electrons. Since the Pauli exclusion principle prohibits two electrons from occupying the same quantum state, positive-energy electrons would be kept above the invisible sea, to form the "excited" states observed in nature. An excited state could also be created by pouring in enough positive energy to raise an electron from the sea, a process that would leave a "hole" into which another negative-energy electron could fall. "These holes will be things of positive energy and will therefore be in this respect like ordinary particles," Dirac wrote.

But with what particle could a hole be identified? At the time, there were two plausible candidates, both of which Dirac considered: the proton and the positive electron. His first choice, the proton, faced two major difficulties almost immediately. First, one would expect an electron occasionally to jump down and fill a hole, in which case the two particles would annihilate in a flash of light (gamma rays). Such proton-electron annihilations had never been observed. Second, it became apparent that the correct candidate needed to be identical to the electron in all respects except for electric charge—yet the proton was known to be nearly 2,000 times more massive than the electron.

Nevertheless, Dirac, prompted by a desire for simplicity, at first favored the proton as the hole. In 1930 the electron and the proton were the only known fundamental particles, and he did not relish introducing a new and unobserved entity. Moreover, if protons could be interpreted as negative-energy states vacated by electrons, the number of elementary particles would collapse to one, the electron. Such a simplification would be "the dream of philosophers," Dirac declared.

But the objections to his initial interpretation of holes soon became overpowering, and in May 1931 he settled, reluctantly, on the second candidate for the hole, the antielectron, "a new kind of particle, unknown to experimental physics, having the same mass and opposite charge to an electron." The complete symmetry between positive and negative charges in his theory further impelled him to admit the antiproton to the realm of theoretical existence. Thus did Dirac double the number of respectable elementary particles and set the stage for speculations about entire worlds made of antimatter. He also argued for the existence of another hypothetical particle, the magnetic monopole, which would carry an isolated

magnetic charge analogous to the electron's or proton's electric charge. Even today there is no conclusive experimental evidence for monopoles [see "Superheavy Magnetic Monopoles," by Richard A. Carrigan, Jr., and W. Peter Trower; SCIENTIFIC AMERICAN, April 1982].

In September 1932 Dirac was elected to the Lucasian Chair of Mathematics at Cambridge, a professorship that Isaac Newton had once held for 30 years and that Dirac would keep for 37 (it is now occupied by Stephen W. Hawking). That same month, a young experimenter at the California Institute of Technology, Carl D. Anderson, submitted a paper to the journal *Science* that described his apparent detection, in cosmic rays, of "a positively charged particle having a mass comparable with that of an electron." Although this discovery was not at all inspired by Dirac's theory, the new particle, dubbed the "positron," became generally equated with Dirac's antielectron. When he accepted his Nobel Prize in Stockholm in December 1933, the 31-year-old Dirac lectured on the "Theory of Electrons and Positrons." Three years later Anderson, also at age 31, received the Nobel Prize for raising Dirac's particle from the realm of the hypothetical.

Quantum electrodynamics (QED) is the name given to a quantum theory of the electromagnetic field. By the mid-1930s, attempts to formulate a satisfactory relativistic quantum field theory had reached a state of crisis, and many physicists concluded that a drastic change in fundamental physical ideas was needed. Dirac had made path-breaking contributions to QED in the late 1920s and was painfully aware of the formal shortcomings of the existing theoretical framework, which was built mainly around a theory advanced

by Heisenberg and Pauli in 1929. Dirac called the theory illogical and "ugly." Moreover, calculations using it led to divergent integrals—infinities—to which no physical meaning could be attached. In 1936 Dirac worked out an alternative theory in which energy was not conserved. Although this radical proposal was quickly refuted by experiments, Dirac continued to criticize the Heisenberg-Pauli theory and to search—almost obsessively—for a better one. Looking back on his career in 1979, he wrote, "I really spent my life mainly trying to find better equations for quantum electrodynamics, and so far without success, but I continue to work on it."

One logical route toward a better QED would be to use, as a springboard, an improved classical theory of the electron. In 1938 Dirac followed this strategy and produced a classical-relativistic theory of the electron that greatly improved on the old theory that H. A. Lorentz had framed near the beginning of the century. Dirac's theory resulted in an exact equation of motion for an electron treated as a point particle. Because the theory avoided infinities and other ill-defined terms, it seemed likely to lead to a divergence-free QED. But creating a satisfactory quantum mechanical version of the theory turned out to be more troublesome than Dirac had anticipated. He fought with this problem for more than 20 years—in vain.

During 1947 and 1948, a new theory of QED emerged that resolved, in a practical sense, the difficulty of the infinities that had previously ruined calculations. The pioneers of the new theory, Sin-itiro Tomonaga in Japan and Richard Feynman, Julian Schwinger and Freeman Dyson in the U.S., proposed a procedure called "renormalization," in which the infinite quantities in theoreti-

cal calculations were effectively replaced by the experimentally measured values for the mass and charge of the electron. This procedure of (in effect) subtracting infinities made possible extremely accurate predictions, and the theory's many empirical successes convinced physicists to adopt renormalization as *the* method for doing QED.

Dirac, however, resisted the renormalization approach, judging it to be as "complicated and ugly" as the older one of Heisenberg and Pauli. A theory that operates with ad hoc mathematical tricks not directly dictated by basic physical principles cannot be good, he argued, no matter how well it matches experimental results. But his objections were mostly ignored. At the end of his life, he was forced to admit not only that he had become isolated in the physics community but also that none of his many proposals to reconstruct QED had succeeded.

Dirac's fight for an alternative quantum field theory did have some significant by-products, however. One of these was his important classical theory of the electron, mentioned earlier. Another was a new notation for quantum mechanics, known as the "bra-ket," or "bracket," formalism, which elegantly introduced into the subject the powerful mathematics of vector spaces (or "Hilbert spaces," as they are sometimes called). This formalism became widely known through the third (1947) edition of his influential textbook *The Principles of Quantum Mechanics* and has been the preferred mathematical language for the subject ever since.

In general, Dirac worked only in rather specialized areas of quantum theory. So it was somewhat surprising when, in 1937, he ventured into cosmology with a new idea and then developed it into a definite model of the universe. His interest was largely inspired by two of his former teachers at Cambridge, Milne and Eddington, and by discussions with the talented young Indian astrophysicist Subrahmanyan Chandrasekhar, whose graduate work at Cambridge Dirac partly supervised. In the early 1930s Eddington had embarked on an ambitious, unorthodox research program, aiming to deduce the values of the fundamental constants of

It seems reasonable to assume that not all the states of negative energy are occupied, but that there are a few vacancies or "holes" which can be described by a wave function, like an X-ray or γ ray. Such a hole would appear experimentally as a thing with +ve energy, since to make the hole disappear (i.e. to fill it up) one would have to put -ve energy into it. Further, one can easily

CONCEPT OF ANTIMATTER, which Dirac introduced in 1931, grew directly from his theory of "holes," outlined here in a letter to Niels Bohr dated November 26, 1929. It illustrates Dirac's characteristic clarity, conciseness and neat handwriting.

nature by bridging quantum theory and cosmology. This quest for a truly “fundamental theory,” as Eddington called it, stretched rational inquiry into the realm of metaphysical speculation—producing, one critic charged, a “combination of paralysis of the reason with intoxication of the fancy.” Dirac was skeptical of Eddington’s imaginative claims but impressed by his philosophy of science, which emphasized the power of pure mathematical reasoning, and by his idea of a fundamental connection between the microworld and the macroworld.

In his first article on cosmology, Dirac focused on the very big “pure,” or dimensionless, numbers that can be constructed by algebraically combining fundamental constants (such as the gravitational constant, Planck’s constant, the speed of light and the charge and masses of the electron and proton) so that their units of measurement cancel in division. He argued that only these large numbers have profound significance in nature.

For example, it was known that the ratio of the electric force between a proton and electron to the gravitational force between the same two particles is a very large number, about 10^{39} . Curiously, Dirac noted, this number approximates the age of the universe (as then estimated) when that age is expressed in terms of an appropriate unit of time, such as the time needed for light to cross the diameter of a classical electron.

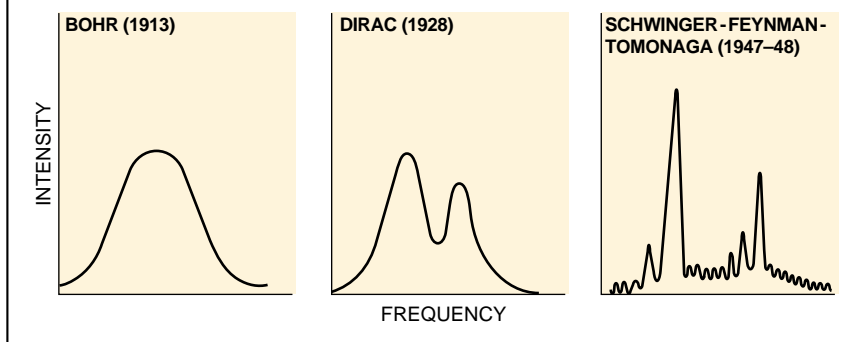
Dirac knew of several such correlations between large pure numbers, but instead of considering them to be mere coincidences, he held that they formed the essence of an important new cosmological principle, which he christened the Large Number Hypothesis: “Any two of the very large dimensionless numbers occurring in Nature are connected by a simple mathematical relation, in which the coefficients are of the order of magnitude unity.”

From this principle, Dirac readily—and controversially—concluded that the gravitational “constant” G is inversely proportional to the age of the universe and hence must be steadily decreasing with cosmic time.

By 1938 Dirac had derived several empirically testable consequences from the Large Number Hypothesis and had outlined his own model of the universe based on that principle. But most physicists and astronomers—who had become increasingly annoyed by the rationalistic approach to cosmology—dismissed his ideas. Only decades later, in the 1970s, did Dirac resume work in cosmology, mostly on the basis of his original theory. He defended the Large

Predicting Hydrogen’s Alpha Line

Hydrogen spectrum’s alpha line illustrates the advances in atomic theory since Niels Bohr first explained it in 1913 as the result of a single quantum transition. When improved experiments revealed a fine structure in the line, Arnold Sommerfeld combined Bohr’s atomic theory with Einstein’s special theory of relativity to explain the components as the result of several transitions. Attempts to derive Sommerfeld’s result from the new quantum mechanics failed until 1928, when Dirac’s theory of the electron proved to reproduce exactly Sommerfeld’s old equation. Later measurements revealed still finer structure, which was explained theoretically in the late 1940s by the modern quantum electrodynamics of Julian Schwinger, Richard Feynman and Sin-itiro Tomonaga. Dirac disliked this new theory because it was, he said, “just a set of working rules,” not a complete theory built on a “sound and beautiful” foundation.



Number Hypothesis and his prediction of a varying gravitational constant against observationally based objections and attempted to modify his model to accommodate new discoveries such as the cosmic microwave background radiation. His efforts failed to gain recognition, and he remained—in cosmology as in QED—a figure estranged from the mainstream of research.

Dirac was wedded to his work, and his colleagues had long considered him an inveterate bachelor. It therefore came as a surprise when in 1937 he married Margit Wigner, sister of prominent Hungarian physicist Eugene Wigner. Margit was a widow; she brought a son and a daughter from her previous marriage, and with Paul she had two girls. Not surprisingly, he remained detached from family life. “It is the irony which only life can produce that Paul suffered severely from his father, who had the same difficulties with his family,” Margit has written. “Paul, although not a domineering father, kept himself too aloof from his children. That history repeats itself is only too true in the Dirac family.”

Dirac never developed an interest in art, music or literature, and he seldom went to the theater. The only hobbies to which he devoted much time were hiking in the mountains and traveling. He was a tireless walker, and on tours

he often demonstrated stamina that amazed those who knew him only from conferences or dinner parties. His travels took him around the world three times, and he climbed some of the highest peaks in Europe and America.

In September 1969 Dirac retired from the Lucasian chair. The next year, he and Margit decided to leave England permanently for the warm climate of Florida, where he accepted a faculty position at Florida State University in Tallahassee. He remained productive and participated in many conferences until his health began to fail. He died in Tallahassee in October 1984.

FURTHER READING

THE HISTORICAL DEVELOPMENT OF QUANTUM THEORY, Vol. 4, Part 1: THE FUNDAMENTAL EQUATIONS OF QUANTUM MECHANICS, 1925–1926. Jagdish Mehra and Helmut Rechenberg. Springer-Verlag, 1982.

PAUL ADRIEN MAURICE DIRAC. R. H. Dalitz and Sir Rudolf Peierls in *Biographical Memoirs of Fellows of the Royal Society*, Vol. 32, pages 137–185; 1986.


REMINISCENCES ABOUT A GREAT PHYSICIST: PAUL ADRIEN MAURICE DIRAC. Edited by Behram N. Kursunoglu and Eugene P. Wigner. Cambridge University Press, 1987.

DIRAC: A SCIENTIFIC BIOGRAPHY. Helge Kragh. Cambridge University Press, 1990.

TRENDS IN ASTROPHYSICS

INCONSTANT COSMOS

by Corey S. Powell, *staff writer*



Space-based telescopes endowed with x-ray and gamma-ray vision observe an ever restless, dynamic universe.

An air of excitement pervades Building 2 at the National Aeronautics and Space Administration Goddard Space Flight Center in Greenbelt, Md. Hallway conversations, punctuated by waving hands, often end in a rush to the blackboard. A colloquium about active galaxies turns into a back-and-forth discussion as various members of the audience jump in to question the speaker, offer corrections or add information of their own. The activity feeds on a steady stream of data flowing to Goddard from scientific satellites, most notably NASA's *Compton Gamma Ray Observatory* and the *Roentgen Satellite*, a joint U.S.-U.K.-German project.

GRO and *ROSAT*, in astronomical parlance, are beaming back information about the objects that generate x-rays and gamma rays, nature's most potent forms of electromagnetic radiation. Each x-ray carries hundreds to tens of thousands of times as much energy as photons of visible light; gamma rays are more energetic yet. By capturing and analyzing these rays, the space-based observatories are offering unprecedented insight into the nature of some of the most violent and baffling phenomena in the universe: cannibal stars, blazing quasars and enigmatic bursts of gamma rays that pop off seemingly at random. "All of a sudden we have these nifty new toys to look with," explains Charles D. Bailyn of Yale University. "They're pushing things forward that had been stagnating for a long time."

The new instruments paint a picture of the heavens that boggles even these extraordinary stargazers. "When you look at the sky at high energies, it's an amazingly inconstant place," reflects Neil Gehrels, the project scientist for *GRO*. On time scales ranging from weeks to thousandths of a second, objects brighten and dim, flicker and oscillate. Such rapid changes imply that the sources of the radiation are minuscule on a cosmic scale (otherwise it would take far too long for a physical change to affect a large part of the emitting region). Yet those same objects are emitting tremendous quantities of energetic radiation.

Astronomers think they have identified the likely culprit behind many of these erratic cosmic beacons: a black hole—a collapsed mass so dense that nothing, not even light, can escape its fierce gravity. A black hole the mass of the sun would stretch only about six kilometers across. The hole itself cannot produce any radiation, of course, but matter close to it can. Theoretical calculations show that anything unlucky enough to approach the hole too closely is spun into a flattened ring, technically known as an accretion disk. As gas in the disk spirals in toward its doom, frictional heating raises its temperature to millions of degrees; beams of particles may also emerge from the disk before matter disappears forever into the hole.

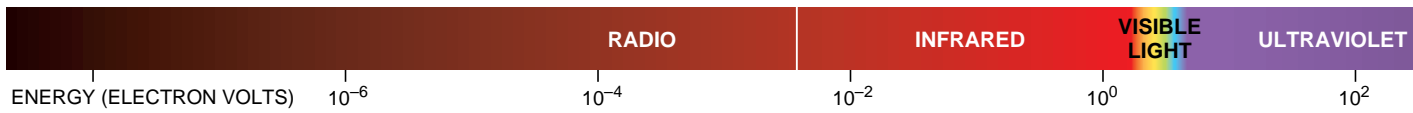
BRILLIANT QUASAR shines as a disk of gas swirls madly about a black hole containing a billion times the mass of the sun. Astronomical satellites sensitive to x-rays and gamma rays, the most energetic forms of radiation, are lifting the veil of mystery from these celestial dervishes.

Exploring the Electromagnetic Spectrum

RADIO WAVES and **MICROWAVES** have been studied using earth-based telescopes since the 1930s. The finest radio observations come from large antenna arrays or huge dishes.

INFRARED RAYS at most wavelengths (energies) are absorbed before they reach the ground. Infrared astronomy came of age in 1983, when NASA lofted the sophisticated *Infrared Astronomical Satellite*.

