SCIENTIFIC AMERICAN

SYNTHETIC VACCINES

$2.50

February 1983

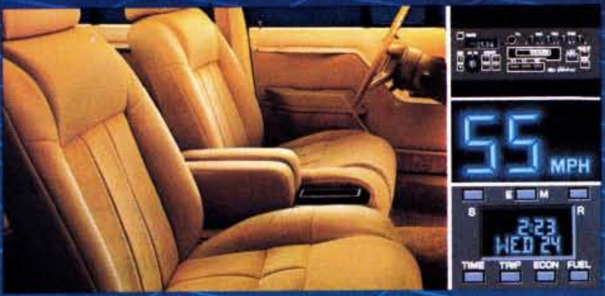
© 1983 SCIENTIFIC AMERICAN, INC

© 1983 SCIENTIFIC AMERICAN, INC

1983 FORD LTD

new Ford LTD. Totally Redesigned. inch an LTD.

For comparison. Your mileage may differ depending upon speed, distance and weather. Actual highway mileage and California ratings lower.



The LTD Brougham standard interior with reclining front seats and fold-down armrests may be enhanced with optional electronic instrumentation.

From the smooth ride to the new driver-designed seating position, the quiet new LTD has one target: your comfort. The steering wheel falls to hand naturally. The instruments are placed logically. The feel of the seats. Everything in the 1983 LTD has been human engineered to give you a uniquely comfortable driving experience.

Reshaped.

The slope of the hood. The cant of the rear window. The aerodynamically raised trunk lid. Smoothness everywhere gives LTD an amazingly low drag coefficient. All of this means excellent highway mileage, reduced wind noise and improved high-speed handling. 6.7 horsepower can push LTD through the air at 50 mph.

Refined.

The interior offers luxurious form and function. Soft cloth covers the standard reclining seats. Electronic options include: digital instrumentation • diagnostic warning lights • new electronic station searching radio • Tripminder® computer.

And behind you rides a streamlined version of our famous deep-well trunk.

Totally Redesigned.

The new 1983 Ford LTD represents more state-of-the-art automotive technology than ever before: smooth-riding gas-filled shocks, an optional 3.8 liter V-6 with automatic overdrive, a propane engine (available for order) and European-style wraparound tread radial tires.

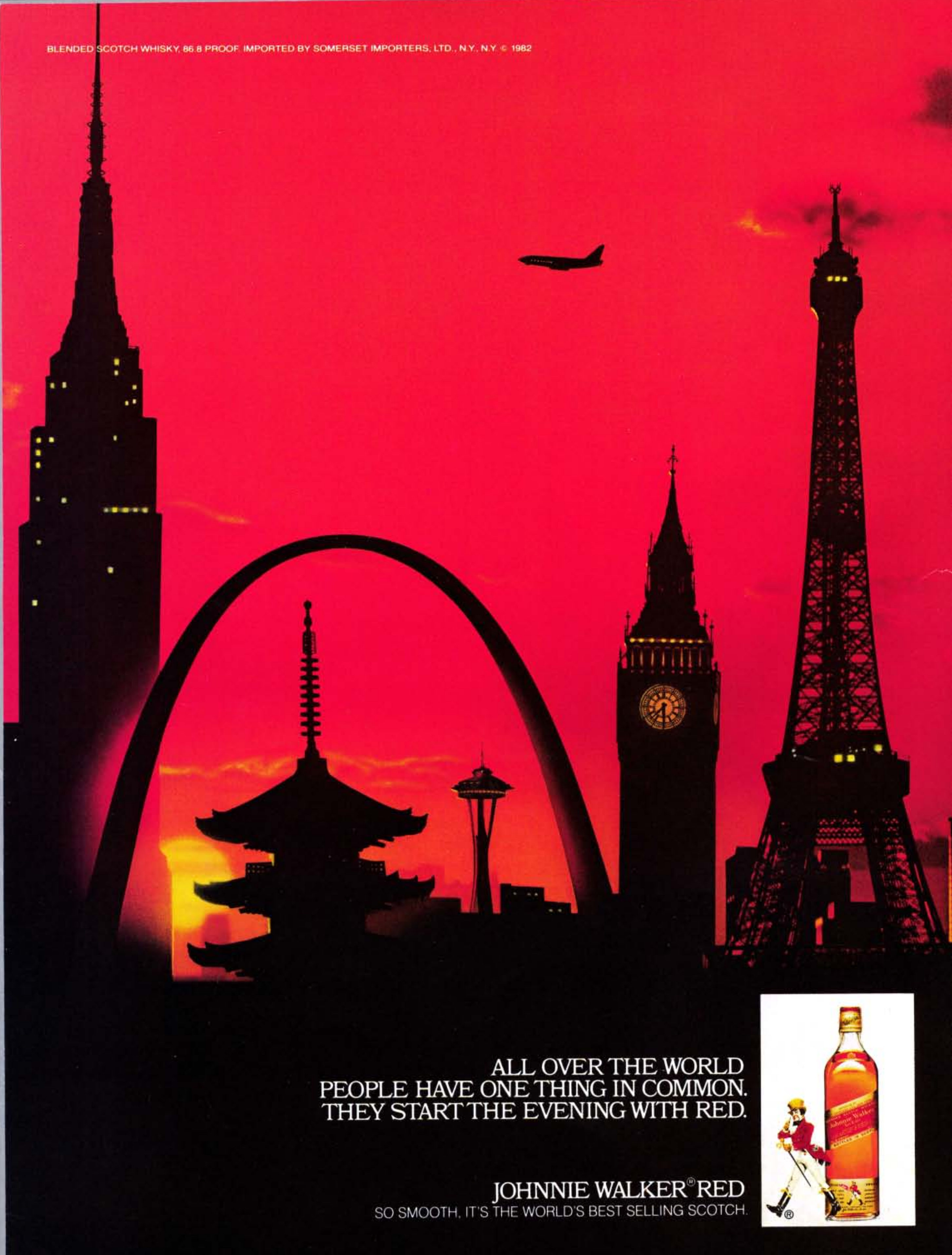
And all of this advanced technology adds up to a totally pleasurable driving experience. Because '83 LTD is, after all, every inch an LTD.

HAVE YOU DRIVEN A FORD...LATELY?

Ford

FORD DIVISION

© 1983 SCIENTIFIC AMERICAN, INC



© 1983 SCIENTIFIC AMERICAN, INC

Established 1845 SCIENTIFIC AMERICAN February 1983 Volume 248 Number 2

ARTICLES

DEPARTMENTS

BOARD OF EDITORS Gerard Piel (Publisher), Dennis Flanagan (Editor), Brian P. Hayes (Associate Editor), Philip Morrison (Book Editor), Francis Bello, John M. Benditt, Peter G. Brown, Michael Feirtag, Diana Lutz, Jonathan B. Piel, John Purcell, James T. Rogers, Armand Schwab, Jr., Joseph Wisnovsky
ART DEPARTMENT Samuel L. Howard (Art Director), Steven R. Black (Assistant Art Director), Ilil Arbel, Edward Bell
PRODUCTION DEPARTMENT Richard Sasso (Production Manager), Carol Hansen and Leo J. Petruzzi (Assistants to the Production Manager), Carol Eisler (Senior Production Associate), Karen O'Connor (Assistant Production Manager), Carol Albert, Lori Mogol, Martin O. K. Paul, Julio E. Xavier
COPY DEPARTMENT Sally Porter Jenks (Copy Chief), Debra Q. Bennett, Mary Knight, Dorothy R. Patterson
GENERAL MANAGER George S. Conn
ADVERTISING DIRECTOR C. John Kirby
CIRCULATION MANAGER William H. Yokel
SECRETARY Arlene Wright

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

© 1983 SCIENTIFIC AMERICAN, INC

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Student sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/83. $2.00 + 0.00.