VISIBLE LIGHT covers only a minuscule portion of the electromagnetic spectrum. Observations at other energies have enabled astronomers to fill in many missing details about the cosmos.



At least, that is the theory. Nobody has ever actually seen a black hole. Although Einstein's theory of relativity predicts that they should exist, and nearly all astronomers believe in them, proving the proposition is another matter entirely. Because the holes themselves would be invisible, astronomers can search for them only by watching what happens in their environs. Here is where x-ray and gamma-ray observations become indispensable. Because these rays carry so much energy, they must originate in the most tortured regions of space, possibly right beside

a hole. Hence, x-rays and gamma rays divulge information crucial for finding likely black holes and for learning how they interact with their surroundings.

Some of the most persuasive data come from observations of x-ray novae, which are among the most volatile denizens of the x-ray sky. Within a few days these objects can soar a millionfold in brightness. Then, over the course of months, the novae gradually recede into meek obscurity. Their behavior generally resembles that of ordinary novae, in which a normal star is slowly consumed by its companion, a collapsed

stellar remnant known as a white dwarf. Gas accumulates on the dwarf's surface until it reaches a critical point and then detonates like a giant hydrogen bomb.

But because x-ray novae shine at much higher energies than conventional novae, the collapsed object is probably something much denser than even a white dwarf. That something could be a black hole. "X-ray novae are a great resource," says Jeffrey E. McClintock of the Harvard-Smithsonian Center for Astrophysics. "They provide the strongest evidence of black holes."

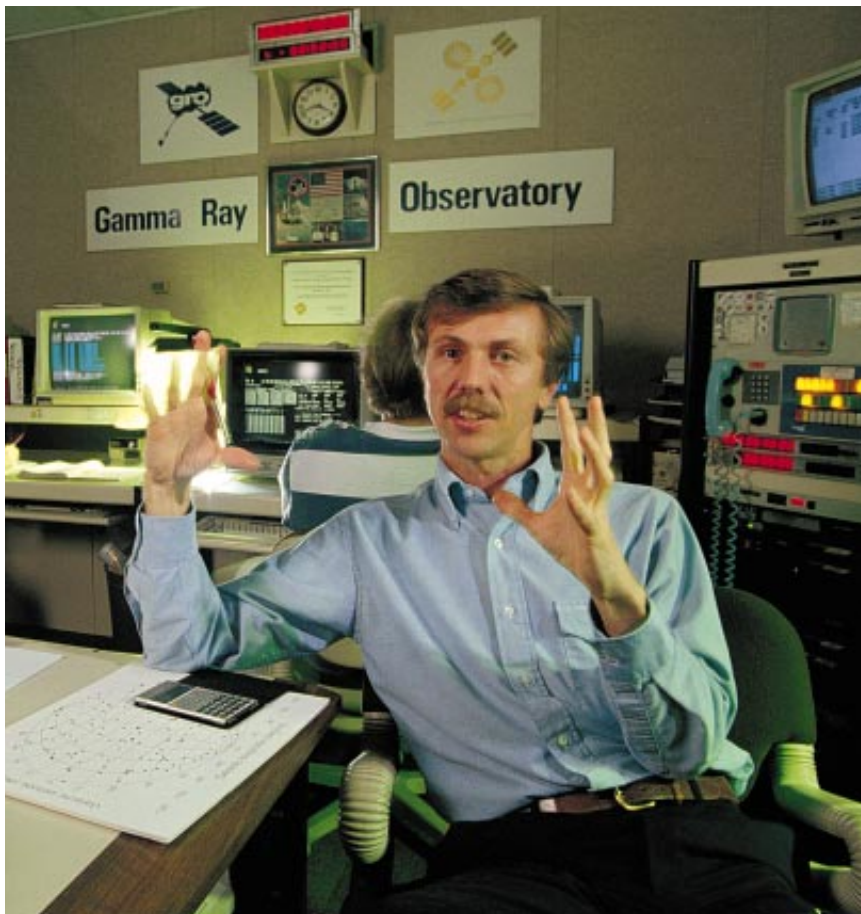
How to Hunt a Hole

McClintock notes that the ephemeral nature of x-ray novae, the property that first caught astronomers' notice, makes them particularly fruitful to study. During an outburst, the glare obscures the individual components of the novae, but once the x-ray source fades it becomes easy to observe the normal companion star, measure its motion and thereby deduce the mass of the strange body it is orbiting.

According to theory, any collapsed body containing more than three times the mass of the sun would generate such an irresistible gravitational field that it would contract into a black hole. For more than a decade, McClintock and others have methodically hunted for an object that surpasses that magical limit.

Within the past two years the data have started to look increasingly promising. Last year, for example, J. Casares of the Astrophysical Institute of the Canary Islands, Spain, and a group of collaborators analyzed V404 Cygni, an x-ray nova (located 7,000 light-years from the earth) that erupted in 1989. They announced that the unseen star must possess a minimum of 6.3 times the mass of the sun, making V404 Cygni, in its discoverers' words, "the most persuasive case yet for the existence of a black hole."

The wildly irregular behavior of x-ray novae may hold important clues to what happens when a black hole lives in proximity to a normal star. "It's not crystal clear why x-ray novae turn off—but they do," McClintock says. One ex-



NEIL GEHRELS of the NASA Goddard Space Flight Center spent years flying balloon-borne gamma-ray telescopes that could glimpse the sky for only a few hours at a time. Now he is the project scientist for the *Compton Gamma Ray Observatory* (GRO), which has pieced together the first comprehensive portrait of the gamma-ray universe and has greatly aided astronomers' efforts to understand the vigorous and highly variable phenomena that may be powered by black holes.

ULTRAVIOLET RAYS and more energetic radiation cannot penetrate the atmosphere. Satellites such as the *International Ultraviolet Explorer*, launched in 1978, have opened up the ultraviolet sky.

X-RAYS from beyond the solar system were discovered in 1962 using sounding rockets. The *Roentgen Satellite*, which began operation in 1990, offers unprecedentedly sharp, sensitive x-ray views.

GAMMA RAYS from cosmic sources were detected by satellite in 1967. The *Compton Gamma Ray Observatory* is now making the first detailed map of the sky in gamma rays.

X-RAYS

GAMMA RAYS

10^4

10^6

10^8

10^{10}

10^{12}

planation for the switching mechanism holds that the black hole starves for fuel most of the time, except during sporadic episodes when the companion star swells, spilling its outer layers onto the hole. Another model posits that gas from the normal star accumulates in a disk around the black hole until the disk reaches a critical state, at which point it abruptly spirals inward, unleashing a flash of radiation.

Theoretical work by Wan Chen of Goddard, in conjunction with Gehrels and with Mario Livio of the Space Telescope Science Institute in Baltimore, supports the latter explanation. Chen's group notes that x-ray novae often undergo one or two secondary brightenings a couple of months after the initial outburst; *GRO* beautifully captured this behavior in Nova Persei 1992, an unusually bright x-ray nova that erupted last year. "We now think we know what causes that," Gehrels says. He and his colleagues suspect that an instability in the accretion disk unleashes the initial blast of x-rays. Radiation from the immediate neighborhood of the black hole then frees material from the surface of the companion star; that material eventually falls into the hole, creating the subsequent flare of x-rays.

As the quality and quantity of x-ray observations have improved, researchers have discovered an intriguing, more repetitive aspect of the flickering of x-ray novae. A group led by William S. Paciesas of the University of Alabama at Huntsville used *GRO* to watch Nova Persei 1992 and found that its brightness varies in a repetitive but not entirely regular way; astronomers refer to such variations as quasiperiodic oscillations. The Japanese *Ginga* satellite recorded similar fluctuations in several other x-ray novae. During the past year, a number of workers have noticed quasiperiodic changes also in Cygnus X-1, a binary-star system long suspected of containing a black hole.

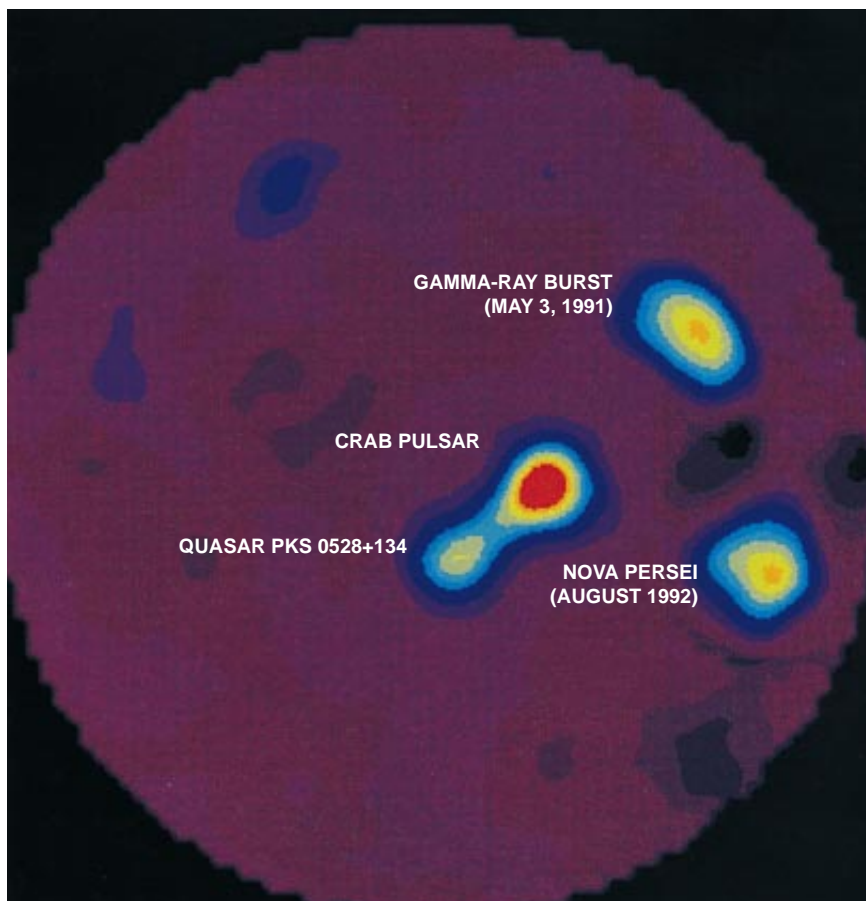
"All of a sudden, quasiperiodic oscillations are popping out all over the place. One wonders what the hell is going on," says Jay P. Norris of Goddard. Until recently, most researchers believed oscillations could occur only around objects

such as white dwarfs and neutron stars, which have solid, rotating surfaces. Black holes have no surfaces at all in the conventional sense, however. Process of elimination implies that oscillations must originate within the accretion disk, but "there's not a single good theory about where exactly the black hole oscillations come from," Norris confesses.

Those reported so far have periods ranging from 10 to 100 seconds—too slow to be caused by rotation of the inner parts of the disk. Eric Gotthelf of Columbia University suggests that the oscillations may result as magnetic-

field lines tear apart and reconnect in the swiftly swirling disk.

The search for black holes in x-ray binaries has generated considerable interest and a few surprises, but it does not come close to matching the passions aroused by claims that the center of the Milky Way is home to a monstrous black hole, possessing about a million times the mass of the sun [see "What Is Happening at the Center of Our Galaxy?" by Charles H. Townes and Reinhard Genzel; *SCIENTIFIC AMERICAN*, April 1990]. One might think that spotting a giant eating machine sitting at the mid-



ERRATIC NATURE of the gamma-ray sky is captured in this composite image made by COMPTEL, one of the instruments on board *GRO*. The Crab pulsar is the brightest continuous source of gamma rays; it flashes 30 times each second. Last August, Nova Persei 1992 brightened by a factor of 100,000 but is now fading back into obscurity. The quasar varies over the course of days. The gamma-ray burst shone with exceptional brilliance yet vanished after just 10 seconds.

de of our home galaxy would be a simple matter, but in fact the indications remain equivocal. Even Sir Martin J. Rees of the University of Cambridge, one of the original proponents of the idea, admits that "the evidence is not overwhelming" but avers, with a laugh, that "there's no strong evidence *against* it."

Rees and others advance a general argument in favor of the existence of a black hole in the heart of the galaxy. Many, perhaps most, galaxies pass through a turbulent early phase, during which they possess brilliant active regions in their centers. Joachim Trümper of the Max Planck Institute for Extraterrestrial Physics in Garching estimates that *ROSAT* has sighted 25,000 of these so-called active galactic nuclei. Most theorists believe those objects are powered by black holes far more mas-

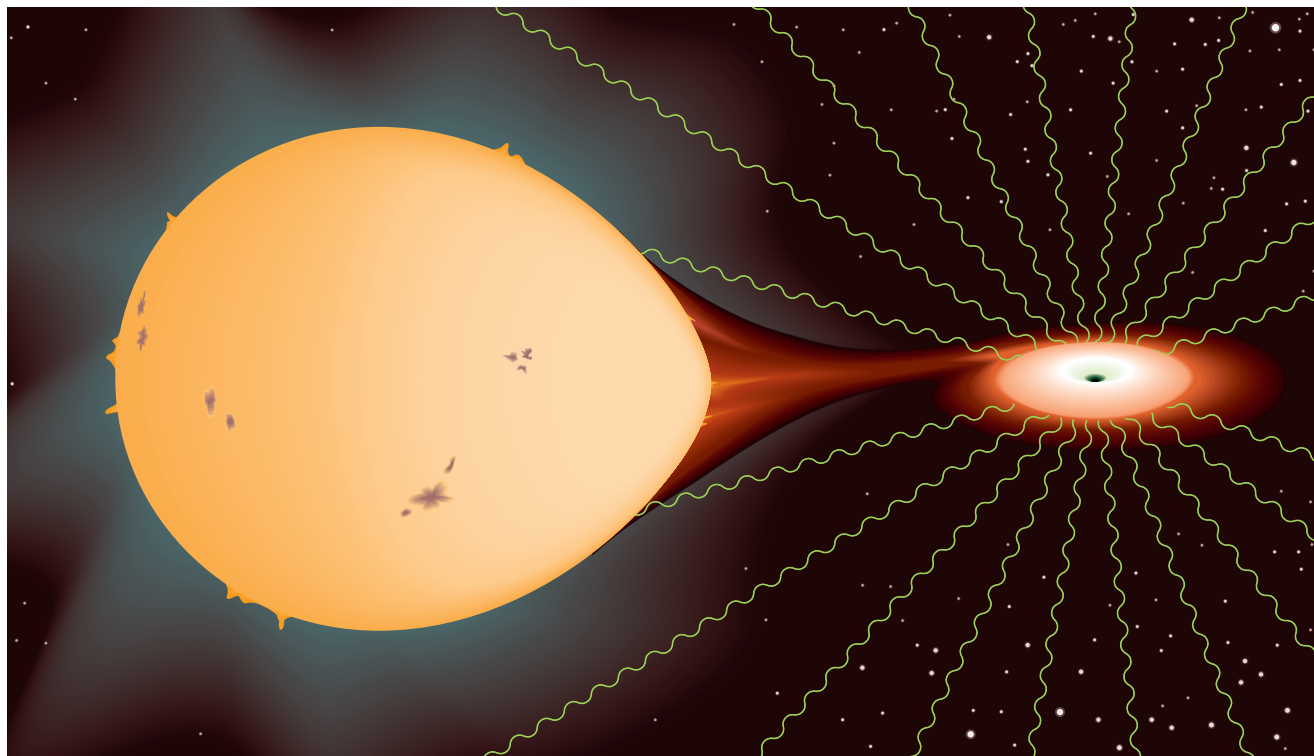
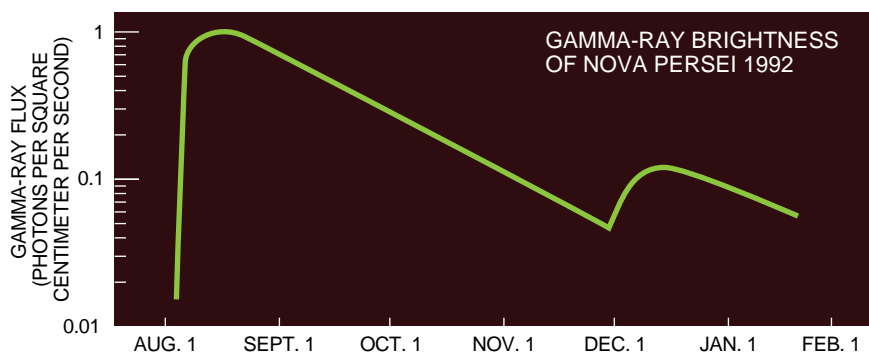
sive than the ones associated with x-ray binaries. Such collapsed giants must therefore be lying dormant in older, more proximate galaxies. If that theory is correct, then it is likely that the Milky Way contains a black hole as well.

There are some tantalizing clues that it does. Astronomers have identified a strong radio source, Sagittarius A*, lying at or very near the precise center of the galaxy in the midst of a region of disturbed, rapidly moving gas. But, unlike x-ray novae, Sagittarius A* proved a tough target to spot with x-ray and gamma-ray telescopes. The paucity of high-energy radiation from Sagittarius A* did not surprise Rees, who notes that "most of the time the hole may not be doing very much"—in other words, the hole may be fairly quiet simply because very little gas is falling in.

Finally, in 1991, the French-Soviet *GRANAT* satellite managed to detect faint x-ray emissions from Sagittarius A*. And at the American Astronomical Society meeting this past January, a team led by John R. Mattox of Goddard related that the EGRET telescope on board *GRO* had detected a weak gamma source within 50 light-years of the dynamic center of our galaxy.

Fulvio Melia of the University of Arizona has constructed a self-consistent model in which the unsteady x-ray, infrared and radio emissions of Sagittarius A* emerge from a disk of hot gas spiraling in toward a black hole. He proudly claims to have derived a "best fit" mass of about 900,000 solar masses and an accretion-disk diameter of about 100 million kilometers, roughly the size of Mercury's orbit around the sun. Other, less committed researchers take a more skeptical view; Robert Petre of Goddard, the project scientist for the U.S. fraction of the *ROSAT* mission, declares that "the jury's still out—there's so much conflicting evidence regarding the black hole at the galactic center."

That confusion derives in part from the fact that not one but many x-ray and gamma-ray sources are near the hub of our galaxy. One of the most intriguing and controversial of these sourc-



X-RAY NOVA erupts when material from a normal star flows onto its collapsed companion, possibly a black hole. Gas spiraling around the black hole grows hot and emits torrents of x-rays before being swallowed (*bottom*). Nova eruptions are

remarkably abrupt, as seen in this brightness curve of Nova Persei 1992 (*top*). The nova's secondary brightening may have been caused by instabilities in the normal star that were triggered by radiation from the vicinity of the black hole.

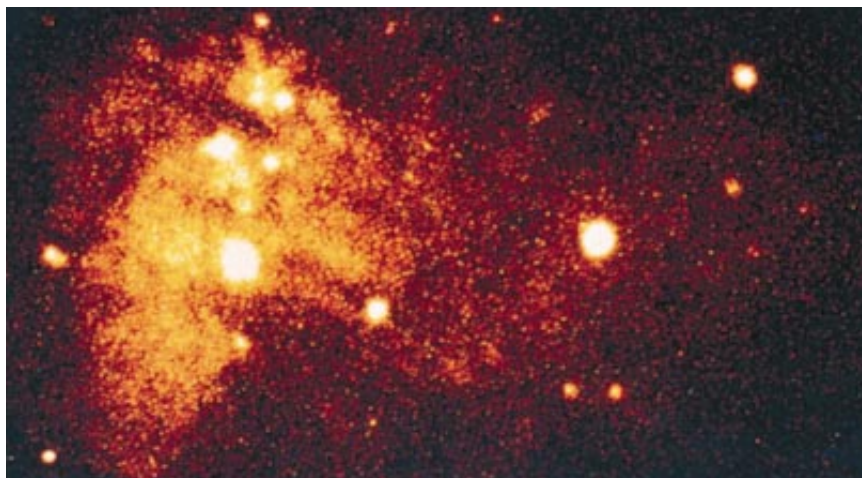
es goes by the unmemorable name of 1E 1740.7-2942. Three years ago Rashid Sunyaev of the Institute of Cosmic Research in Moscow, along with a flock of Soviet and French collaborators, reported that *GRANAT* had observed a brief but tremendously bright pulse of gamma rays coming from 1E 1740.7-2942. Much of the radiation had an energy near 511,000 electron volts, the amount liberated when an electron meets and annihilates its antimatter twin, the positron. (In comparison, visible light carries an energy of about two electron volts.) For this reason, Marvin Leventhal, now at Goddard, nicknamed the source the "Great Annihilator."

In his clipped, slightly detached style, Leventhal outlines how the Great Annihilator might work. It probably consists of a massive normal star and a stellar-mass black hole locked in close orbit around each other. Gas falling into the hole grows so hot that it emits gamma rays. Because the radiating region around the hole is so small, many of those rays collide with one another to create pairs of electrons and positrons (this effect is a manifestation of Einstein's famous equation stating that matter and energy are equivalent). Jets of electrons and positrons squirt out of the system; positrons eventually annihilate with electrons in a nearby, dense cloud of gas, creating the 511,000-electron-volt gamma rays.

Like other suspected black holes, the Great Annihilator has proved maddeningly capricious. Now that *GRO* is keeping a watchful eye on the galactic center region, the Great Annihilator is behaving like a naughty child caught in her parents' gaze: it is keeping politely quiet. William Purcell, Jr., of Northwestern University reports that, much to his dismay, the *OSSE* instrument on *GRO* sees "no evidence" that the Great Annihilator is doing any annihilating at all. "It would be much more exciting if we did see the 511,000-electron-volt line," he agrees, "but I call 'em as I see 'em."

On the other hand, a radio map assembled by I. Felix Mirabel of the Center of Nuclear Research in Saclay, France, along with several colleagues, affirms that the Great Annihilator is truly something rare and spectacular. Two opposed radio-emitting jets emerge from a compact source that coincides with the location of the x-ray and gamma-ray emissions. Mirabel's interpretation, which Leventhal heartily endorses, is that the radio jets trace out the motion of the high-speed positrons before they perish in the surrounding gas clouds.

Such jet structures are uncommon among objects in the Milky Way but evoke a sense of déjà vu among astron-



X-RAY VIEW captured by the *Roentgen Satellite (ROSAT)* shows remarkably fine details in the Large Magellanic Cloud. The bright spot at the middle left represents the intense x-ray glow from LMC X-1, a binary-star system believed to contain a black hole. Some of the other sources seen here are stars and clouds of hot gas. This image stretches about four times the diameter of the full moon.

omers who study the greatest cosmic powerhouses of all—quasars and their ilk, collectively known as active galactic nuclei. In many cases, jets extend hundreds of thousands of light-years from the brilliant but compact source at the centers of these galaxies. Active galactic nuclei also resemble the Great Annihilator in their seemingly paradoxical combination of quick variation and staggering luminosity.

Making Sense of Quasars

Quasars can outshine the entire Milky Way by a factor of 1,000, and yet their brightness in visible light can vary by 50 percent over the course of a single day. The latest round of high-energy observations—most notably by *GRO*—emphasizes just how extreme the behavior of some quasars can be. The new data also lend impressive support to the leading theoretical models of how such objects can pull off their almost miraculous stunts.

Those models hold that the most luminous quasars contain holes packing a billion solar masses into a region that would fit neatly inside Pluto's orbit around the sun. As in x-ray binaries, circling clouds of gas would grow hot and radiate furiously immediately before being sucked into the central black hole. In some cases, jets of charged particles tens of thousands of light-years long shoot from the accretion disk. The angle at which the disk is seen by terrestrial astronomers determines the appearance of the object and hence the way it is classified. Quasars and their similar cousins, BL Lac objects, are the most luminous active galactic nuclei;

their more sedate relatives are known as Seyfert galaxies.

Observations at x-ray energies serve up persuasive, if circumstantial, evidence that active galactic nuclei do indeed consist of accretion disks around supermassive black holes. For several years now, astronomers have known that the x-ray brightness of the nuclei can change particularly abruptly, over the course of hours or even minutes. "Many people are very keen on x-ray variability," explains Andrew Lawrence of Queen Mary and Westfield College in London, "because the speed of it tells you that you must be looking right down to the core"—that is, to the immediate vicinity of the black hole. Only in the inner regions around a hole would the distances be so small and the motions so fast that changes could occur over such very short time scales.

Previous investigations have reported that the x-ray variations appeared random, but when Lawrence and his graduate student Iossif Papadakis reanalyzed archived x-ray observations of the Seyfert galaxy NGC 5548, they saw something quite unexpected: a quasiperiodic oscillation of about eight minutes. When Lawrence and Papadakis checked observations from *Ginga*, they saw the same time pattern there as well, "which proves we're not dreaming it!" Lawrence says cheerily. He was also heartened that similar oscillations had just been detected in binary-star systems thought to contain black holes. Lawrence suspects that "the same mechanism explains both kinds of objects."

The periodicity of NGC 5548 may denote a blob of hot material orbiting in and out of sight in an accretion disk, or

it may be an oscillating instability in the disk itself. If either interpretation is correct, one can deduce an approximate mass for the black hole from the period of the variation. Lawrence infers a mass of between 100,000 and one million solar masses, 10 to 100 times less than what other researchers had estimated from the optical variability of NGC 5548. "It makes me a bit nervous," Lawrence admits.

Of course, the black-hole guessing game still contains lots of room for error. Lawrence hopes his observations will push the theoretical models "right to the limit," thereby elucidating how black hole systems operate.

Theorists are coming to accept that quasiperiodic oscillations probably occur wherever accretion disks are found, but so far only one exceptional active galaxy displays regular changes, clock-like in precision. In 1985 the *EXOSAT* x-ray satellite found that the flaring behavior of another Seyfert galaxy, NGC 6814, seemed to repeat every 3.3 hours. Chris Done of the University of Leicester analyzed recent data from *Ginga* and derived a more precise period of 12,130 seconds. She also showed that the period remains "very stable," suggesting that it does not originate in an accretion disk, where conditions probably change continuously.

Done judges that the most plausible cause of the regular variation is a star trapped in a tight orbit around the hole.

Such an interpretation requires that the intense x-rays emitted by NGC 6814 come from a single object, not a collection of smaller sources. That argues in favor of the presence of a heavyweight black hole; the period of the variation even gives a hint at the hole's mass.

Clearly, the star has not yet been devoured by the hole, and so it must be orbiting at a somewhat safe distance. If the star circles about 80 million kilometers from the hole—about 50 times the radius of the hole itself—then the hole would have one million times the mass of the sun. Done explains that the central object must contain less than 100 million solar masses to avoid swallowing the star, but it must possess more than 100,000 solar masses to produce the observed x-ray luminosity of NGC 6814. That range lies well within the values that astrophysicists expected on the basis of models of how active galactic nuclei shine.

"Photons from Hell"

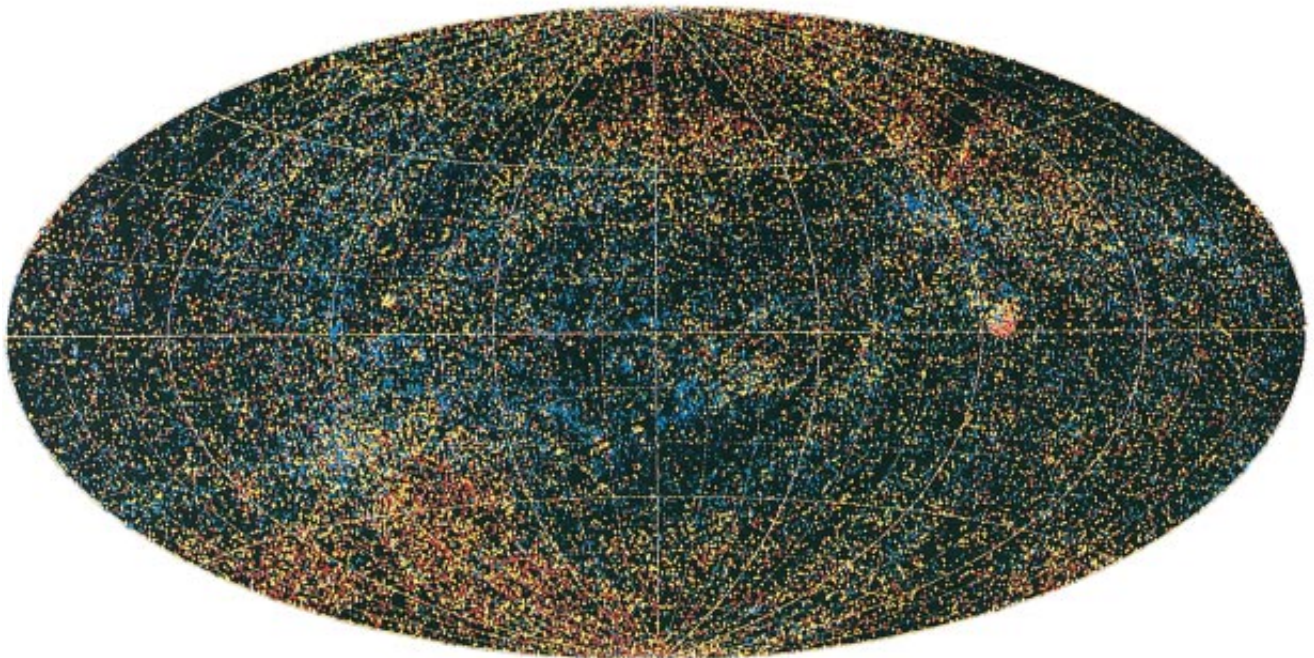
GRO is testing some of the details of those models by providing an immensely improved picture of what active galactic nuclei look like at very high gamma-ray energies. The new data "have completely reinvigorated the field," says Charles D. Dermer of the Naval Research Laboratory in Washington, D.C. So far, Dermer is happy to relate, the *GRO* results are "fairly consistent with the stan-

dard scenario." Active galaxies that produce no radio emission seem to have no jets, so all their radiation must come from the material surrounding the black hole. He reports that emission from these "radio quiet" active galaxies fizzles out at energies above about 100,000 electron volts. The high-energy spectrum of these objects resembles what one would expect from a disk of hot gas circling a supermassive black hole, he asserts.

The active galaxies that have radio-emitting jets should be able to produce outpourings of gamma rays at much higher energies. Before *GRO*'s launch in 1991, only one active galaxy, the quasar 3C 273, had been observed to emit gamma rays carrying greater than 100 million electron volts, so researchers could not draw any general conclusions.

That situation has changed now that *GRO*'s sensitive scans of the gamma-ray sky have turned up 23 such galaxies, a number that increases weekly. All these objects are strong radio sources, and nearly all display what astronomers call a "blazar-type spectrum," in which the spectrum at the highest energies looks decidedly unlike the radiation emitted by hot gas. Dermer concludes that "we seem to be looking nearly straight down the jet" of these galaxies.

The existence of a jet beamed toward the earth would help explain many of the attributes of the remarkable quasar 3C 279. Although it lies roughly six billion light-years away—halfway to the



SKY MAP produced by *ROSAT* shows roughly 50,000 sources of x-rays. More than half of these are thought to be active galactic nuclei, galaxies that have anomalously bright and dynamic central regions. Many of the other sources are normal stars. The active galaxies are, on average, a million times

more distant than the stars, indicating the tremendous luminosity of those galaxies. Colors denote the relative flux of high-energy ("hard") and low-energy ("soft") x-rays from each object. Red indicates the softest sources, yellow intermediate ones and blue the hardest.

visible limit of the universe—3C 279 showed up clearly in *GRO*'s detectors. Based on those observations, 3C 279 holds the record as the brightest gamma-ray source in the universe. Whatever central engine produces the gamma rays is extremely efficient: 3C 279 emits 10 times as much energy in gamma rays as it does in all other parts of the spectrum. The engine must also be small. Robert C. Hartman of Goddard and his collaborators, working with the EGRET telescope on board *GRO*, watched 3C 279's gamma-ray brightness double and then fall fourfold over a two-week period last year.

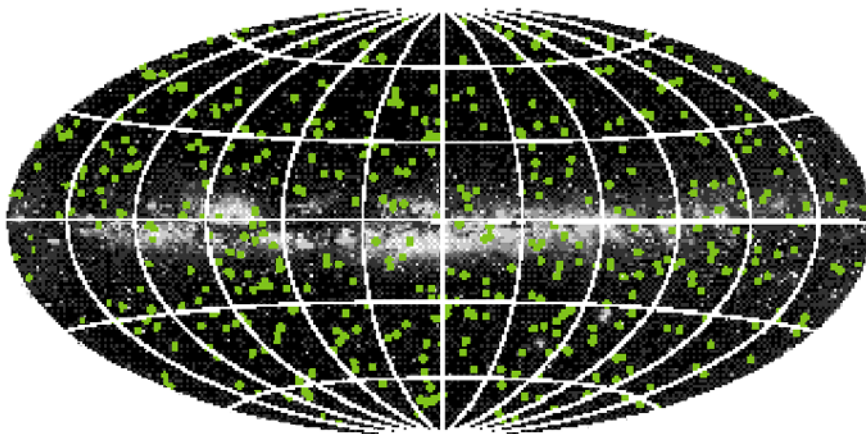
A beam of particles traveling close to the speed of light makes the attributes of 3C 279 at least somewhat explicable. If the earth lies in the sights of a focused beam, then the quasar need not radiate as intensely in other directions, and so its total luminosity can be somewhat more modest than one might naively assume. Particles accelerated to near-light velocities in the beam could carry enough energy to generate the high-energy gamma rays.