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

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

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, N.Y. 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address

Street ______________________

City ______________________

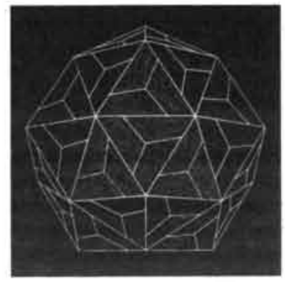
State and ZIP ______________________

Old Address

Street ______________________

City ______________________

State and ZIP ______________________



THE COVER

The illustration on the cover shows (in blue) the polyhedral structure of the capsid, or protein shell, of the tomato bushy-stunt virus. The capsid is composed of 180 copies of a protein, with three copies occupying each of the 60 triangular faces of the shell; 12 copies of the surface domain of the protein are depicted in orange. The illustration is based on a computer display generated by Arthur J. Olson to demonstrate a step in the design of a synthetic vaccine (see "Synthetic Vaccines," by Richard A. Lerner, page 66). Whereas a vaccine directed against viral infection ordinarily contains intact viruses, a synthetic vaccine contains only a peptide (a short segment of a protein chain) synthesized in the laboratory to mimic part of a viral surface protein. One way to ensure that the peptide mimics a site exposed on the surface, and therefore is likely to elicit antibodies that recognize the intact virus, is to exploit computer-graphics methods developed by T. J. O'Donnell of Abbott Laboratories and Michael L. Connolly and Olson of the Research Institute of Scripps Clinic.

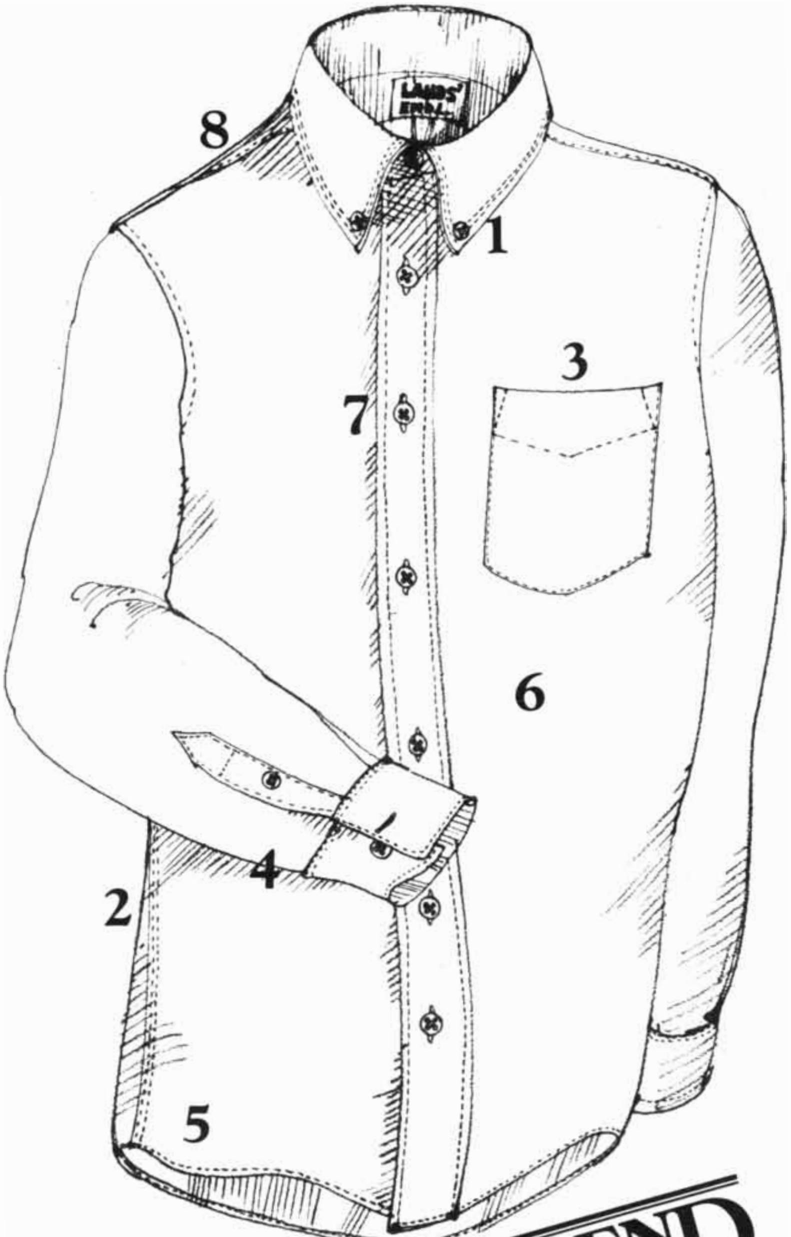
THE ILLUSTRATIONS

Cover illustration by Gabor Kiss

Page	Source
46–53	Andrew Tomko
55–65	Ilil Arbel
67	Arthur J. Olson, Research Institute of Scripps Clinic
68–69	Patricia J. Wynne
70–73	Arthur J. Olson, Research Institute of Scripps Clinic
74	Patricia J. Wynne
86–93	Walken Graphics
95	R. F. Bonifield, courtesy of the Granger Collection
96–102	George V. Kelvin
105	High Altitude Observatory
106	Jay M. Pasachoff, Williams College (*top*); High Altitude Observatory (*bottom*)
108	High Altitude Observatory
110	American Science & Engineering, Inc.
111	Goddard Space Flight Center (*top*), High Altitude Observatory (*bottom*)
112	Naval Research Laboratory
114–119	Allen Beechel
121	David M. Dennis
124	Patricia J. Wynne
125	Charles W. Myers, American Museum of Natural History
127	Lois W. Tice, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases
128	Patricia J. Wynne
132	Enid Kotschnig
134	John Moss, Black Star
136–139	Alan D. Iselin
142	Walter C. Leedy, Jr., Cleveland State University (*top*); Alan D. Iselin (*bottom*)
143	Alan D. Iselin
144	John Moss, Black Star
146–152	Michael Goodman

© 1983 SCIENTIFIC AMERICAN, INC

We set out to make the world's best buttondown. This one comes close. At $24.



LANDS' END
DIRECT MERCHANTS
of fine wool and cotton sweaters, Oxford buttondown shirts, traditional dress clothing, snow wear, deck wear, original Lands' End soft luggage and a multitude of other quality goods from around the world.

This is the Hyde Park—the latest addition to our impressive Oxford Collection, featuring both shirts of 100% cotton and our Lands' End reverse cotton blends.

Check it out feature for feature, beginning with the knowledge that it's made of imported 100% cotton Oxford. Heavier. More densely woven for a nicer drape. It launders better, resists wrinkles; best of all, it wears longer than normal.