In fact, 3C 279 was already known as a "superluminal" quasar, one in which parts of the object seem, in radio images, to be moving faster than the speed of light. The superluminal situation is thought to be an illusion caused by relativistic effects in a pointed beam traveling close to the speed of light, an idea that is consistent with the other observations.

Particles in the jets move so furiously that they give rise to gamma rays too energetic for even *GRO* to see. Last August a group of researchers led by Trevor C. Weekes of Whipple Observatory in Amado, Ariz., discovered the most vigorous gamma rays ever seen from an active galaxy. They found that Markarian 421, a nearby BL Lac object (it is only about 400 million light-years away, quite close for an object of its type), emits gamma rays having trillions of electron volts of energy.

These rays are so energetic that they can be observed from the earth's surface. When they collide with atoms in the atmosphere, the rays create a spray of secondary subatomic particles, which in turn give rise to a visible flash as they decelerate. The gamma rays from Markarian 421—referred to as "photons from hell" by some researchers—indicate that particles in the jet must shoot out with at least as much energy as the gamma rays they eventually emit. *ROSAT* recently demonstrated that Markarian 421 is also a highly variable x-ray source.

The finding that many quasars and BL Lac objects shine intensely and unsteadily at high energies further buttres-



GAMMA-RAY BURSTS (green dots) flare for anywhere from a fraction of a second to a couple of minutes; no burst has even been seen to repeat. Many astronomers thought bursts originated from neutron stars in our galaxy, in which they should appear to line up along the band of the Milky Way (seen in this visible-light photograph). Instead *GRO* showed that they come from random directions.

ses the argument that these objects must be built around supermassive black holes. One long-standing opposing theory holds that active galaxies radiate as the result of widespread star formation and supernovae explosions. But as Andrew Robinson of the University of Cambridge notes, such a model cannot readily account for the energy or the fast variability of the radiation from many active galactic nuclei.

Explaining the Inexplicable

Given their success at using black holes to explain x-ray novae and active galactic nuclei, some astrophysicists also invoke black holes to explain gamma-ray bursters, perhaps the most enigmatic and variable phenomena known. For anywhere from a fraction of a second to a couple of minutes, they gleam with an intensity that dominates every other gamma-ray source in the sky—often including the sun. No two flashes look exactly alike, and no two originate from the exact same direction. Moreover, bursts intensify and dim astonishingly fast even by the standards of high-energy astrophysics. The briefest variations occur in much less than a thousandth of a second, indicating that the burst originates in a region only a few tens of kilometers across. Nobody has ever been able to observe the objects from which the bursts emerge.

By the mid-1980s astronomers had come to a general agreement about the origin of these odd gamma-ray eruptions. The standard theory held that the bursts resulted from some kind of disruption on or around a neutron star, an ultradense remnant left over from a supernova explosion. A neutron star's

gravity is so great that a mass roughly that of the sun is compacted into a sphere only 20 kilometers across. Because of the crushing gravity, even the slightest disruption—a seismic glitch, the impact of a small asteroid—could liberate a tremendous amount of energy, thereby explaining the fleeting flashes of gamma radiation.

Researchers hoped that observations from *GRO* would confirm the standard theory and lay the mystery of gamma-ray bursts to rest. *GRO* contains an instrument called the Burst and Transient Source Experiment, or BATSE, which scans the entire sky for sudden gamma-ray blips. If bursts are associated with neutron stars, then the faintest ones should appear to line up in the plane of the Milky Way, just as the faintest stars in the night sky fall along that band.

Much to everyone's astonishment, BATSE showed that faint bursts are spread randomly around the sky, just like the bright ones. Compounding the conundrum, there is a net deficit of faint bursts, indicating that BATSE is seeing to the edge of a spherical but bounded population. And, unlike previous instruments, BATSE does not see the kind of gamma-ray spectra that researchers expected neutron stars might produce.

Overnight, the neutron-star model began to look shaky, and theorists scrambled to concoct alternative ideas. Not every researcher was caught flat-footed, however. "It was all speculation from the beginning," says Bohdan Paczynski of Princeton University, his voice thick with exasperation.

So what lies around the earth more or less equally in all directions and yet is finite in size? One answer, almost blindingly obvious, is the universe it-



BOHDAN PACZYŃSKI of Princeton University has spent years struggling to solve the mystery of gamma-ray bursts. During the 1980s, he remained skeptical of the prevailing theories explaining the bursts and so was not particularly surprised when *GRO* proved them incorrect. He suspects that the bursts could arise from collisions between neutron stars in distant galaxies.

self, in which case the perceived “edge” in the gamma-ray-burst population corresponds to the visible limit of the cosmos. Paczyński had been needling his colleagues with that possibility for years before the BATSE results surfaced; now it is becoming the majority opinion.

The question that Paczyński and others must face is how an object the size of Bermuda, lying millions or even billions of light-years away, could outshine everything else in the gamma-ray sky. Not surprisingly, a black hole quickly enters the picture. Paczyński has worked out a scenario in which a neutron star collides with a black hole or in which two neutron stars coalesce, forming a hole. The initial collision, and leftover debris that falls into the hole, would liberate a brief but intense blast of gamma radiation. Such a model “satisfies the basic observational constraints” on gamma-ray bursts, although he hastens to add that “there’s nothing in the data to show that it’s right.”

Bradley E. Schaefer of Goddard claims that, on a decent day, “I can come up with eight fairly good refutations” of Paczyński’s model. His sharpest criticism is that, in the simplest version of the model, colliding neutron stars should produce a very different kind of gamma-ray spectrum than bursts actually display. On the other hand, Dermer points out that gamma-ray bursts do

look “amazingly similar to gamma-ray emission from quasars.” He therefore tosses out another, quite different black hole model. In his favored scenario, bursts occur when a quiescent, massive black hole in the center of an obscure galaxy consumes an entire star in a single, abrupt gulp. Here, too, there is a problem. Rees has calculated that the resulting flare-up should last for a few years, not a few seconds.

But as Paczyński notes, one need not resort to black holes at all. If gamma-ray bursts do not occur in the far reaches of the universe, then theorists could return to the neat neutron-star-burst models developed before BATSE came along. The problem then becomes one of explaining why the bursts show up equally in all directions. One way to account for that even distribution is to assume that the Milky Way is surrounded by a halo of neutron stars more than about 100,000 light-years in radius.

This kind of “what if” immediately raises another question: How did the neutron stars get into the halo, where no normal stars are seen? Joseph Silk of the University of California at Berkeley, in collaboration with David Eichler of Ben Gurion University in Israel, offers an intriguing possibility. The Milky Way may be surrounded by a swarm of white dwarfs, relics of the earliest population of stars that formed before our

galaxy had collapsed down to its present, flattened spiral shape. Pairs of these white dwarfs would occasionally merge, forming a neutron star. Each newborn neutron star would emit a burst of gamma rays as its interior settled down.

A search now under way for massive clumps of unseen material surrounding the Milky Way (the notorious “dark matter”) should settle within the next few years whether Silk’s proposed cloud of white dwarfs really exists. But in general, the evidence from BATSE does not look promising for models that place gamma-ray bursts in the Milky Way’s halo. Jerome James Brainerd of the NASA Marshall Space Flight Center, for example, abandoned the previous thrust of his work because “I find it difficult to make a galactic halo model work now.”

The resounding lack of a suitable explanation for the gamma-ray bursts has led some workers to consider some rather flaky models. R. Stephen White of the University of California at Riverside is trying to squeeze gamma-ray bursts out of colliding comets. His model requires the existence of a new, previously unknown cloud of magnetized comets surrounding the sun. “The theorists are having a field day,” sighs Gerald Fishman of the Marshall Space Flight Center, who is the principal scientist on BATSE. “We just don’t know what these things are.” More than 100 gamma-ray-burst models have appeared in peer-reviewed journals within the past two years.

The gamma-ray-burst puzzle serves as a cautionary reminder that, despite *ROSAT* and *GRO*, astronomers are just beginning their exploration of the high-energy sky. Even then, there is a long road from observation to understanding of the fickle objects that may contain black holes. Paczyński muses that “astronomers were observing stars for millennia” before they understood what makes them shine. “What is it? and How does it work? are very different questions,” he comments. The desire to bridge that gap, of course, accounts for much of the passion now so evident in the astronomical community. Schaefer sums up the situation with enthusiastic brevity: “The payoff is huge, so you keep plugging away. You’ve just got to be an optimist.”



Gigabit Gestalt

Clinton and Gore embrace an activist technology policy

In my office I have four cubic feet of reports that have been written since 1985, all saying about the same thing, recommending how we can do better with technology policy in competitiveness and productivity." The speaker was Craig Fields, head of the Microelectronics and Computer Technology Corporation, talking to Bill Clinton and Al Gore at the pre-Inauguration economic summit in December.

Unlike the previous two administrations, President Clinton's team seems to have taken those recommendations to heart. Just two months after the summit, it performed what appeared to be a massive cut-and-paste operation that stitched together the "executive summary" sections from the many dust-gathering reports. The result was a distinct statement about where the new administration stands on the matter of technology.

In introducing his plan, Clinton proved that he, unlike his predecessor, who seemed to be awed by a supermarket price scanner and was uncomfortable with technology policy, could master the nuances of technospeak. Working a crowd of employees at a Silicon Valley computer manufacturer, the president recounted how as governor of Arkansas he had introduced "total quality management" in state government. And he remarked that the vice president is the only person holding national office who can explain the meaning of "the gestalt of a gigabit," a reference to Gore's obsession with building high-speed computer networks.

Rhetoric aside, the Clinton-Gore plan marks a clear break with the past. Gone is the ambivalence or outright hostility toward government involvement in little beyond basic science and the development of such military technology as Star Wars. In fact, the proposal establishes a goal over the course of Clinton's term for achieving an equal balance between defense and civilian research and development. Today defense R&D consumes nearly 60 percent, or \$41 billion, of federal research spending. "The nation urgently needs improved strategies for government-industry coopera-

tion in the support of industrial technology," the document states.

Still to come, however, are specifics; the plan was cobbled together in three weeks. And it remains to be seen whether the technology policy lapses into an industrial policy in which the government picks companies or entire industries to become "winners" by favoring them with federal largesse. "There was so little time," says John H. Gibbons, the president's science and technology adviser, who came to his new job after serving as director of the congressional Office of Technology Assessment. "Now we have to go back and stake out what's missing."

One of the biggest challenges Gibbons faces will be to implement this vision in an administration that has vowed to shrink government. The plan carefully ignores the incessant entreaties that resonate through the stack of studies, calling for an agency to direct the civilian technology agenda. Perhaps substituting form for substance, Gibbons describes the present arrangement of having several agencies responsible for disbursing funds for technology a "virtual" technology office. In doing so, he borrows the notion, now trendy in management consultant circles, of a "virtual" corporation—one made up of many companies that come together for a specific project and then disband once their work is done. The White House Office of Science and Technology Policy under Gibbons will try to orchestrate these wide-ranging efforts, in part by bolstering an existing cross-agency panel, the Federal Coordinating Council on Science, Engineering and Technology.

By spreading the responsibility around, the White House resolved a much pondered question about the fate



HIGH-DEFINITION PRESIDENCY came into focus in late February as Bill Clinton and Al Gore responded to questions about their technology policy at Silicon Graphics, a California computer workstation manufacturer.

of the Defense Advanced Research Projects Agency (DARPA). The agency is often cited as a model for government management of technology projects in the private sector and universities. It played a leading role in fostering parallel computing, artificial-intelligence computer networking and materials science. But because of its deep roots in the Department of Defense, DARPA will not receive all-encompassing responsibility for technology oversight, a possibility that was often considered by the numerous study panels on competitiveness.

Instead the Clinton plan formalizes a role that DARPA has had to pursue quietly in recent years as a backer of "dual use" technologies: those that have both military and commercial signifi-

cance. The "D" will be removed from the name, restoring it to the original the agency received 35 years ago when it was established in the panic following the launch of *Sputnik*.

DARPA has also managed Sematech, the new government's paradigm in helping industry to generate technology. The semiconductor-equipment research consortium has taken a measure of credit for the turnaround in the fortunes of the U.S. chip industry. It has also served as a focus of debate over the value of government collaborations with industry. Some conservatives criticize DARPA's having put up half of the \$1 billion Sematech has spent over the past five years. They say the industry's improved performance does not result from government sub-

sidies but from the U.S. companies' creative prowess in designing custom logic circuits.

The Clinton administration will give Sematech a chance to continue to tell its side of the story by extending funding for a second year beyond DARPA's original five-year obligation, which drew to a close in 1992. Not surprisingly, the consortium has already begun to inspire eager imitators. Alliances that might be dubbed Aerotech, Autotech, Shiptech—even Fibertech for the textile industry—are all being contemplated or are in the works.

Another participant on the industrial-policy debating team will be the Department of Commerce, elevated from the second-tier status it held for the past 12 years. Within that department,

a big winner is the National Institute of Standards and Technology (NIST), which in 1988 changed its name from the National Bureau of Standards after Congress made it responsible for aiding industry to develop commercial technology. By the 1997 budget year Clinton proposes increasing by more than 10-fold the \$68 million for NIST's Advanced Technology Program (ATP), which provides funding for research into high-risk technologies. The administration will also give the department additional support for expanding NIST's seven industrial assistance centers into a nationwide network numbering more than 100.

The influx of money has caused concern that ATP will balloon uncontrollably and perhaps eclipse NIST's basic

Silicon Glen's Semi-Success

The likes of Dewar's and The Glenlivet may be Scotland's best-known export, but whisky is no longer its largest: electronic products are. More than 20 years ago researchers at the University of Edinburgh decided their future lay in high technology. They persuaded Hughes Aircraft Company to build a silicon foundry at Glenrothes, north of Edinburgh, and the seeds of what was to become known as Silicon Glen were sown.

Today the glen—actually a large area of the lowlands around Edinburgh and Glasgow—is home to seven out of the world's 10 biggest computer manufacturers. Scotland Enterprise, a government development agency, proudly declares that Scotland now produces one third of the personal computers sold in Europe and a tenth of the world total. Manufacturers of other consumer electronics products, such as videocassette recorders and mobile telephones, also cluster in the region.

Yet Silicon Glen could never be mistaken for Silicon Valley. Despite growing revenues, the proportion of the Scottish electronics industry owned by British companies fell from 52 to 42 percent during the 1980s. More worrisome, from a Scottish point of view, is the trend in value added, the amount of work done in Scotland on product development and manufacturing. According to Ivan Turok, an economist at the University of Strathclyde in Glasgow, the share of value added in electronics gross output plunged from 39 to 24 percent between 1983 and 1989.

Turok notes that at least two factors probably explain why value added fell even as the industry grew. First, firms tend to import assembled components; only 12 percent of materials come from local sources. Second, many new plants are simply final-assembly operations, derisively called "screwdriver" plants by their critics.

One of them is Jim Rigby, controller of the telecommunications division of Hewlett-Packard in South Queensferry, near Edinburgh. The spacious open-design factory, which develops and makes microwave testing equipment, is responsible for its own product development and marketing and even for raising capital. That strategy is good for development, Rigby says. "Where Silicon Glen has failed is that not enough companies have brought in chains of

research and development and marketing expertise. If all you do is put something in a box, there'll always be somebody who can do it cheaper," he declares.

Crawford W. Beveridge, chief executive of Scottish Enterprise, concurs. The former top manager of Sun Microsystems, who returned to his home ground two years ago, cites the closings of major plants owned by Wang and Unisys to illustrate the dangers of being a branch economy. A new "inward investment" operation of Scottish Enterprise is trying to convince firms that they should establish integrated operations as well as manufacturing plants. Beveridge has also started a crusade to urge enterprising Scots to expand into export markets and to start up companies.

In California the proximity of science-based universities and major industry spawned countless enterprises and not a few legends. In Scotland, it did not. "There have been some spin-out companies, but none of them have got very far," says David Simpson, who established the South Queensferry Hewlett-Packard plant in the 1960s.

Now in semiretirement, Simpson is chairman of Spider Systems, a successful all-Scottish computer company that was founded in 1983. The company, which specializes in network software and hardware, now has revenues of \$30 million a year. "It is doing well by Scottish standards," Simpson says. But it is an exception.