For those of you interested in more specifics, we've provided this step-by-step "tour" of the shirt—available in pink, ecru, blue, maize, helio and white in our catalog.

1. **Double-track stitching highlights the softly-rolled collar, keeps it neater with a tie.**
2. **The shirt is single-needle stitched, with strong double-needle side seams.**
3. **Note the placement and detailing of the left breast pocket.**
4. **Barrel cuffs have a long, buttoned placket.**
5. **The shirt has extra long tails, so it stays neatly tucked in.**
6. **The cut is full, for comfort. Yet civilized, too. Tapered just enough to avoid being sloppy.**
7. **Note that it has a generous 7-button front.**
8. **In back, a box pleat with hanging loop adds function and style.**

Why make so much of a single shirt? We may have told you more than you ever wanted to know about a shirt. But only to make a point about the Lands' End philosophy of doing business. It is a simple philosophy really:

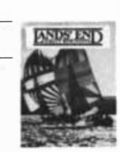
First, *quality*. Then, *price*. And always, always *service*.

A quality item at a reasonable price represents a Lands' End value. Anything less is someone else's ballgame—not ours. What's more, every item we offer you—from soft luggage to sweaters to snow wear to shoes—is unconditionally guaranteed.

We don't ask you to trust us, just try us. Mail the coupon for a free catalog. Better still call our toll-free 800 number, 800-356-4444. 24 hours a day. Except in Wisconsin, 608-935-2788.

☐ **Please send free catalog.**
Lands' End Dept. Q-08 Dodgeville, WI 53533

Name ______
Address ______
City ______
State ______ Zip ______



Or call Toll-free
800-356-4444
(Except Wisconsin call 608-935-2788)

NordicTrack

Jarless Total Body Cardiovascular Exerciser Duplicates X-C Skiing for the Best Motion in Fitness

The enjoyable sport of cross-country skiing is often cited by physiologists as the most perfect form of cardiovascular exercise for both men and women Its smooth, fluid, total body motion uniformly exercises more muscles and higher heart rates seem easier to attain than when jogging or cycling. NordicTrack closely simulates the pleasant X-C skiing motion and provides the same cardiovascular endurance-building benefits—right in the convenience of your home Makes a year round, consistent exercise program easily attainable Eliminates the usual barriers of time, weather, chance of injury, etc Also highly effective for weight control

Better Than Running
NordicTrack gives you a more complete work out—conditions both upper body and lower body muscles at the same time. Fluid, jarless motion will not cause joint or back pains as jogging or running often does.

More Effective Than Exercise Bikes
NordicTrack's stand-up skiing motion more uniformly exercises large leg muscles and also adds important upper body exercise. Higher pulse rates, necessary for building fitness, seem easier to attain because the work is shared by more muscle mass. The NordicTrack also keeps more muscles in tone.

Even Better Than Swimming
NordicTrack more effectively exercises the largest muscles in the body, those located in the legs and buttocks. These muscles' main function is to lift and propel the body in the standing position. However, when swimming, the body is supported by the water, thus preventing these muscles from being effectively exercised. The stand up exercising position on the NordicTrack much more effectively exercises these muscles.

A Proven, High Quality Durable Product
Ruggedly constructed, NordicTrack is quiet, motorless and folds compactly for storage. Separately adjustable arm and leg resistances. NordicTrack is in its 6th year of production and is used in thousands of homes and many institutions. We manufacture and sell direct. One year warrantee, 15 day trial period with return privilege.

Call or Write for FREE BROCHURE
TOLL FREE 1-800-328-5888
8 AM-5 PM MON.-FRI.
PSI 124 F Columbia ct. Chaska, MN 55318
Minnesota 612-448-6987

LETTERS

Sirs:

I must congratulate Randall Forsberg on her objective appraisal of the state of the nuclear arms race and its concomitant dangers ["A Bilateral Nuclear-Weapon Freeze," by Randall Forsberg; SCIENTIFIC AMERICAN, November, 1982]. For me the article has provided the most convincing case for a nuclear freeze yet seen in print.... To Europeans, who fear the outbreak of nuclear war on their soil as keenly as, if not more than, citizens of the U.S., and particularly in view of the weapon escalation planned by NATO and the British government for next year, this article has been most timely.

Whilst a bilateral nuclear freeze would be strongly welcomed, Ms. Forsberg did not mention a complementary move that could be tied to a freeze and that could considerably improve the security of all nations: the establishment of a European battlefield-nuclear-weapon-free zone (BNWFZ) on both sides of the Iron Curtain, as has been recommended by the Independent Commission on Disarmament and Security Issues under the chairmanship of Olof Palme (*Common Security: A Blueprint for Survival,* Simon and Schuster, 1982). The existing dangers of current battlefield nuclear weapons have been clearly stated in the Palme report: "Because of their deployment in forward areas battlefield nuclear weapons run the risk of being overrun early in an armed conflict. Maintaining command and control over such weapons in 'the fog of war' would be difficult. Pressures for delegation of authority to use nuclear weapons to local commanders and for their early use would be strong. The danger of crossing the nuclear threshold and of further escalation could become acute."

GRAEME G. WILKINSON

University of Reading
Reading, England

Sirs:

I was interested in the "default assumption" story about the son whose mother was a surgeon ["Metamagical Themas," by Douglas R. Hofstadter; SCIENTIFIC AMERICAN, November, 1982]. When I translated the story into Latvian, my native language, the default assumption failed. The rules of grammar revealed the surgeon to be female. I did construct some default-assumption stories in the language and tried them on my countrypersons. My reward was reprimands for improper grammar, distorted syntax, wordplay and tricksterism.

There are compensations. Consider the plight of two Latvian couples, Freda and Fredis Ozols and Vila and Vilis Liepa. Fredis was pleased with his surname Ozols, the oak tree, a symbol of manly strength and fortitude. Freda, like many other wives and daughters, was not. There was a national debate on the question, and eventually women were allowed to feminize their surnames. Thus Freda Ozols became Freda Ozola. Fredis did not object; Ozola is also the genitive case of Ozols, making Freda Fredis' property.

Meanwhile Vilis Liepa led a compromising life with a feminine-gender surname, the linden tree, a symbol of womanly propriety and virtue. Vila was pleased but secretly envied Freda, who could differentiate her surname from her husband's. Vila took the genitive strategy and became Vila Liepas, which also masculinizes the surname. It is odd, grumbled Vilis; if I were a woman, I could have a proper man's name.

KARLIS REKEVICS

Seattle, Wash.

Sirs:

The mention of Mersenne numbers in "The Search for Prime Numbers," by Carl Pomerance [SCIENTIFIC AMERICAN, December, 1982], brings to mind an anecdote told by E. T. Bell in *Mathematics: Queen and Servant of the Sciences.* Marin Mersenne had claimed that $2^{67} - 1$ was a prime number, but in 1903 F. N. Cole proved it was composite. To quote Bell:

"When I asked Cole in 1911 how long it had taken him to crack M_{67}, he said 'Three years of Sundays.' But this, though interesting, is not the history. At the October 1903 meeting in New York of the American Mathematical Society, Cole had a paper on the program with the modest title 'On the factorization of large numbers.' When the chairman called on him for his paper, Cole—who was always a man of very few words—walked to the board and, saying nothing, proceeded to chalk up the arithmetic for raising 2 to the sixty-seventh power. Then he carefully subtracted 1. Without a word he moved over to a clear space on the board and multiplied out, by longhand, 193,707,721 × 761,838,257,287. The two calculations agreed. Mersenne's conjecture—if such it was—vanished into the limbo of mathematical mythology. For the first and only time on record, an audience of the American Mathematical Society vigorously applauded the author of a paper delivered before it. Cole took his seat without having uttered a word. Nobody asked him a question."