One reason for the poor record on Scottish spin-off companies, Beveridge and Simpson agree, is the difficulty of finding financial backing for small start-up companies in Britain. To help bridge the capital gap, Beveridge is establishing a "one-stop shop" for fledgling enterprises, browbeating banks into backing ventures and setting up networks to advise novices. Optoelectronics is likely to be targeted for special attention.

Rigby of Hewlett-Packard has another idea that he has already put into practice. The Scottish Software Partnership Centre of the South Queensferry factory is a software incubator unit. Hewlett-Packard provides low-cost offices, marketing advice and hardware. Clients show their gratitude by developing software that runs on Hewlett-Packard machines. Harnessing Scottish thrift could yet be a winning strategy.

—Tim Beardsley

mission of performing measurement science. "It will be a real stretch to grow the program that fast," says John A. Armstrong, IBM's recently retired vice president for science and technology.

Whatever its eventual fate, the sudden growth of the ATP will certainly propel it into the center of arguments over how to pick and choose technologies. Today the program handles this problem by allowing any company, whether it be a biotechnology or a computer firm, to submit proposals to separate review teams that make evaluations of technical and business merit. This approach is intended to give the program a grass-roots quality; industry decides what it wants and pays about half of the cost.

But Lewis M. Branscomb, another former IBM vice president who is now a professor at Harvard University's Kennedy School of Government, says the way the program makes selections has given it a haphazard character. "Unless a single grant produces an extraordinary result, nobody knows, because the money is so scattered around," he points out. Program managers at NIST acknowledge that a more focused approach may be needed for certain projects.

Some aspects of the Clinton technology initiative are by no means revolutionary. The program recommends reorganizing but not killing funding for the Superconducting Super Collider and the space station. And it does not call for shutting any of the Department of Energy's huge national laboratories, whose main missions have included nuclear power and bombs.

The administration may give the national laboratories an extra \$100 million to \$200 million in the next budget year to enter partnerships with industry. The goal is for the laboratories to use from 10 to 20 percent of their funding on industry collaborations. The idea that these dedicated behemoths will smoothly switch to serving industry has its skeptics. "Their mind-sets aren't where they need to be," says William F. Brinkman, executive director of the physical sciences and engineering division at AT&T Bell Laboratories and a former vice president of Sandia National Laboratories.

The Clinton team plans to give a needed jolt to anemic private research spending by reviving a tax credit for R&D—which is projected to stimulate corporate research spending by 4 to 6 percent a year. By making the credit a permanent fixture of tax law, corporate managers are encouraged to take a more farsighted approach to planning. Relying on the tax credit in the past of-

ten proved difficult as it was set to expire before a product could be moved down the development pipeline from the laboratory toward the marketplace. There are other measures—investment tax credits and, for small business, capital-gains relief—within the Clinton economic package.

For large, multinational companies, which account for more than half of the \$80 or so billion in annual expenditures on research and development by private industry, the benefits of the credit may be offset by various corporate tax increases. For example, the economic plan dumps a tax break for royalty income that U.S. corporations receive from foreign companies or from their own overseas subsidiaries for rights to use a technology. At worst, the new rule might unintentionally encourage companies to move research operations abroad to avoid having to repatriate this income. "We're troubled that this may leave the R&D incentive landscape dramatically weaker," says Kenneth R. Kay, executive director of the Council on Research and Technology, an industrial and academic lobbying organization.

The administration must recognize that potentially its technology programs can become political pork barrels. The temptation to channel "earmarked" funds to favored projects or agencies will grow as the pot gains in size. "The typical congressman views a project for Silicon Valley the same way as one for downtown Dayton," says Claude Barfield, an analyst with the American Enterprise Institute. "It becomes an economic development program that does not go to the best and brightest but to the chairman of a subcommittee."

Avoiding a crippling debacle will require that Clinton and Gore prevent the infighting that is likely to emerge as congressional committees and different agencies try to establish niches in the new bureaucratic order. Easily overlooked may be the industries the programs are designed to serve. "I think there should be a technology summit to talk about some of these things," says Robert M. White, who served as an undersecretary of commerce during the Bush years.

One immediate goal of the high-tech administration is updating the lines of communication to 1600 Pennsylvania Avenue. Despite the vice president's fascination with digital superhighways, the new White House denizens found they had great difficulty hooking up to the Internet, the electronic communications network. "Most people here aren't even using it," laments Henry Kelly, who works in the Office of Science and Technology Policy. —Gary Stix

Bright Opportunity

Electroluminescent polymers illuminate an old institution

Back in 1989 Richard H. Friend, a physicist at the University of Cambridge, was trying to make transistors out of an organic polymer called polyphenylenevinylene (PPV). He knew that PPV had electrical properties similar to those of semiconductors and that it was a tough material ideal for making devices based on thin films. What he had not foreseen was that a thin layer of purified PPV emits a surprisingly bright light if a few volts are applied across it.

More than three years after that discovery, and after much feverish work, Friend and his backers believe they have developed an important technology for making flat-panel displays. Working through a specially created company, Cambridge Display Technologies, they intend to capture a large share of the market for cathode-ray tubes, light-emitting diodes, fluorescent lights, liquid-crystal displays and other such devices.

Although much engineering remains to be done, the researchers have made substantial progress since the initial discovery. Unmodified PPV emits only greenish yellow light, but Friend, along with Cambridge chemist Andrew B. Holmes and others, has been able to make light-emitting PPV derivatives that span the entire visible spectrum. Most significantly, the investigators have produced a good blue, a color that has proved to be the Achilles' heel of some display technologies. "Until we got a reliable blue, we thought we would only be able to make two-color displays," says Richard J. Artley, a consultant with the Generics Group, one of the investors in Cambridge Display.

Moreover, PPV displays can shine 10 times more brightly than can a television screen, Artley says. Ultimately, Cambridge Display wants to see full-color, bright PPV screens containing pixels of the three primary colors that are thinner than a back-lit liquid-crystal display. Emergency lights and signs, as well as indicators that have to be read in full daylight, also constitute prospective applications.

As if that were not enough, Cambridge Display workers maintain that PPV devices promise to be more efficient than back-lit liquid-crystal displays. Friend and his associates say they have now demonstrated a power-to-light efficiency of 1.5 percent—a 15-fold improvement on the first PPV devices three years ago—and Friend esti-

mates that “10 percent efficiency would be reasonable” as manufacturing techniques are refined to produce materials of higher purity. In comparison, inorganic light-emitting diodes have efficiencies that are typically less than 1 percent. The choice of materials used for electrodes is a key factor. Calcium, for example, has electronic properties that make it work well as one of the electrodes in PPV devices.

Perhaps PPV’s most crucial advantage, according to Artley, is that its chemical precursor can easily be handled in a solution. “You can ship it around the world like paint, put it on a substrate, then just heat it to 200 degrees Celsius,” he declares. Inorganic light-emitting-diode displays, in contrast, have to be fabricated in a vacuum at around 500 degrees C—a far more challenging operation. Because a PPV precursor can be painted onto a surface, individual picture elements can be made any size, from very big to very small. Regions can even be modified after painting so that they emit a different color. Two-color signs of PPV will, the materials’ boosters assert, be a relatively simple proposition. And because the polymer is flexible, it can be applied to curved or bendable surfaces.

Whether Cambridge Display researchers will be able to translate the natural advantages of PPV into commercial ones is, of course, another matter. Memories

are still keen in Cambridge of how another bright discovery made there—how to produce monoclonal antibodies—was not protected by patents, and consequently millions in potential royalties were lost. Friend and his colleagues have at least avoided that mistake. Broad patent claims covering any use of PPV-like polymers in displays have been filed. In addition, the company is now negotiating with big-name U.S. and Japanese consumer electronics manufacturers who might invest the capital needed to develop a color, high-resolution PPV display.

Despite the promising portents, PPV devices built so far exist only as small laboratory prototypes. Moreover, they have not yet been subjected to long-life-time tests, although informal observations (samples left on the window ledge at the Cavendish Laboratory where Friend works) suggest they are stable under ordinary conditions. PPV is certainly stable for up to 1,000 hours, Artley says, and similar materials are known to be stable for up to 10,000 hours. He expects no difficulty in reaching the target of 100,000 hours (over a decade) needed for long-lifetime products.

Then the market could, by Artley’s estimate, be more than \$30 billion at 1993 prices. Artley expects a simple prototype PPV display within 18 months. Then the bright lights could hit Cambridge. —Tim Beardsley

Three Swings

Hoping for a home run against multiple sclerosis

To the 2.6 million people around the world afflicted with multiple sclerosis (MS), medicine has offered more frustration than comfort. Time after time, researchers have discovered new ways to cure laboratory rats of experimental autoimmune encephalomyelitis (EAE), the murine model of MS, only to face obstacles in bringing the treatments to humans. It was thus cautious optimism that greeted three recent announcements of promising results in human MS studies.

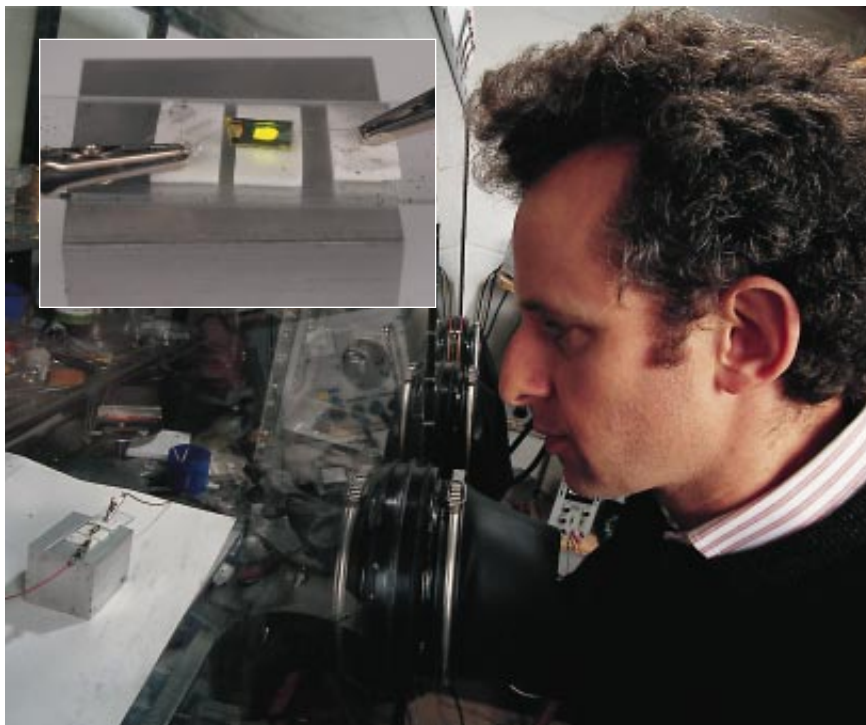
In March a federal advisory committee recommended that the Food and Drug Administration approve a new drug that appears to slow the progress of MS. The drug, trade-named Betaseron, was developed by Chiron and Berlex Labs, a subsidiary of German drug-maker Schering A.G. Meanwhile two other research groups—one at Stanford University School of Medicine led by Lawrence Steinman, who is also chief immunologist at Neurocrine Biosciences, and another headed by Howard L. Weiner and David A. Hafler of Harvard Medical School and backed by AutoImmune—reported results that they also hope will lead to therapies.

No one knows what causes multiple sclerosis, although there are statistical clues. About 60 percent of MS patients share a genetic tissue type called DR2, 60 percent are female, and many spent their adolescence in a region where the disease seems to be unusually common. Scientists have determined that MS is an autoimmune disease, like arthritis and insulin-dependent diabetes. When these illnesses strike, part of the body’s defense system turns against normal cells.

In MS patients, the immediate target of this self-destruction is myelin, the protective covering that electrically insulates nerves. The attack on myelin is led by helper *T* cells. These white blood cells are monomaniacal creatures: each one bears particular *T* cell receptors (TCRs) that enable it to bind and react only to a specific protein sequence, its antigen.

When a helper *T* cell bumps into its antigen, it generally sounds the alarm, reproducing furiously and calling in killer *T* cells to destroy what it assumes to be an invader. But sometimes *T* cells respond to one of the constituents of myelin, usually myelin basic protein (MBP) or proteolipid protein (PLP). In healthy humans, such *T* cells are switched off. In MS patients, they evidently are not.

Chiron’s researchers found that they



DAVID LEVENSON/Black Star

GLOWING POLYMER (inset) could be the key to a new generation of displays for computers, signs and other devices. Richard H. Friend of the University of Cambridge discovered the electroluminescent properties of polyphenylenevinylene in 1989.

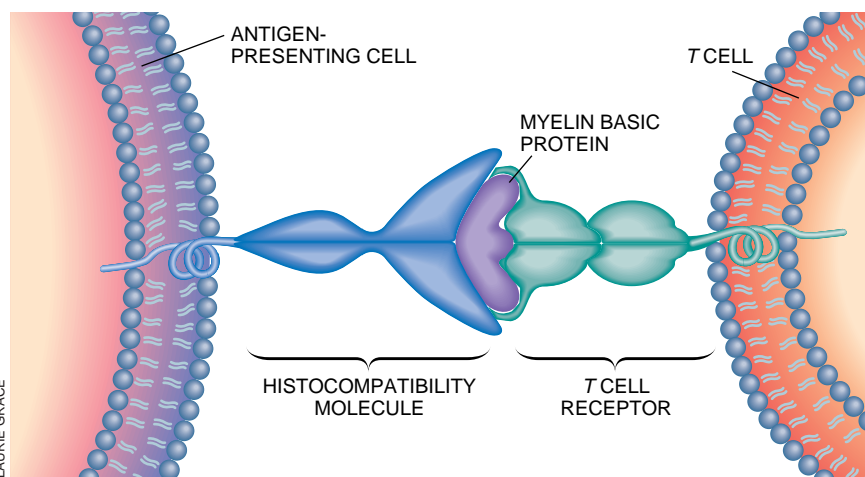
could reduce the severity of patients' relapses by injecting high doses of beta interferon. That immune system hormone, or lymphokine, stimulates suppressor *T* cells to call off immune attacks. In a clinical study involving 372 patients, those receiving high doses of beta interferon experienced one third fewer acute episodes than those given a placebo. After two years, the area of the brain in which lesions appeared decreased by 4.2 percent in those taking high doses of interferon; those given a placebo showed a 19.4 percent increase.

But Chiron's drug sheds little light on the primary point of debate among MS researchers: whether the disease is caused by a clan of rogue *T* cells that are genetically very similar to one another or whether a diverse population of *T* cells is at fault. Steinman's group endorses the former hypothesis. They took tissue from brain lesions in 16 deceased MS patients, extracted a handful of *T* cells from each sample and determined the DNA sequence of their receptors. Comparing the sequences with one another and with DNA data bases, they observed that *T* cell receptors that included a component called $\text{V}\beta 5.2$ appeared at least once in nine of the patients (and in seven of the eight who were DR2-positive).

That so many patients had at least some $\text{V}\beta 5.2$ -bearing *T* cells is important, Steinman says, because *T* cells that carry DNA specifying that same component have been shown to attack human myelin *in vitro* and murine myelin in some EAE rats. Steinman argues that $\text{V}\beta 5.2$ *T* cell reaction is "one of the major responses—if not *the* major response—in MS-affected brains."

If he is right, treatment should be relatively straightforward: with positive identification in hand, simply arrest the culprit. Neurocrine hopes to have a drug that will do just that. "It's an immunologic decoy," explains Kevin C. Gorman, Neurocrine's director of corporate development. "Now that we know exactly what *T* cell receptor is being used by the immune system to destroy myelin and we know precisely which portion of myelin basic protein the *T* cell is recognizing, we can block that interaction with a peptide." Gorman expects human toxicity testing to be under way by the end of 1994.

The peptide Neurocrine is betting on consists of fewer than a dozen amino acids—small enough to synthesize in a laboratory. Because it will bind with only a select few TCRs, the drug should avoid the side effects often caused by interferons and other substances that act on the immune system. "It's an exquisitely specific immunotherapy," Gorman boasts.



T CELLS that carry receptors for myelin basic protein (MBP) cause neurological damage in multiple sclerosis patients. When presented with MBP, they create clones of killer T cells that attack myelin. Researchers found a specific type of receptor component, known as $\text{V}\beta 5.2$, on some T cells extracted from MS brain lesions.

Too specific, argues Weiner: "A great deal of theoretical evidence indicates that the immune response involves other cells, such as those that react with PLP or that respond to MBP with different *T* cell receptors. Steinman's work is interesting, but I don't think the immune response in MS is so restricted that you can affect it by targeting just one subset of cells."