ROBERT C. GEBHARDT

Hopatcong, N.J.

© 1983 SCIENTIFIC AMERICAN, INC

HOME IS THE LAST PLACE YOU SHOULD LEARN ABOUT A HOME COMPUTER.

WANT TO LEARN SOMETHING ABOUT HOME COMPUTERS? HERE, IT'S FREE. AT HOME, IT COULD COST YOU.

No one expects you to know everything about a home computer before you buy it. A fact which is not lost on our competition.

They know that an impressively low price can divert your attention from some depressingly cheap features. So that you won't know what you may be missing with their home computer until after it's been in your home for a while.

At which point, naturally, it'll cost you to change your mind.

IT'S EASY TO TELL THE DIFFERENCE.

Fortunately, you don't have to be a computer engineer to tell what makes the Commodore VIC 20™ superior to the competition.

All you have to do is take advantage of three of your five senses.

Use your sense of vision and read this comparison chart. You can see in black and white where two of our major competitors have skimped. Use your sense of touch in the store. You'll feel the VIC 20's superiority immediately. It feels a lot more expensive than it is.

If these two senses don't convince you that the VIC 20 offers more for the money than any other home computer, simply rely on common sense.

NOW THAT YOU KNOW HOW EASY A COMMODORE HOME COMPUTER IS TO OWN, FIND OUT HOW EASY IT IS TO EXPAND.

One thing about home computers that you're bound to discover at home is that, once you learn what they can do, you'll want them to do more and more. To do this, you may need accessories called peripherals. These let you get more out of a home computer by letting you put more into it.

They include items like cassette recorders and disk drives to input data, modems for telecomputing and printers. And all VIC 20 peripherals are fully compatible with the powerful Commodore 64™ personal computer.

PLAN AHEAD.

When you start looking at your first home computer, you may think it's too early to start planning to add peripherals. If that's what you think, you're once again playing right into the hands of our competitors.

Because once they've gotten you to buy their home computer, for what seems to be a reasonable price, they have you hooked on their system.

The costs of which, if you'll examine the chart below, can really start getting unreasonable. For example, while these computers may seem to be close to the same price to start, an expanded system can cost you twice as much with TI or Atari as with the Commodore VIC 20 or Commodore 64.

EXPANSION COSTS	VIC 20™ or COMMODORE 64™	TI99/4A®	ATARI 400®
BASIC	Included	Included	$59.95
Peripheral Expansion System	Not Necessary	$249.95	Not Necessary
Disk Drive	$399.00	399.95	599.95
Disk Controller Card	Included	249.95	Included
Modem	109.95	224.95	199.95
Modem Interface	Included	174.95	219.95
TOTAL	$508.95	$1299.75	$1079.80

Manufacturer's suggested list prices: Prices per TI June-December 1982 U.S. Consumer Products Suggested Price List. Atari prices effective July 1, 1982 Suggested Retail Price List

THINK OF IT AS BUYING A TOASTER.

It's easy to fill up a computer ad with RAM's and ROM's, numbers and technical jargon. But when it comes right down to it, buying a home computer is just like buying anything else. It's important to know just what you're getting for your hard-earned money.

And we hope we've accomplished that here by telling you about the cost of expanding your Commodore VIC 20 or Commodore 64 computer.

COMPUTER FEATURES	VIC 20	TI 99/4A	ATARI 400
Typewriter Keys	Yes	Yes	No
Typewriter Feel	Yes	No	No
Color Control Keys	Yes	No	No
Graphics on Keys	Yes	No	No
Reverse Letters	Yes	No	Yes
Programmable Function Keys	Yes	No	No
Works with TV or Monitor	Yes	Yes	No
True Lower Case Letters	Yes	No	Yes
DISK FEATURES			
Capacity	170K	90K	88K

commodore COMPUTER

U.S.A.—P.O. Box 500, Conshohocken, PA 19428; Canada—3370 Pharmacy Ave., Agincourt, Ontario, Canada M1W2K4.

© 1983 SCIENTIFIC AMERICAN, INC

Quattro Turbo

4000 Sports Sedan or Diesel
Gas model shown here.

© 1983 SCIENTIFIC AMERICAN, INC

Audi offers more high-performance German luxury cars than anyone. Surprised?

AUDI In a sea of newly minted "performance" cars, automakers are suddenly extolling the pleasures of "varroom."

Performance is, once more, "in."

At Audi, it was never "out."

After decades of refining performance automobiles, our engineers in Germany find all this amusing.

To those discerning car buyers, the most sophisticated high-performance luxury cars continue to come from Germany.

The most complete line of which comes from Audi, for 1983. The recent addition of the 5000 Turbo Diesel, the Quattro Turbo and GT Coupe illustrates Audi's commitment to performance.

Without hyperbole, we can say that no luxury Sports Coupe can compete with our Quattro Turbo.

Ask those rally drivers who have tried to, in America and in Europe. The Quattro Turbo's powerhouse tandem of all-wheel drive and turbocharged five-cylinder engine has led to eight straight victories in SCCA competition in the United States—following one stunning victory after another in Europe.

Turbocharged engine power is also readily available in our 5000 flagship line. The 5000 Turbo Gas and the 5000 Turbo Diesel. Two full-sized, high-performance German luxury cars with turbo exhilaration. A rare blending of elegance, power and remarkably quick reflexive handling.

In a mere 7.5 seconds, the 3000-pound 5000 Turbo can bring you comfortably to 50 miles an hour. Even the Turbo Diesel is a top performer with its 2.0-liter, five-cylinder power plant.

When we introduced our GT Coupe, it was named, without question, by the editors of *Road & Track* "Best Sports Coupe For The 80's."

We weren't surprised. It is one of the few genuine Grand Touring Coupes. And the only front-wheel drive, five-cylinder Coupe in the world. Aerodynamically designed, it thrives on tight, hairpin curves.

Some of the most enjoyable performance driving you'll ever do will be in our 4000 Sports Sedan. It is extremely agile, the steering very light. Front-wheel drive. Four-cylinder, 1.7-liter engine.

Available as a two-door, four-door, or '82 diesel model, the 4000s offer many exciting features as standard equipment and have a low starting base price of only $10,205.*

No matter which Audi you buy, you get the high level of luxury you expect from motor cars in this class. Because, here in Ingolstadt, Germany, we engineer thoughtful luxuries into every Audi. For example, a few of the luxurious appointments available in certain Audi models include: air conditioning, power windows, central door/trunk locking system, sumptuous seats (genuine leather available) and AM/FM stereo/cassette radio.

Simply stated, in a luxury Audi, performance means something today. Because it always has.

For your nearest Porsche Audi dealer or details on the Audi Delivery in Europe Program, call toll-free (800) 447-4700. In Illinois, (800) 322-4400.

© 1983 Porsche Audi

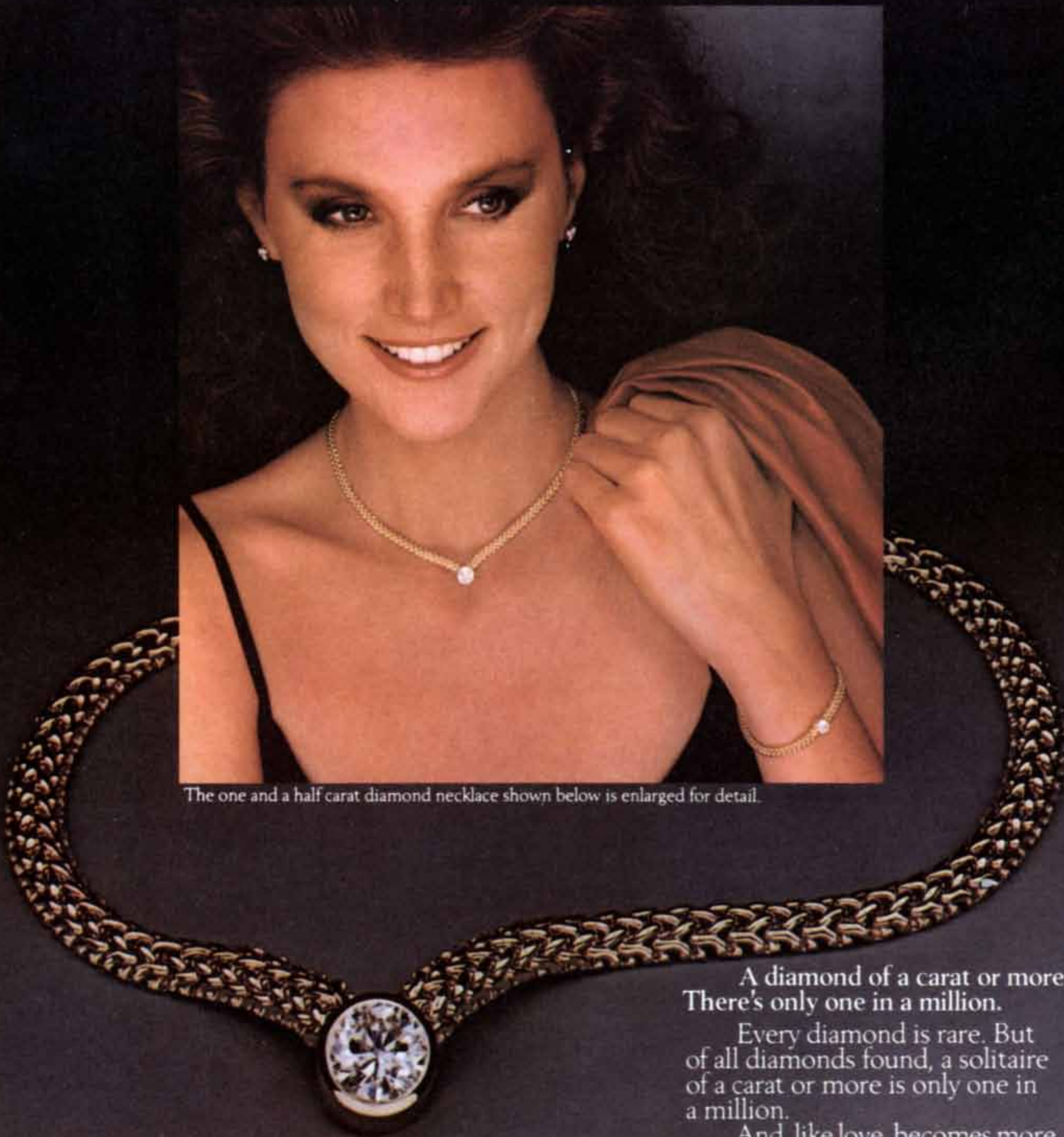
PORSCHE + AUDI

Audi: the art of engineering.

Use "est. mpg" for comparison. Mpg and range vary with speed, trip length, weather. Actual hwy. mpg and hwy. range will be less. *Use mfr's sugg. retail price. Title, taxes, transp., registration and dealer prep add'l.

© 1983 SCIENTIFIC AMERICAN, INC

A carat or more.
A rare diamond. For that rare individual who never doubted you'd make it.



The one and a half carat diamond necklace shown below is enlarged for detail.

**A diamond of a carat or more.
There's only one in a million.**

Every diamond is rare. But of all diamonds found, a solitaire of a carat or more is only one in a million.

And, like love, becomes more precious with time.

A miracle among miracles. Born from the earth. Reborn on a woman.

The extraordinary diamond of a carat or more.

Show the world you couldn't have made it without her.

A diamond is forever. De Beers.

© 1983 SCIENTIFIC AMERICAN, INC

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

FEBRUARY, 1933: "A million dollars a night! Dr. Irving Langmuir, who has just received the Nobel award for chemistry, made possible this saving to the American people in electricity costs by his development of the high-intensity incandescent lamp, which contains small quantities of either nitrogen or argon. He is only the second American chemist to have been honored by the Swedish Academy in 31 years. Growing out of his work with the incandescent lamp came momentous discoveries affecting the then new and immature vacuum tube. Dr. Langmuir's work with vacuum tubes contributed largely to the vast development of modern radio broadcasting. Equally important was his discovery of the atomic-hydrogen method of arc welding. He has been particularly interested in the mechanism of chemical reactions taking place on the surfaces of solids. He is now engaged in working out the laws according to which atoms and molecules distribute themselves over surfaces to form monatomic layers—laws of great importance in understanding many simple phenomena."

"Stars that appear reddish in color have been thought to be very far away, but now Dr. Joel Stebbins of the Mount Wilson Observatory reports to the National Academy of Sciences meeting at the University of Michigan that some red stars located in the Milky Way are four times closer to us than has been estimated. A layer of 'cosmic dust' that splits the Milky Way makes the stars on the far side appear red, just as our sun appears red at sunset, when its rays come to the observer through a longer distance of the dust-filled atmosphere of our planet."

"The rocket may someday become a very dangerous military weapon. In fact, many authorities think that its utilization in warfare is likely to come much sooner than its application in aerial navigation, and they visualize a powerful rocket guided by wireless and capable of destruction at long range. Perhaps one reason the Germans are so interested in rocket development is that they are debarred from the construction of military aircraft. At any rate, more rocket experimentation is carried on in Germany than in all other countries put together. The latest German rocket design is that of a young engineer named Tilling, of Osnabrück. Tilling recently gave a striking demonstration at the Tempelhof aerodrome in Berlin. The Tilling rocket is made entirely of aluminum and consists of a central shell to which are attached four long tail fins. The over-all length is about nine feet. The rocket motor is inserted into the shell between the fin roots; this motor is filled with a special powder that is relatively slow-burning. As Tilling wished his craft to alight within the aerodrome, he only employed a 12-pound charge in his experiment. The powder was ignited from a distance by means of an electric spark, and the reaction of the exhaust gases drove the rocket upward to a height of about 2,600 feet. Two wings were hidden in the hollow fins until the rocket had reached its maximum altitude. Then a simple automatically acting hydraulic device gradually moved the wings out from their sheaths. The rocket, now converted into a glider, made its descent in a fairly tight spiral. The landing was only 400 yards from the starting point."