Hafler, who has long promoted the idea of a restricted immune response in MS, nonetheless tends to concur. "It's clear that *T* cell receptor usage varies from individual to individual." He points to his own work sequencing TCRs found in the blood of MS patients. "We have cloned more than 100 *T* cells that react to MBP, yet we have found only one $\text{V}\beta 5.2$. To argue that this *T* cell alone causes MS is kind of ludicrous."

Weiner and Hafler think they have a better idea: feed myelin to MS patients. Mammals tend to tolerate antigens better when they are ingested rather than injected. Weiner found that when he gave rats shots of MBP, they were quickly crippled by EAE; when he then fed them the protein, they got better. How this response occurs is not entirely clear, but Weiner's animal research suggests that the ingested protein is processed by lymph nodes in the small intestine, which then generate MBP-specific suppressor *T* cells. When they reach the brain and encounter MBP, the suppressor cells sound the retreat.

In a recent trial using 30 early-stage patients with relapsing-remitting MS (the most common form of the disease), Weiner's group gave half of the patients daily capsules of bovine myelin; the other half were given a placebo. During the year-long study, 12 members of the con-

trol group suffered exacerbations of their MS; only six myelin-treated patients did. The condition of six of the treated patients improved significantly over the year, true for only two control patients. And blood tests showed a significant decrease in the number of MBP-reactive *T* cells in treated patients. No side effects or opportunistic infections were observed.

Hafler warns against overenthusiasm: "This is a clinical experiment; by no means is it proven." But, he adds ebulliently, "the beauty and importance of this is that one does not need to know the inciting antigen. We may have come up with a very simple way of turning off immune responses." Weiner points out that the treatment is by its nature already in the ideal form for a drug: a natural, orally administered protein.

AutoImmune hopes that fact will smooth the path toward approval. Robert C. Bishop, the company's chief executive officer, says AutoImmune plans to repeat the myelin trial with 200 to 300 DR2-negative patients, segregating the males and females. The 30-patient trial showed the oral treatment to be much more effective in men than in women. But Weiner emphasizes that the number of patients in each group was too small to draw any meaningful conclusions.

AutoImmune is not waiting for definitive confirmation. This fall Bishop expects to release results of a 60-patient trial in progress for rheumatoid arthritis (using chicken collagen). Similar treatments for uveitis, an eye disorder, and insulin-dependent diabetes will soon be tested. And all involved expect even better results in MS trials when recombinant human MBP, already cloned, is approved as a drug. —*W. Wayt Gibbs*

Garbage in, Gravel out

Plasma torches transmute waste into harmless slag

The casual reader—inundated by reports of toxins leaching from landfills, orphan garbage barges aimlessly wandering the oceans and incinerators belching carcinogens—might be forgiven for wishing we could simply zap our refuse and turn it into something useful. That idle wish is spawning an industry. Several companies are developing systems to transform garbage into slag, although the technology owes more to 19th-century steelmaking than to *Star Trek*.

The systems are based on a new generation of highly efficient plasma arc torches, simpler forms of which have been used in metallurgy since the 1880s. They pass a strong electric current through a rarefied gas, ionizing it to produce a flame that, at up to 8,000 degrees Celsius, is much hotter than any fire.

Faced with such intense heat and denied oxygen, matter does not burn; it dissociates, in a process called pyrolysis. Toxic hydrocarbons break down into simple gases. Metals melt and disperse. Soil solidifies. Some see in plasma pyrolysis a technological panacea for America's landfill woes. "This could be an ultimate solution to municipal solid-waste disposal," announces Louis J. Circeo, director of the Georgia Insti-

tute of Technology's Construction Research Center.

As justification for his exuberance, Circeo points to a 1990 study funded by the Electric Power Research Institute (EPRI) and the Ontario Ministries of Energy and Environment and performed by Resorption Canada Limited (RCL). When workers at RCL turned a 150-kilowatt plasma torch on shredded garbage, they found it reduced the mass of the trash by 80 percent; volume was cut more than 99 percent. The missing mass emerged as a fuel-grade gas composed mostly of hydrogen, nitrogen and carbon monoxide; when burned, the gas could generate 25 percent more electricity than the torch consumed. Whatever metallic slag remained was safely inert.

Despite such promising results, plasma torch manufacturers express more modest ambitions for the technology. "People are always looking for a better way of doing things," says David Camacho, marketing director of Plasma Energy Corporation (PEC) in Raleigh, N.C. "But most would rather take evolutionary steps than revolutionary steps."

So far those taking steps toward plasma pyrolysis have been fleeing hazardous waste regulations and the costs they impose. In Japan, Mitsui and Company has added PEC torches to four incinerators to convert their effluent from ash, which is classified as hazardous, to slag, which is not. The city of Bordeaux, France—pleased with the results of Circeo's tests on two tons of ash the city shipped to Georgia Tech—decided to

install a plasma furnace next to its incinerator combustor. Bordeaux hopes to save more than \$2 million each year by selling slag to a cement producer instead of shipping ash 500 kilometers to an expensive state-owned landfill.

But in North America, where most incinerator ash is not treated as hazardous waste, plasma vitrification has yet to leave the lab. One reason is cost: pyrolysis is still about twice as expensive as landfilling. "Until the cost to dump goes up to \$150 or \$200 per ton, I don't see widespread use of plasma technology for municipal waste," Camacho says. "But biologically or chemically hazardous waste is already in that neighborhood."

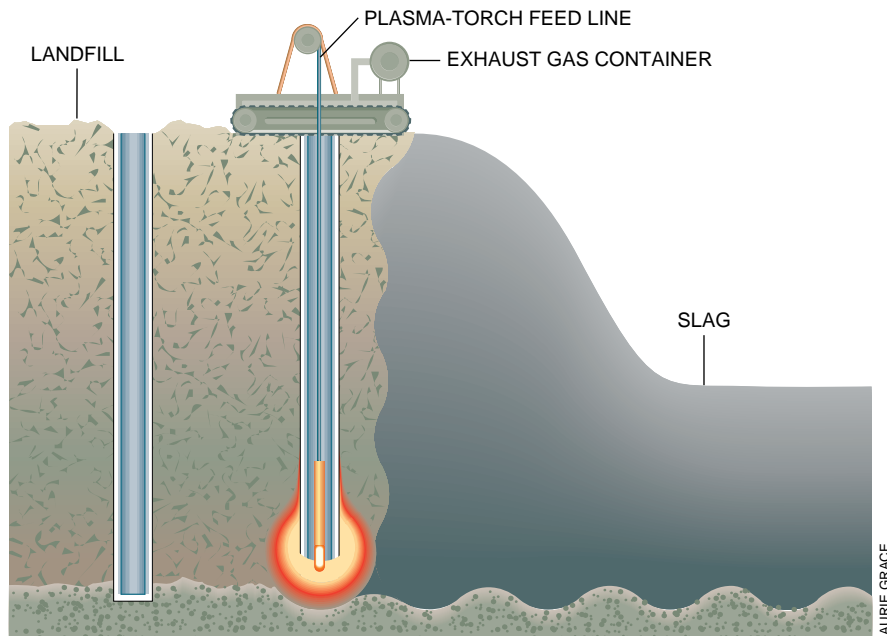
It is hardly surprising then that the first commercial plasma furnace in the U.S. will probably treat medical waste. Kaiser Permanente has teamed up with EPRI and several southern California utilities to build a pilot plant later this year at its San Diego hospital. Kaiser plans to use a 500-kilowatt torch to vitrify 12 tons of waste per day, while pumping the gas emissions into a fuel cell to generate electricity.

But W. Eugene Eckhart of EPRI notes that recently adopted regulations, which force incinerators to cut emissions more than 90 percent by 1995, may push incinerator operators to add cleaner-burning plasma furnaces and to use their combustors just for preheating. "This is a relatively conventional approach using proven technologies and well-understood economics," Camacho adds.

Proponents of plasma vitrification already look ahead to the eradication of landfills. Not everyone thinks this is feasible, however. "Cleaning up a leaking landfill would be a monstrous undertaking. We haven't yet dreamt of that large a scale," Eckhart moans. But Circeo has. He envisions mobile torches crawling over dumps, vitrifying waste a column at a time. Last year Circeo's proposals attracted the attention of Semen S. Voloschuk, the Russian minister charged with cleaning up the environmental havoc wreaked at Chernobyl. After visiting Atlanta to see the technology firsthand, Voloschuk has set about garnering money to set up a trial on site.

Circeo is also soliciting funds, though for a much more humble project. He wants to build a "showcase" plant to vitrify the estimated 20 tons of garbage that will be produced each day during the 1996 Summer Olympics. If he can raise the \$10 million necessary for construction in time, Circeo may well be the first on his continent to prove that this new brand of alchemy is not only technically possible but economically viable.

—W. Wayt Gibbs



PLASMA TORCH may be a solution to leaky and overflowing landfills. Lowered down a borehole and fed electricity, coolant and gas, the torch would melt and vaporize waste into a harmless, glassy slag. Exhaust gases could be burned to generate power.



Why Do Some Industries Pay Better?

Standard economic theory says that employees who do the same job, have the same skills and work under the same conditions should all earn about the same wage. No employer can offer a salary less than the market rate and expect an applicant to accept it, and none need pay more as long as other workers are willing to take a job at that price. So why do jobs of all kinds in the petroleum industry, for example, pay 30 percent more than the U.S. average for jobs with equivalent skill requirements? And why do those in private household services make 36 percent less than average?

Economists in the U.S. first noticed this deviation from classic market behavior more than 40 years ago. Since then, similar patterns have been seen in many other countries. The same industries pay high (or low) wages everywhere. Moreover, the pattern has been stable over time. "Your grandma knew which firms had good jobs with high wages, and these same firms are paying high wages today," observes Erica L. Groshen, an economist at the Federal Reserve Bank of Cleveland.

The debate over the causes of these persistent wage differentials has been heightened by its potential impact on Clinton administration economic policy. Some economists advocate subsidizing high-wage industries so that there will be more well-paid jobs. Others believe the same result can be achieved by narrowing the gap between high- and low-wage versions of the same job. These alternatives have serious implications for so-called industrial policy. The first option corresponds to the long-discussed notion that the government should pick "winners" in the high-tech sweepstakes and leave "loser" industries to fend for themselves. The second harkens closely to Secretary of Labor Robert B. Reich's call for wide-ranging improvement of workers' skills.

According to traditional economic thinking, wage differentials should reflect differences in employees' skills,

education and effort. Wages might also vary because of regional differences in the cost of living; industries located mainly in the South, for example, generally pay less. And workers may be paid a premium (or accept discounted wages) for doing the same job under unusually bad (or good) circumstances.

But none of these factors explains why employees in breakfast-cereal manufacturing, cosmetics or petroleum products should be paid more than their counterparts at the same skill level in garment manufacturing or education services. Not only do engineers in the petroleum industry earn more than they would in another sector, so do service employees. It seems unlikely that washing floors or replacing light bulbs at Exxon should require such extra skill or effort as to command a 30 percent premium over the wages of the same job at the corner dry cleaner. Fur-

Replacing light bulbs at Exxon pays 30 percent more than the same job at the corner dry cleaner.

thermore, people who move from one industry to another tend to find that their earning power changes suddenly even though there has been no change in their skills.

Recently some economists have turned to the idea of "efficiency" or "equity" wages to find the explanation for differentials. They argue that some firms will "overpay" their employees to keep them from slacking off, quitting or becoming disgruntled. Equity means that firms pay wages workers will perceive as "fair" (in particular, above market wages when profits are high). Such a policy is "efficient" for the company because it cuts the costs of close monitoring and supervision, reduces turnover and enables the company to fill vacancies quickly.

There is evidence, moreover, that efficiency wages are more than a theorist's figment. Alan B. Kruger of Princeton University found, for example, that fast-food entrepreneurs who managed their own restaurants pay lower wages

than those who hired managers for each site. He takes this difference as evidence that the higher pay scales serve as insurance for those owners who are unable to monitor their employees' performance directly.

High-wage industries, meanwhile, tend to be capital-intensive, so that labor represents a small fraction of their costs. They are also those dominated by a few firms that can effectively set prices. Indeed, firms often compete over product design or advertising rather than price. Furthermore, the typical company in a high-wage industry is very profitable. It not only has more leeway to pay efficient wages, but it may also have more to lose from worker defection.

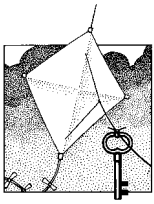
If efficiency wage theory is correct, then the U.S. ought to be able to create more high-paying jobs simply by giving aid to companies in industries that fit the appropriate criteria (not to mention helping lower-paying service industries to restructure themselves into more concentrated, capital-intensive forms). Groshen and Kruger both assert, however, that economists are not yet in a position to recommend that such policies be implemented. The conundrum is that no one really knows whether such industries are highly profitable because they pay well or whether they pay well because their businesses are highly profitable.

David G. Blanchflower of Dartmouth College thinks public policy could be used to stimulate employment in high-wage jobs, but he warns that "there will be difficulty in knowing exactly where the good jobs will be." As a result, he favors subsidies to firms that offer job training rather than to those in specific industries (thus coming down on the side of trickle-up profitability).

Beyond the issue of creating remunerative jobs, however, there lies the problem of the growing gap between high- and low-wage workers in the U.S. "There is nothing natural about these industry wage differentials," says Edward N. Wolff of New York University.

Germany, he points out, has smaller industry differentials than does the U.S., even though average wages there are quite high. "We should think hard about the wage-setting institutions we have in this country and about how we can bring our differentials more in line with those in European countries," he contends. —*Judith Fields and Paul Wallich*

JUDITH FIELDS is a fellow at the Jerome Levy Economics Institute at Bard College in Annandale-on-Hudson, N.Y.



Mapping to Preserve a Watershed

Every map reflects the interests of its makers and users. A road map guides drivers to their vacation spots. A zoning map shows planners where they may construct houses, parks and shopping plazas. A congressional district map determines a representative's constituency. By depicting information in a graphic way, maps become potent instruments of communication and persuasion [see "The Power of Maps," by Denis Wood, page 88].

We had the opportunity to do some amateur cartography with the Westchester Land Trust, an environmental organization in Westchester County, N.Y. The maps we created illustrated a watershed system that went beyond political boundaries. We undertook the project to show how different types of land use affect the quality of the Mianus River, the main source of water for 130,000 residents in contiguous parts of New York and Connecticut. The maps highlighted environmentally sensitive areas and revealed those sections of the river that are particularly vulnerable to pollution.

A watershed, the area drained by a specific stream or river system, resembles a bowl whose rim is marked by hilltops or ridges and whose base is a river or other body of water. Besides providing water for drinking and recreation, watersheds are essential for wildlife habitats, climate control, agriculture, industry and sometimes even transportation and electric power. Each of us lives in a watershed, although no

one actually sees it. A map is probably the best way to make a watershed visible. It does so by shifting our frame of reference from man-made districts to a landscape defined by physical boundaries. It reveals the anatomy and physiology of the region.

Constructing the watershed base map meant combining information from several different kinds of maps. First, we used state highway maps to determine the path and full length of the Mianus River. We found that this 37-mile-long river crosses five towns, two counties and two states before emptying into Long Island Sound.

Once we knew where the river was and where it went, we assembled a large-scale paper map of the Mianus from large-format county road atlases. Such atlases are the ideal choice from which to make watershed maps because, in addition to being inexpensive and readily available, they render information at a useful scale and depict the natural and man-made features in color. We compiled enough pages to show all the town, county and state borders, and we carefully matched grid lines on the borders of the maps and connected the ends of the roads. Our finished base map required 19 atlas pages. Because we used both sides of some pages, we needed two copies of each atlas.

Also used were drainage basin maps, which show the location of the drainage divides. The divides define the watershed boundaries—the rim of the bowl down which water drains into the Mianus River. Drainage basin maps show rivers and streams as solid lines and drainage divides as dotted lines. We needed four drainage basin maps to make a complete composite of the Mianus drainage basin: one Westchester drainage basin map and three U.S. Geological Survey quadrangle basin maps for southwestern Connecticut (the Mianus basin includes part of Fairfield County, Conn.). Drainage basin maps are generally available from some local government agencies. We obtained ours from the Westchester County Planning Department and from the Connecticut Department of Natural Resources.

If drainage basin maps had been unavailable, we could have used topographical maps, or topos. In addition to

showing rivers and streams, topos represent the hilltops and ridges as closely spaced lines. Finding the watershed boundaries on a topo requires an extra step. One would have to plot the drainage divides by connecting the highest elevations shown—a laborious and time-consuming task.

Because a drawn river is by nature long and skinny, we had to be careful that our paper map did not get too big to handle. We decided that the maximum size would be 48 by 36 inches. As the Westchester part developed, it became clear that its scale of three inches to one mile would produce a base map of about the right size. We adjusted all other maps to this control scale.

We quickly found out how difficult it is to work with rolled maps. So we made a map storage case by taping together two large pieces of scrap cardboard to keep the work flat. The finished paper maps were kept rolled for easy transport and long-term storage.

Making composite maps is easy if all the pieces are drawn to the same scale. All our maps were drawn to different proportions, so we had three of them rescaled ("statted") to our control scale of three inches to one mile. A copy shop that reproduces blueprints was able to stat large maps as well as copy them. A small map must be enlarged in sections and then reassembled.

To help the copy shop, we drew a precisely measured control line on the bottom of each map, showing the inch-to-mile ratio. One map was not printed at the scale indicated by its legend, so we had to draw a corrected line. Another had no legend at all, so we calculated the scale arithmetically.

The copy shop statted the maps by enlarging the control lines until each was three inches long. Once the scales matched, we cut and taped the pieces together to make two maps: one of the river and one of the drainage basin. We laid a 48- by 36-inch sheet of Mylar over the drainage map and traced the outline of the Mianus watershed.

MIANUS RIVER WATERSHED serves residents in New York and Connecticut. This composite map was generated by tracing the drainage basin (black outline) on a transparent overlay. The discrepancies between county borders were caused by improper scaling of the Connecticut map.

KAREN JESCAVAGE-BERNARD and ANDERS CROFOOT are Westchester Land Trust associates. Jescavage-Bernard writes articles that focus on environmental issues, and Crofoot is a consultant who specializes in computer-aided drafting and visualization. The Westchester Land Trust, a nonprofit organization, was formed to protect open space and to promote responsible use of land and water in Westchester County, N.Y. The watershed mapping project was carried out under the direction of Crofoot and Bice C. Wilson, an architect and planner. Alice Bamberger directed the full project, which included a water-quality monitoring program, a public awareness campaign and a school-based educational program.

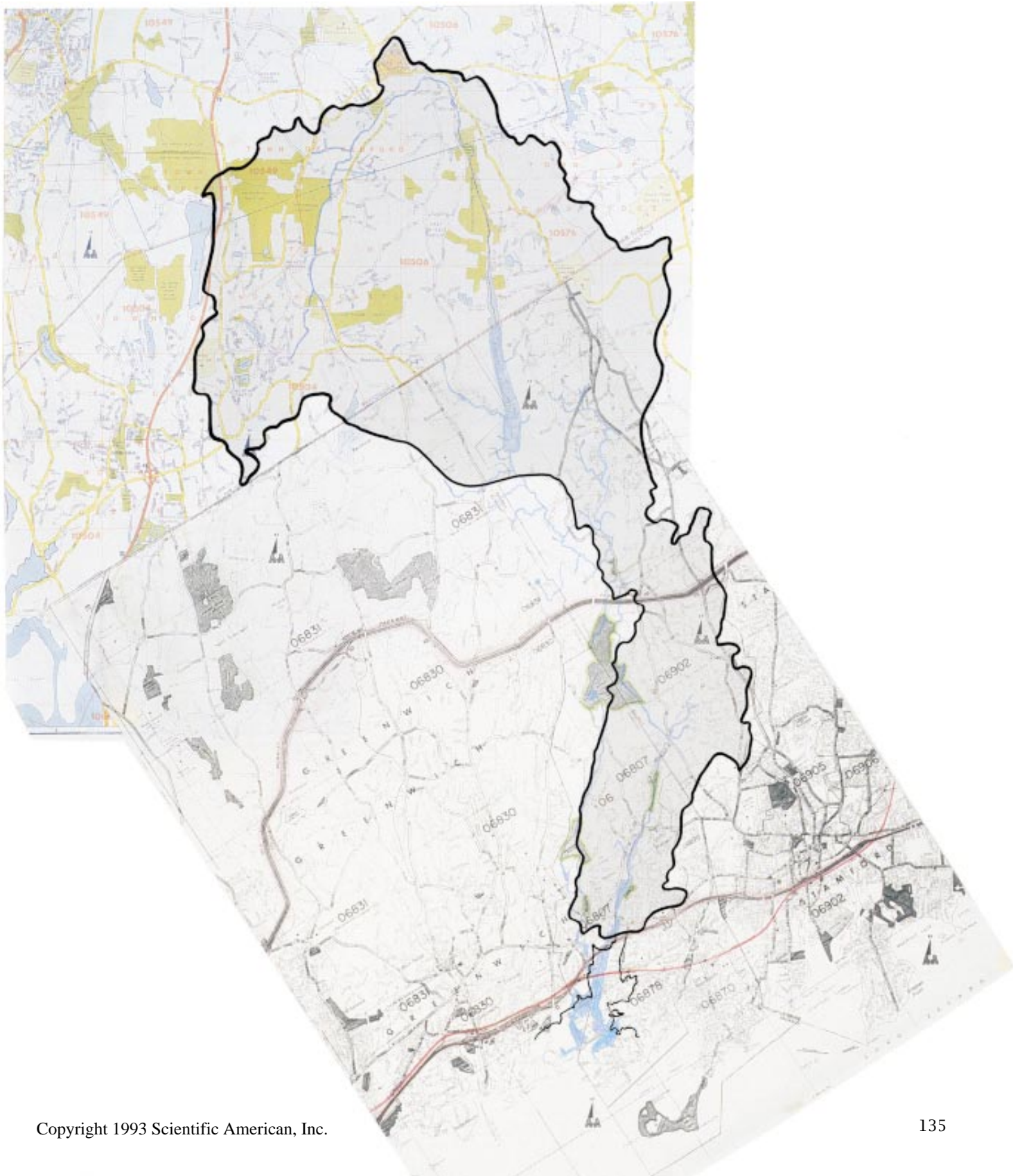
Although the Fairfield County chunk had been correctly statted to scale, it did not line up with the Westchester chunk at the county border: roads, rivers and even the reservoir were slightly askew. We therefore adjusted the river map to make the pieces fit as closely as possible. We suspect that incorrect scal-

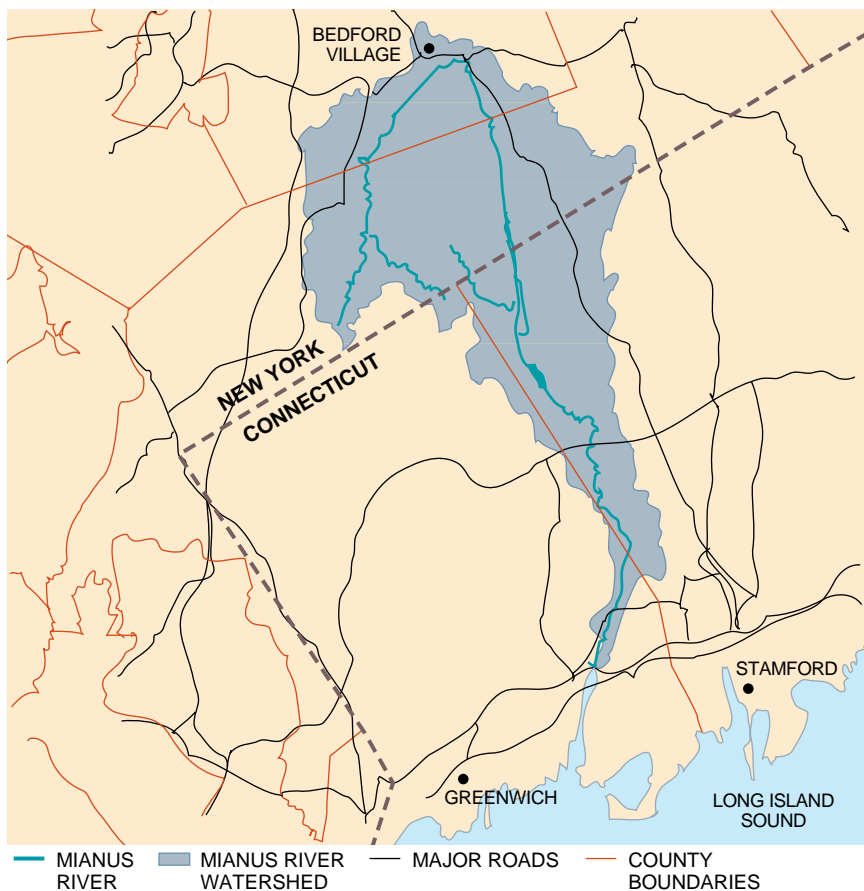
ing is more common than is generally thought; one should maintain a healthy skepticism toward the presumed accuracy and objectivity of maps.

Because the copy shop could not produce statted maps in color, we lost valuable information from the Fairfield section. Blue and green waterways and

parks just disappeared from the black-and-white copy. We reoriented ourselves by hand-coloring major features with felt-tip markers.

When we placed the transparent drainage basin Mylar over the paper map of the river, the Mianus watershed immediately appeared on the landscape.





COMPUTER-GENERATED MAP, created by a geographic information system (GIS), offers a clear and accurate picture of the Mianus River drainage basin. GIS technology can also incorporate such nonmap information as statistics and satellite imagery.

We saw a completely new, more comprehensive picture: the relations among the boundaries of the natural region, the man-made region of roads and the abstract region defined by politics. Although our simple experiment ended with the creation of this base map, one could build a much more sophisticated watershed map by combining the transparent overlay of the watershed with other kinds of maps, such as those depicting hydrography, topography, soils, property taxes and land use.

The total cost for this project was approximately \$145: about \$60 for the maps, \$24 for the Mylar and \$60 for reproduction. You can save the copying costs if you can obtain maps drawn to the same scale. Paper maps from government sources may be free.

A more high-tech mapping procedure, perhaps best suited as a team project, would be to use a computer and software called a geographic information system, or GIS. A GIS can combine maps of wetlands, slopes and soils with such nonmap data as aerial photographs, satellite images, statistical tables, water-quality measurements and census records. GIS technology enables those who

have no training in cartography to make maps and models. Its jazzy graphics allow one to see, analyze and anticipate the effects of alternative courses of human activity on the watershed. The GIS used for the Mianus project was ARC/INFO, which is produced by Environmental Systems Research Institute, a software company in Redlands, Calif.

Although GIS technology keeps becoming cheaper and more widely available, it is still so expensive that it is used mainly by government departments and large corporations. Anyone tackling this version of the project must have access to a source of digitized data, a GIS and trained personnel who are willing to dedicate the time to design and generate a watershed base map. Furthermore, amateur mapmakers need a personal computer able to run PC Paintbrush or equivalent graphics software. At least one team member would have to be conversant with the software and be able to design a map with it. Although a color printer would produce the best maps, any printer that can generate graphic patterns can be used to create a base map.

Amateur cartographers should know

some GIS jargon. A GIS works with digitized data, not maps. Each layer of data is called a coverage. The U.S. Geological Survey subdivides coverages into geographic regions called quads (short for quadrangles). The Mianus watershed spreads over four quads. We needed four coverages for each quad. The coverages were for municipal boundaries, watershed boundaries, roads and rivers. Coverages are available from the county planning department or the state department of natural resources.

Because learning to use any GIS software package requires a significant commitment of time, amateur mapmakers should consider working with a government or academic planning department that has a GIS. Its staff would help design and generate a preliminary watershed base map. After assembling the coverages, they would display the base map on the screen for the project team to review. The base map would then be saved on a floppy disk in a file format on which the mapmakers could work with their own graphics software.

Both the simple and the advanced exercises produced informative maps, but each contained different messages. The simple map was troublesome to make because the data were hard to find and harder still to combine. Even with precise cutting and fitting, major discrepancies are glaringly apparent. The accuracy and clarity of the GIS map are a major leap forward. Unfortunately, it is also much more expensive. Each coverage costs about \$20, and we needed 16 of them. Then there are the costs of the personal computer and the graphics software.

Of course, variations on a watershed map can lead to other kinds of applications. A community group can plot recorded sightings of migratory birds on our simple base map to make a persuasive case for preserving or restoring seasonally critical habitats. A town can identify a suitable site for a new well by taking our computer-generated base map and plotting existing water service districts, surface and subsurface geology, area soils, zoning and land cover.

For additional information about making a watershed base map, contact the Westchester Land Trust, 31 Main Street, Bedford Hills, NY 10507.

FURTHER READING

THE MAP CATALOG: EVERY KIND OF MAP AND CHART ON EARTH AND EVEN SOME ABOVE IT. Joel Makower. Vintage Press, 1986.
GIS WORLD. 12 times per year. 155 East Boardwalk Drive, Suite 250, Fort Collins, CO 80525. Telephone: (303) 223-4848.



The Legacy of *Homo sapiens*

ORIGINS RECONSIDERED: IN SEARCH OF WHAT MAKES US HUMAN, by Richard Leakey and Roger Lewin. Doubleday, 1992 (\$25).

reviewed by Christopher B. Stringer

Richard Leakey's own origins are by now well known to most people interested in the earliest history of our species. At the time of his birth, his parents, Louis and Mary Leakey, were already famous for their archaeological and paleontological discoveries in East Africa. When he was still a teenager, they achieved even greater prominence through their excavations of early human fossils at Olduvai Gorge in Tanzania. Although for a number of years Richard turned away from the field of human beginnings, his origin became his destiny. Despite a lack of academic training, he has inherited his father's mantle as (in the words of the book jacket) "the world's most famous living paleoanthropologist."

This book is the third collaboration between Leakey and science writer Roger Lewin, a relationship that stretches back to *Origins* in 1977. A lot has happened in paleoanthropology over the past 15 years, and a lot has happened to Richard Leakey. He nearly died from kidney failure in 1979 but was saved by a transplant from his brother, Philip. He recovered and continued leading fossil-hunting expeditions as director of the National Museums of Kenya until 1989. At that time, his life entered an even more challenging and dangerous phase. He became director of the Kenya Wildlife Service, in charge of combating (by virtually any means necessary) the rampant poaching of elephants and rhinoceroses. Through it all he has continued to be involved in paleoanthropology, collaborating with such scientists as his wife, Meave Leakey, and Alan Walker, as well as with writers such as Lewin.

The *Origins* volume capitalized on a succession of marvelous finds made in the 1970s in East Africa. The most famous are the Kenyan "skull 1470," assigned to the primitive human species *Homo habilis*, and the "skull 3733," assigned to *H. erectus*, both found by



Homo erectus



Homo sapiens neanderthalensis



Homo sapiens sapiens

Richard's own team; the Laetoli discoveries, including trails of footprints, found by Mary's group in Tanzania; and the excavations of the French and American workers at Hadar in Ethiopia (which include the "Lucy" skeleton).

The study of human origins seems to be a field in which each discovery raises the debate to a more sophisticated level of uncertainty. Such appears to be the effect of the Kenyan, Tanzanian and Ethiopian finds. The rich new information has opened and gradually widened a rift between the Leakeys and Donald Johanson and Timothy White as well as other American workers involved in the discovery and interpretation of fossils. True to the traditions of the field, the arguments swirl around the questions of the correct classification of the fossils and of the presumed relationships between the species of humans and prehumans. Was one group ancestral to another, or were they contemporaries, perhaps contending for evolutionary advantage? Such discussions have made the field fascinatingly contentious, and they have probably won it more newsprint than it might otherwise have received.

In *Origins Reconsidered*, Leakey tries to avoid resurrecting such disputes. Yet he still believes he was right to question the validity of Johanson and White's creation of the new prehuman species *Australopithecus afarensis* to encompass the whole range of Hadar material, including Lucy. Indeed, he suggests that even Johanson and White may now doubt that only one species of hominid was present in East Africa three million years ago.

A discovery made in 1985 at west Turkana in northern Kenya serves as an important part of the foundation for his reasoning. As with so many of the Turkana fossils, this one was found by someone (in this case, Alan Walker) looking for bone eroding out of the ground. Many fragments were discovered and painstakingly reassembled. From this work, a new kind of australopithecine emerged in the shape of the "black skull" (so called because of its distinctive fossilization). The skull combines some primitive aspects found in *afarensis* with many more of those of the specialized later robust australopithecines, but its importance also lies in its estimated age of 2.5 million years.

CORRESPONDENCE

Reprints are available; to order, write Reprint Department, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111 or fax inquiries to (212) 355-0408.

Back issues: \$6 each (\$7 outside U.S.) prepaid. Most numbers available. Credit card (Mastercard/Visa) orders for two or more issues accepted. To order, fax (212) 355-0408.

Index of articles since 1948 available in electronic format. Write SciDex™, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111, or fax (212) 355-0408.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the fee of \$3.00 per copy of each article plus \$0.50 per page. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 27 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/93. \$3.00 + 0.50.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017-1111, or fax (212) 754-1138.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. 51537. The date of the last issue of your subscription appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (mailing label, if possible) and your new address.