SCIENTIFIC AMERICAN

FEBRUARY, 1883: "At a recent meeting of the Académie de Médecine, M. Bouley communicated, in the name of M. Pasteur, a series of conclusions regarding rabies at which the distinguished investigator has arrived. The first was to enunciate the familiar truths that the dumb madness and the furious madness, and in short all varieties of rabies, are caused by the same virus. Mention was also made of the cases of three dogs inoculated in 1881: two quickly died from rabies; the third, after manifesting the early systems, recovered. The latter animal was inoculated by trephining in 1882 on two separate occasions, but without effect. M. Pasteur asserts that he now has four dogs that will not contract rabies, whatever the method of inoculation adopted or the proved virulence of the material employed. These facts he believes to be the first step toward the discovery of a method of the prevention of rabies by its inoculation. He confesses, however, that the end seems to be at present far distant."

"It is simply for lack of muscular power that man cannot fly. There is no combination of wings or arrangements of any kind that will compensate for this fact. Whether man can produce a machine to supplement his own want of force remains to be seen. Such a motor cannot, we think, be driven by steam. It is not impossible that a machine might be made that would be caused to fly by means of a small electric motor run at a very high speed and worked by the aid of a couple of wires from the ground. This, however, would hardly be flying in the true sense of the word. That wings and such things can be made we have no doubt; and experiments enough have been done to prove that, if power enough be available, flight can be achieved. When a machine can be made each pound of which will develop as much energy as each pound of a bird, flying may be possible—not till then."

"The Hudson River ice crop of the season just closed is rated as one of the largest, if not the largest, ever harvested. It amounted to nearly 3,000,000 tons. It is said that not so many new ice houses were built last fall as the previous one; still, those constructed were quite large, and they increase the storing capacity 200,000 tons. The ice gathered ranges in thickness from eight to 20 inches, and, owing to the low condition of the river when it froze over, is as clear as crystal. Notwithstanding the cry of short crop last winter, about 100,000 tons remained in ice houses unsold at the beginning of the year. Nearly 50,000 tons of this was ice gathered in 1880, which had been held for a good market."

"It is gratifying to know that the Western Union Telegraph Company have decided to begin in earnest the work of putting all the wires of the company in New York underground. The first underground main will extend from the company's building at Broadway and Dey Street, up Broadway to 23rd Street. The wires, it is said, will be laid in a tube large enough to contain 200 wires. The right to use the streets for this purpose was secured a few years ago. The promise is that if the underground system adopted works well, it will be extended to all parts of the city."

"Recently newspapers in France announced that the police of Bordeaux had forbidden the exhibition of the automaton Az Rah, one of the attractions of the Exhibition Theater, because it had been discovered that the manikin was set in motion not by mechanical arrangements but by a youth of 18 years enclosed within a cavity behind the wheelwork, and whose health was gravely compromised by this daily torture. This automaton recalls the famous Turkish chess player that was constructed in Hungary by Baron Kempelen in 1769 and exhibited in Germany, Russia, France, England and America without the public's succeeding in ascertaining its mechanism. In 1819 and 1820 a man named Melzer showed it anew in England. Robert Houdin saw it in 1844 at the house of a mechanician of Belleville, named Cronior. Since then its fate has been unknown, and it is very probable the Az Rah of Bordeaux is nothing other than the Turk of Vienna."

© 1983 SCIENTIFIC AMERICAN, INC

THE AUTHORS

SANDRA S. BATIE and ROBERT G. HEALY ("The Future of American Agriculture") are economists with a special interest in natural resources and land-use planning. Batie is associate professor of agricultural economics at Virginia Polytechnic Institute and State University. A graduate of the University of Washington, she has an M.S. and a Ph.D. from Oregon State University. She has been a member of the faculty at Virginia Tech since 1973. Healy is associated with the Conservation Foundation in Washington, D.C. He got his B.A., M.A. and Ph.D. degrees, all in economics, from the University of California at Los Angeles. Before joining the staff of the foundation in 1975 he taught regional planning at Harvard University and U.C.L.A. and was a member of the research staff of the Urban Institute and Resources for the Future. The present article is based in part on an interdisciplinary study sponsored by the Federal Emergency Management Agency and the Conservation Foundation; the study resulted in the book *The Future of American Agriculture as a Strategic Resource,* published by the foundation. Contributors to the book included (besides Batie and Healy) Otto C. Doering III, Kenneth D. Frederick, Jack R. Harlan, Philip M. Raup, Vernon W. Ruttan, Robert H. Shaw and Frederick N. Swader.

CLAUDIO REBBI ("The Lattice Theory of Quark Confinement") is an elementary-particle theorist on the staff of the Brookhaven National Laboratory. Born in Trieste, he studied physics at the University of Torino, where he received the Italian equivalent of a Ph.D. in 1967. He did his postdoctoral work at the California Institute of Technology, taught at the University of Trieste and did research at the European Organization for Nuclear Research (CERN) in Geneva before returning to the U.S. in 1974 as a visiting associate professor at the Massachusetts Institute of Technology. He moved to Brookhaven in 1977. During the academic year 1980–81 he was again at CERN as a visiting scientist. This is Rebbi's second article in SCIENTIFIC AMERICAN; his previous one, "Solitons," appeared in February, 1979.

RICHARD A. LERNER ("Synthetic Vaccines") is chairman of the molecular biology department at the Research Institute of the Scripps Clinic in La Jolla, Calif. He obtained his M.D. from the Stanford University School of Medicine in 1964 and spent the next year at Stanford as a medical intern. He has been affiliated with the Scripps Clinic ever since, except for two years at the Wistar Institute of Anatomy and Biology in Philadelphia in 1968–70. Among his hobbies is writing fiction; his latest published novel is *Epidemic 9,* and he is currently at work on another novel.

EITAN ABRAHAM, COLIN T. SEATON and S. DESMOND SMITH ("The Optical Computer") work together at Heriot-Watt University in Edinburgh. Abraham is a research associate in the physics department, where he lectures on quantum statistics, thermodynamics and computational physics. A native of Israel, he was graduated from the University of Buenos Aires in 1975. He got his Ph.D. from Manchester University in 1979. Seaton received his undergraduate degree in applied physics from Heriot-Watt in 1979 and is currently completing work on his Ph.D. His research interests include nonlinear optical effects in semiconductors and optical switching and signal-processing devices. Smith is professor of physics at Heriot-Watt, where he also serves as dean of the faculty of science. He did his undergraduate work at the University of Bristol and went on to obtain two doctoral degrees: a Ph.D. from the University of Reading (1956) and a D.Sc. from the University of Bristol (1966). Before taking up his present posts at Heriot-Watt in 1970, he worked at the Royal Aircraft Establishment at Farnborough, Imperial College London and the University of Reading. In addition to his academic duties he is chairman of Edinburgh Instruments Ltd., a manufacturer of laser systems. Smith is a fellow of the Royal Society of Edinburgh, the Institute of Physics (which awarded him the C. V. Boys Prize in 1976) and the Royal Society of London. From 1977 to 1980 he was chairman of the quantum electronics division of the European Physical Society.