This age is close enough to Lucy's three million years to suggest to Leakey and others that the separate evolutionary lineage of the robusts stretches back at least as far as her time. Moving on to about two million years ago, Leakey finds yet more reason to question the conclusion of Johanson's research teams that an ancestral relationship unites the two forms. He also disagrees with the assertion by his American rivals that the partial *H. habilis* skeleton they found at Olduvai belongs to the same species as "1470." (He is certainly not alone in his doubts about this, because the body skeleton of the Olduvai find seems far more like an australopithecine than those from Kenya that have been linked to "1470.")

The second half of *Origins Reconsidered* describes the major shift of interest in the field over the past 10 years toward recent events in human evolution and, in particular, the evolution of our own species, *H. sapiens*. The authors describe in some detail the marvelous 1984 discovery in west Turkana of the skeleton of a boy who was about 12 years old at death and is about 1.6 million years old geologically. He is regarded by Leakey and his colleagues as representing the early human species *H. erectus* (first known from Java and China), and the completeness of his skeleton far exceeds later examples. As a result, it gives us an important starting point to consider the subsequent evolution of our own species from *erectus*. The authors paint an atmospheric picture of the country around Lake Turkana and of the excitement of an excavation in progress. This discovery is the pretext for much further discussion of the boy himself, in life and death, as well as the history of his species.

The fate of *H. erectus* seems to have been to evolve into *H. sapiens*, but the details of that process are hotly disputed by the experts. Some believe (as Leakey, and *Origins*, favored in 1977) that *H. erectus* developed into *H. sapiens* across the entire Old World range of the species. The alternative to this view, the "Noah's Ark" hypothesis of 1976, has become the rival "out of Africa" theory of today. In its contemporary form the hypothesis holds that only one kind of *erectus* evolved into our species, and this was the African form. The line of *erectus* may have become the Neanderthal line in Europe, but it ultimately died out, as did all the other lineages of *H. erectus* outside of Africa.

The idea of a recent and exclusively African origin for modern humans

clearly appeals to Leakey now (no doubt encouraged by his co-author's known enthusiasm for it), although he is still very cautious. In the interests of disclosure, I should underscore the fact that I am a much stronger advocate of "out of Africa" than is Leakey [see "The Emergence of Modern Humans," by Christopher B. Stringer; SCIENTIFIC AMERICAN, December 1990]. But equally, I am also a little more cautious about some of the evidence than I am quoted as being in *Origins*. I have never said that the rather modern-looking Border Cave fossils from South Africa could be 130,000 years old; in fact, I co-authored a paper in 1990 indicating that they were probably younger than 90,000 years. Perhaps Lewin has confused Border Cave with the Omo Kibish skeleton from Ethiopia, which I do regard as a very early example of a modern human from Africa.

Several chapters on the possible evolutionary origins of aspects of modern human behavior such as language and consciousness conclude the book. In them the authors draw on broad expert opinion from anatomists, primatologists, linguists, archaeologists and psychologists. In a final section, which has the same title as the book, Leakey and Lewin deal with the presumed inevitability of the evolution of people like us (they strongly doubt it); the supposedly unbridgeable gap between us and the rest of nature (they believe the gap certainly exists but has developed only gradually over the past two million years); and finally the "Sixth Extinction." With that topic the book comes full circle to Leakey's prologue about his new responsibilities for Kenya's wildlife. The term "Sixth Extinction" refers to the apparent and alarmingly rapid disappearance of species as a result of human economic activity. The authors take the view that this event, the sixth such crisis in the history of life on the earth, is unique in that it is the result of the activities of a single species. The crisis is also unique in that it is a potentially controllable process. Our humble and serendipitous origins, the authors argue, make us part of nature. We should not be acting as if we are its arrogant and selfish overlords.

Such an observation provides a fitting end to this very readable book. Yet the work is not free from flaws. For example, the authors are not completely successful in avoiding the temptations that beset those of us who seek to write the history of events in which they participated. Leakey and Lewin are occasionally somewhat wise after the event. Readers would probably be bet-

ter served if the authors were more frank about how they (like many of us) have had to change position on a number of important issues in paleoanthropology. Furthermore, the new book is rather longer than its predecessor but not so well illustrated; nor does it contain even a token list of suggested further reading.

Nevertheless, as a very approachable introduction to some of the current issues in the debate over the early history of humanity, it is a worthy successor to *Origins* and an excellent starting point for anyone who comes to the field innocent of its past. Whatever one may think of the Leakey family's extensive contributions to theory in the field, there is no doubt that the debate could not have advanced to its present level were it not for the evidence that they have coaxed and prized from the dusty African soil where we began.

CHRISTOPHER B. STRINGER is a member of the Human Origins Group of the Natural History Museum in London.

A Gallery of Landforms

COLOUR ATLAS OF THE SURFACE FORMS OF THE EARTH, by Helmut Blume. Translated by Björn Wygrala. Edited by Andrew Goudie and Rita Gardner. Harvard University Press, 1992 (\$75).

Nowadays even the couch potato, eyes on the tube, knows the look of the global landscape pretty well. This thin, serious volume by a reflective globetrotter from Tübingen arrays more than 200 small, sharp photographs that he himself has taken on all continents save icebound Antarctica and on many a remote island as well. He drew only on a little help from his friends and from the literature at large. With each scene, he includes a thoughtful paragraph to explain the nature and origin of the visible landform. Every view stands among a platoon of 20 or so of its geomorphological counterparts, and a brief introduction to each category unites the forms we see. A few introductory—though not elementary—pages and a couple of awesomely detailed specimen maps set the processes within the broadest geologic chronicle. You will inspect a tapestry of sere red deserts, storm-cut white sea-cliffs, fertile green valleys, even wide pits opened dark and deep to mine Mesabi ore.

Evidently, diversity rules, yet the fas-

To preserve your copies of SCIENTIFIC AMERICAN

A choice of handsome and durable library files or binders. Both styles bound in dark green library fabric stamped in gold leaf.

Files Each holds 12 issues. Price per file \$7.95
• three for \$21.95 • six for \$39.95

Binders Each holds 12 issues. Issues open flat.
Price per binder \$9.95 • three for \$27.95 •
six for \$52.95

A GREAT GIFT IDEA, TOO!

(Add \$1.00 per unit for postage and handling in the U.S. Add \$2.50 per unit outside the U.S.)

To: Jesse Jones Industries, Dept. SA, 499 East Erie Ave., Philadelphia, PA 19134



Send me _____ SCIENTIFIC AMERICAN

Files Binders

For issues dated through 1982 1983 or later.

I enclose my check or money order for \$ _____ ;
(U.S. funds only)

Charge my credit card \$ _____ (Minimum \$15)
 Amex VISA MasterCard Diners

Card # _____ Exp. date _____

Signature _____

or call toll-free 7 days, 24 hours, 1-800-825-6690

Name _____ *(please print)*

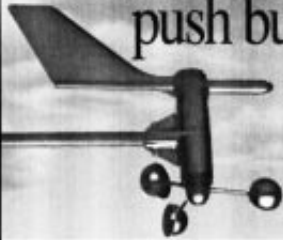
Address _____ *(No P.O. Box please)*

City _____

State _____ Zip _____

SATISFACTION GUARANTEED. Pennsylvania residents add 7% sales tax.

In case of weather,
push button.



Now you can have complete weather information literally at your fingertips with the Weather Monitor II. Sophisticated in design, yet easy to operate, it's as miraculous as the weather itself.

FEATURES INCLUDE:

- Inside & Outside Temps
- Wind Speed & Direction
- Inside Humidity
- Time & Date
- Barometer
- Wind Chill
- Alarms
- Highs & Lows
- Instant Metric Conversions
- Outside Humidity & Dew Point Option
- Rainfall Option
- Optional PC Interface



WEATHER MONITOR II

THE PROFESSIONAL HOME WEATHER STATION

Order today: 1-800-678-3669 • SC629X

M - F 7 a.m. to 5:30 p.m. Pacific Time • FAX 1-510-670-0589
MC and VISA • One-year warranty • 30-day money-back guarantee

DAVIS INSTRUMENTS 3465 Diablo Ave., Hayward, CA 94545

**CHOICE
MAGAZINE
LISTENING**

This **FREE** service—for anyone deprived of the joy of reading by loss of vision or other handicap—provides 8 hours of audio tapes every other month with unabridged selections from over 100 publications such as **THE NEW YORKER, SMITHSONIAN, ATLANTIC, and SCIENTIFIC AMERICAN**. CML subscribers have been reading the world's best minds for over a quarter century in selections by writers such as Saul Bellow, John McPhee, Annie Dillard, Grace Paley, William Styron, Seamus Heaney and Russell Baker.

The special 4-track cassette player is provided free, on permanent loan, by the Library of Congress.

For information, write:

CML, DEPT. 15
85 Channel Drive
Port Washington, NY 11050,
or call: (516)883-8280



MAGAZINE LISTENING
A TALKING MAGAZINE



PRECIPITATION PINNACLES are limestone pillars of cemented sand grains, exposed as loose sand blows away. These examples formed in Western Australia.

cination here is that nonetheless the theme is unity. All landforms are shaped by motion, large or small, slow or fast, out of preexisting structure during the stages of geologic time. Beginnings may be subtle. On icy Spitsbergen the rocks at your feet are splintered like firewood by ice wedging open the surficial fractures. In the Western Australia desert we see rounded rocks where first they were split by salt crystals, not ice, and then spalled off centimeter-thick shells to release the thermal stress stored over many tropical days and nights. Above Rio de Janeiro towers the celebrated dome of Sugar Loaf, shaped by exfoliation of granite shells tens of meters thick, shed long ago to release huge internal stresses, once the confining pressure was reduced by slow erosion.

Along the Upper Rhine and in a still more spacious view across the island of Hispaniola, large-scale linear stress- es are grandly revealed: deep-lying tensions opened long, parallel fault cracks, and the slabs between them slid downward, so that now elongated valleys lie flat for tens of miles within high, steep walls. The flow of fluids, all the mineral fluids—even to “fluidized” rock fragments and soil—has worked even more widely. Molten lava flowed once in a long tunnel under an already solidified lava surface in the Canaries; today the old channel, its roof collapsed to rubble, snakes across a quiet landscape. Climate opens to that abundant mineral, water, an option: crystal or liquid. Rivers cut their valleys into V's, as along the Saar; the glacier came down the Alps to plough out a wide U in the

Upper Engadin; and the sea itself in glacial Norway flooded a deep, narrow U-shaped valley, a neat village now at the head.

Tool and workpiece together determine the form. In arid Mali a perpendicular limestone step rises high above the softer marls; erosion and weathering rarely scar the riser of this near-perfect step. But in softly rainy Ireland, the Benbulbin Range shows almost as sharp a tread edge; the caprock limestone is much harder than what underlies it. Rocks are soluble in water, especially when the solvent is acidified. The process can eat away a wide, limey pavement, to leave the haystacks as it has in Guilin, China, or the higher cones of tropical Puerto Rico.

But solution can give as well as take; in the Western Australia desert a complex process has left behind limestone pinnacles up to two meters high, precipitated underground from dissolved shell fragments and eventually exposed by the winds. “Pseudokarsts” are here to see, not of calcite but of much more resistant sandstone, even of granite. Molecular teeth may grind even where the solvent carries no reactive ions; certain old ice surfaces mimic limestone corrosion pits, although the reactant that ate them was no ionic impurity but only heat.

The pleasures of these world-long gallery walls are marred only by the high cost of the monograph. Indeed, we readers are all a little spoiled by today's glossy magazines, whose wide circulation has made the luxury of abundant printed color seem all but commonplace. —Philip Morrison



Why Do Things Become More Complex?

Fifty years ago our technologies, our organizations and our lives were less complicated than today. Things were simpler. Most of us prize this plainness, this simplicity. Yet we are fascinated by complexity. Lately I've been wondering why the simple becomes complex. Is there a general principle causing things to get more complicated as time passes? Is complexity useful?

One good place to look for answers to these questions is the history of technology. The original turbojet engine, designed by Frank Whittle in the early 1930s, was beautifully simple. The idea was to propel aircraft by a jet of high-speed air. To do this, the engine took in air, pumped up its pressure by a compressor and ignited fuel in it. It passed the exploding mixture through a turbine to drive the compressor, releasing it through an exhaust nozzle at high speed to provide thrust. The original prototype worked well with just one moving part, the compressor-turbine combination.

Yet over the years, jet engines steadily become more complicated. Why? Commercial and military interests exert constant pressure to overcome limits imposed by extreme stresses and temperatures and to handle exceptional situations. Sometimes these improvements are achieved by using better materials, more often by adding a subsystem. And so, over time, jet designers achieve higher air pressures by using not one but an assembly of many compressors. They increase efficiency by a guide-vane control system that admits more air at higher altitudes and velocities and prevents engine stalling. They increase combustion temperatures, then cool the white-hot turbine blades by a system that circulates air inside them. They add bleed-valve systems, afterburner assemblies, fire-detection systems, fuel-control systems, deicing assemblies.

But all these additions require subsystems to monitor and control them and to enhance their performance when they run into limitations. These subsystems in turn require subsystems to enhance their performance. All this indeed improves performance—today's jet engine is 30 to 50 times more powerful than Whittle's. But it ends up encrusting the original simple system with subsystem upon subsystem and subassem-

bly upon subassembly in a vastly complicated array of interconnected modules and parts. Modern engines have upwards of 22,000 parts.

There's nothing wrong with this increase in complexity. We can admire it. On the outside, jet engines are sleek and lean; on the inside, complex and sophisticated. In nature, higher organisms are this way, too. On the outside, a cheetah is powerful and fast; on the inside, even more complicated than a jet engine. A cheetah, too, has temperature-regulating systems, sensing systems, control functions, maintenance functions—all embodied in a complex assembly of organs, cells and organelles, modulated not by machinery and electronics but by interconnected networks of chemical and neurological pathways. The steady pressure of competition causes evolution to "discover" new functions occasionally that push out performance limits. There's something wonderful about this—about how, over eons, a cheetah forms from its simple multicellular ancestors.

But sometimes the results of growing complexity are not so streamlined. For example, 60 years ago in most universities, bringing in and managing research grants might have occupied only a few people. These functions now require a development department, legal department, sponsored-projects office, dean-of-research office, grants accounting department, budget-control office, naval research office, technology licensing office. In part, such growth is necessary because the research-grant world itself is more complicated (and so complexity engenders further complexity). But often, new bureaucratic offices and departments become entrenched because the career interests they create overpower any external competitive forces that might pare them away. In 1896 my own university, Stanford, had only 12 administrators. It is still leaner than most, yet now it has more administrators than the British had running India in the 1830s.

It's that way with our lives, too. As we become better off, we gain more ways to squeeze more performance from our limited time. We acquire a car, profession, house, computers, fitness programs, pets, a pool, a second car. Fine.

But all these bring with them maintenance, repairs, appointments, obligations—a thousand subactivities to keep them going. In this case again, the overall result is increased complexity of debatable effectiveness.

So in answer to the original question, I believe there is a general law: complexity tends to increase as functions and modifications are added to a system to break through limitations, handle exceptional circumstances or adapt to a world itself more complex. This applies, if you think about it, not just to technologies and biological organisms but also to legal systems, tax codes, scientific theories, even successive releases of software programs. Where forces exist to weed out useless functions, increasing complexity delivers a smooth, efficient machine. Where they do not, it merely encumbers.

But, interestingly, even when a system gets lumbered down with complications, there is hope. Sooner or later a new simplifying conception is discovered that cuts at the root idea behind the old system and replaces it. Copernicus's dazzlingly simple astronomical system, based on a heliocentric universe, replaced the hopelessly complicated Ptolemaic system. Whittle's jet engine, ironically, replaced the incurably complicated piston aeroengine of the 1930s before it also became complex. And so growing complexity is often followed by renewed simplicity in a slow back-and-forth dance, with complication usually gaining a net edge over time.

The writer Peter Matthiessen once said, "The secret of well-being is simplicity." True. Yet the secret of evolution is the continual emergence of complexity. Simplicity brings a spareness, a grit; it cuts the fat. Yet complexity makes organisms like us possible in the first place. Complexity is indeed a marvel when it evolves naturally and delivers powerful performance. But when we seek it as an end or allow it to go unchecked, it merely hampers. It is then that we need to discover the new modes, the bold strokes, that bring fresh simplicity to our organizations, our technology, our government, our lives.

W. BRIAN ARTHUR is Morrison Professor of Economics at Stanford and external professor at the Santa Fe Institute, where he investigates the economy as an evolving, complex system.