JEREMY M. WOLFE ("Hidden Visual Processes") teaches introductory psychology at the Massachusetts Institute of Technology and does research in the institute's Center for Cognitive Science. Born in London, he was graduated summa cum laude from Princeton University in 1977. He did his graduate work on binocular vision with Richard Held at M.I.T., getting his Ph.D. in 1981. He writes: "In high school I thought I would become a lawyer or perhaps a rabbi, but I was 'converted' to the experimental psychology of vision by a summer job at Bell Laboratories. I spent my time sitting in a dark room looking at barely visible flashes of colored light and came out determined to do the same to other people." When Wolfe is not working in the classroom or the laboratory, he adds, "I live in Cambridge, Mass., with one wife, two cats (one without a tail) and one six-foot Burmese python."

RICHARD WOLFSON ("The Active Solar Corona") is associate professor of physics at Middlebury College. He spent his undergraduate years at the Massachusetts Institute of Technology and Swarthmore College, receiving a B.A. in philosophy and physics at Swarthmore in 1969. After a brief stint doing research in medical physics at the Massachusetts General Hospital he took a master's degree in environmental studies at the University of Michigan. He subsequently went to Dartmouth College, where he obtained his Ph.D. in physics in 1976. His research interests, he reports, "center on the behavior of gases and plasmas in astrophysical situations and include studies of accretion flows in X-ray sources as well as solar and stellar winds and the solar corona. In 1980–81 I was a visiting scientist at the High Altitude Observatory of the National Center for Atmospheric Research, where I carried out theoretical studies of coronal dynamics. I should like to acknowledge support from the staff of the High Altitude Observatory and from many others who provided photographs."

CHARLES W. MYERS and JOHN W. DALY ("Dart-Poison Frogs") have worked together since 1966, "dividing our time between jungle camps and modern laboratories in a multidisciplinary approach to the biology and biochemistry of poisonous frogs." Myers is chairman of the department of herpetology at the American Museum of Natural History. He received a B.S. at the University of Florida in 1960 and an M.A. at Southern Illinois University in 1962. He interrupted his doctoral studies to spend three years (1964–67) as a visiting scientist at the Gorgas Memorial Laboratory in Panama. He joined the staff of the American Museum in 1968 and got his Ph.D. from the University of Kansas in 1970. Daly is chief of the laboratory of bioorganic chemistry at the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases. He obtained a B.S. in 1954 and an M.A. in 1955 at Oregon State College and a Ph.D. in chemistry from Stanford University in 1958.

WALTER C. LEEDY, JR. ("Fan Vaulting"), is an architectural historian and medievalist at Cleveland State University. He majored in architecture as an undergraduate and graduate student at the University of Michigan, taking time out to spend a year as a Kościuszko Foundation Scholar in Poland in 1966–67. He was awarded his Ph.D. in art history in 1972 by the Courtauld Institute of Art of the University of London.

© 1983 SCIENTIFIC AMERICAN, INC

explore the world below

Oceanus®

Published by
Woods Hole Oceanographic Institution
one of the world's leading centers of ocean research

Marine Scientists talking directly to the serious student of the sea

SUBSCRIPTION DEPT. BOX H
1172 COMMONWEALTH AVENUE
BOSTON, MASSACHUSETTS 02134
Please make checks payable to Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543.

Please enter my subscription to OCEANUS for one year at $15.00 (four issues).
☐ payment enclosed ☐ bill me

Name..Address ....

City...State.......................................Zip ....

© 1983 SCIENTIFIC AMERICAN, INC

METAMAGICAL THEMAS

The pleasures of Lisp: the chosen language of artificial intelligence

by Douglas R. Hofstadter

In previous columns I have written often about the field of artificial intelligence: the search for ways to program computers so that they might come to behave with such characteristics as flexibility, common sense, insight, creativity, self-awareness and humor. The quest for AI began in earnest more than two decades ago, and since then it has bifurcated many times, so that today it is an active and multifaceted research area. In the U.S. there are perhaps a couple of thousand people professionally involved in AI, and abroad there are a similar number. Although there is among these workers a considerable divergence of opinion concerning the best route to AI, one thing that is nearly unanimous is the choice of programming language. Most AI efforts are carried out in the language called Lisp. The name stands for "list processing."

Why is most AI work done in Lisp? There are many reasons, most of which are somewhat technical, but one of the best is quite simple: Lisp is crisp. Every computer language has arbitrary features, and most languages are in fact overloaded with them. A few, however, such as Lisp and Algol, are built around a kernel that seems as natural as a branch of mathematics. The kernel of Lisp has a crystalline purity that not only appeals to the aesthetic sense but also makes Lisp a language far more flexible than most others. Because of Lisp's beauty and centrality in this important area of modern science, I shall be devoting some columns to presenting some of the basic ideas of the language.

The deep roots of Lisp lie principally in mathematical logic. Mathematical pioneers such as Thoralf Skolem, Kurt Gödel and Alonzo Church contributed seminal ideas to logic in the 1920's and 1930's that were incorporated decades later into Lisp. Computer programming in earnest began in the 1940's, but "higher level" programming languages (of which Lisp is one) came into existence only in the 1950's. The earliest list-processing language was not Lisp but IPL ("information-processing language"), developed in the mid-1950's by Herbert A. Simon, Allen Newell and J. C. Shaw. In the years 1956–58 John McCarthy, drawing on all these earlier sources, came up with the elegant algebraic list-processing language he named Lisp. It caught on quickly with the young crowd around him at the newly formed M.I.T. Artificial Intelligence Project, was implemented on the IBM 704, spread to other AI groups and has stayed around all these years. Many dialects now exist, but all of them share that elegant kernel.

Let us now move on to the way Lisp really works. One of the most appealing features of Lisp is that it is "interactive," as contrasted with most other higher-level languages, which are noninteractive. What this means is the following. When you want to program in Lisp, you sit down at a terminal connected to a computer that has a Lisp system in its memory and you type the word "lisp" (or words to that effect). The next thing you will see on your screen is a "prompt": a characteristic symbol such as an arrow or asterisk. I like to think of the prompt as a greeting spoken by a special "Lisp genie," bowing low and saying to you, "Your wish is my command, and now, what is your next wish?" The genie then waits for you to type something to it. This genie is usually referred to as the Lisp *interpreter,* and it will do anything you want—but you have to take great care in expressing your desires precisely, or you may get some disastrous effects. Shown below is the prompt of Franz Lisp (a particular dialect), the sign that the Franz Lisp genie is ready to do our bidding:

```
—>
```

The genie is asking us for our heart's desire, so let us type in a simple expression,

```
—> (plus 2 2)
```

and then a carriage return. Even non-Lispers can probably anticipate that the Lisp genie will print in return the symbol "4". Then it will also print a fresh prompt, so that the screen will now show:

```
—> (plus 2 2)
4
—>
```

The genie is now ready to carry out our next command or, more politely, our next wish. The carrying out of a wish expressed as a Lisp statement is called the *evaluation* of that statement. The preceding short interchange between human being and computer exemplifies the behavior of the Lisp interpreter: it *reads* a statement, *evaluates* it, *prints* the appropriate value and then signals its readiness to read a new statement. For this reason the central activity of the Lisp interpreter is referred to as the "read-eval-print" loop.

The existence of the Lisp genie (the Lisp interpreter) is what makes Lisp interactive. You get immediate feedback as soon as you have typed a wish—a complete statement—to Lisp. And the way to get a bunch of wishes carried out is to type one, ask the genie to carry it out, then type another, ask the genie again, and so on.

In contrast, in many higher-level computer languages (but by no means all) you must write out an entire program consisting of a vast number of wishes to be carried out in some specified order. What is worse is that later wishes usually depend strongly on the consequences of earlier wishes—and of course you do not get to try them out one by one. The execution of such a program may, needless to say, lead to unexpected results, because so many wishes have to mesh perfectly. If you have made the slightest conceptual error in designing your wish list, then a total foul-up is likely; in fact it is almost inevitable. Running a program of this type is like launching a new space probe untested: you cannot possibly have anticipated all the things that might go wrong, and so all you can do is sit back and watch, hoping it will work. If it fails, you go back and correct the one thing the failure revealed and then try another launch. Such a gawky, indirect, expensive way of programming is in marked contrast to the direct, interactive, one-wish-at-a-time style of Lisp, which allows "incremental" program development and debugging. This is another major reason for the popularity of Lisp.

What kinds of wishes can you type to the Lisp genie for evaluation, and what kinds of things will it print back at you? Well, to begin with, you can type arithmetic expressions expressed in a rather strange way, such as "(times (plus 6 3) (difference 6 3))". The answer is 27, since "(plus 6 3)" evaluates to 9, "(differ-

© 1983 SCIENTIFIC AMERICAN, INC

CIMARRON RESPONDS

You can sense it the moment you first see Cimarron's bold, contemporary road stance. You can feel it the second you slip behind the wheel and into Cimarron's leather-faced lumbar-supported front bucket seats. And you can experience it every time you take to the open road.

Cimarron responds. See for yourself at your Cadillac dealer's soon. And be sure to ask about the special edition Cimarron D'ORO, shown above.

GM

BEST OF ALL...IT'S A CADILLAC. CADILLAC MOTOR CAR DIVISION U.S.A

© 1983 SCIENTIFIC AMERICAN, INC

Exxon introduces offshore oil

Joe Burkhardt helped design a production facility that works on the floor of the sea.

Some of the world's untapped petroleum may be trapped in rocks under water too deep for conventional platforms. That's why Joe Burkhardt and teams of Exxon scientists and engineers have pioneered the development of new technology to take production into deeper and deeper waters.

Breaking new ground beneath the sea

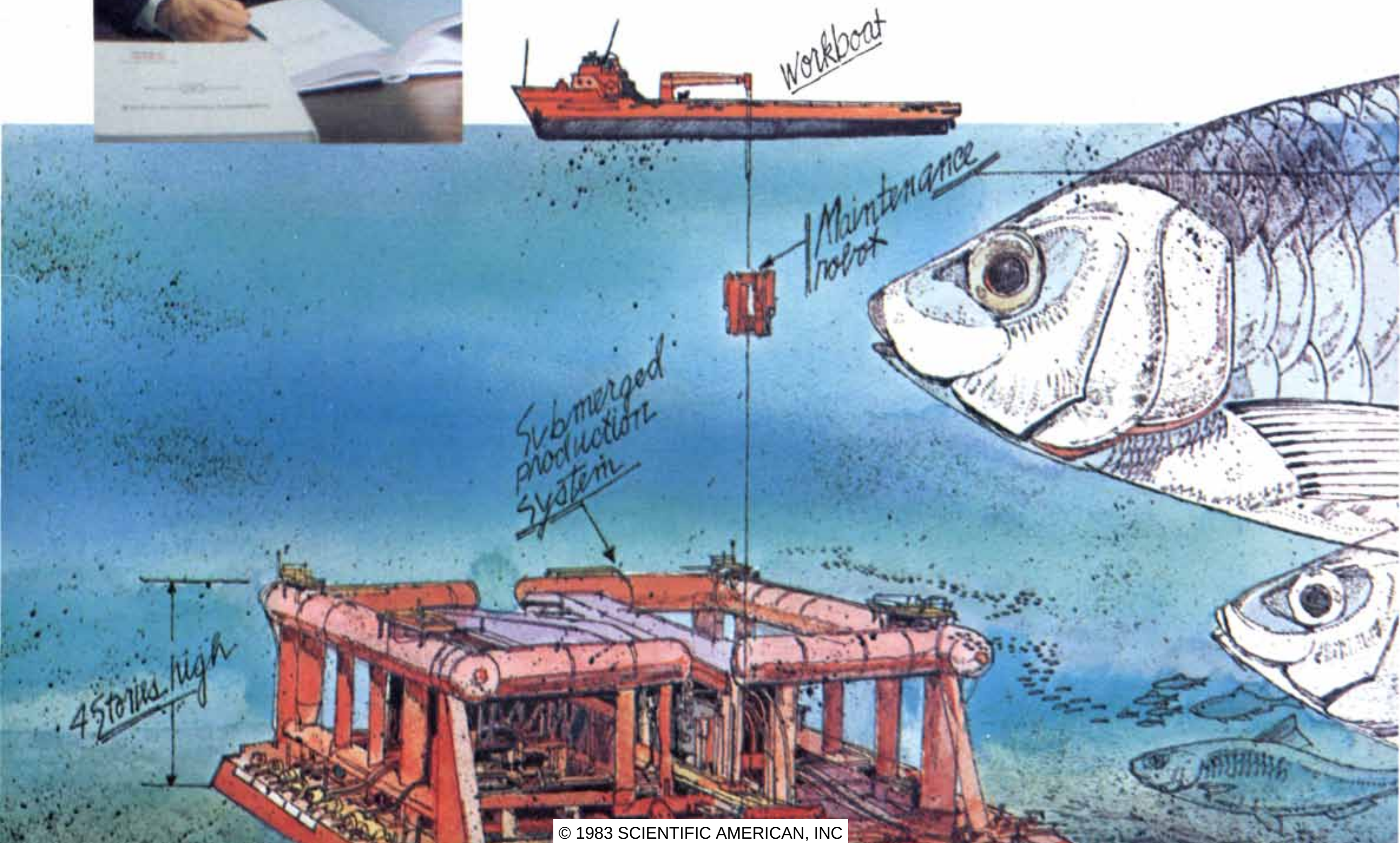
In 1968, Exxon leased tracts off the coast of Southern California, including some under 1800 feet of water. This made it necessary to study ways of producing oil in waters too deep to build platforms. These studies began in a flooded pit near Ventura, California, and culminated in the Gulf of Mexico with the 1974-79 pilot test of Exxon's Submerged Production System—a massive manifold set on the floor of the Gulf.

This prototype SPS controlled three producing wells drilled through openings in its base and delivered the oil to a nearby platform. Although the test was conducted in only 170 feet of water, it proved that the system could produce oil almost anywhere in the sea where a well could be drilled.

Now, three years later, the first commercial application of SPS technology is on the floor of the North Sea, installed by Exxon and another company.

Subsea maintenance by robot

SPS's have the potential to operate in water depths inaccessible to human divers. So Burkhardt and his



© 1983 SCIENTIFIC AMERICAN, INC

a new step in technology.

colleagues developed sophisticated robotics to perform routine maintenance.

The robot Remote Maintenance System (RMS) lowers itself along a buoyed cable from a surface vessel to the SPS. It then travels along a track which gives it access to critical valves and control components. The RMS can replace faulty parts, perform pressure tests and carry out various other maintenance functions. The entire robot operation is observed on closed-circuit television and remotely controlled from the surface vessel.

The SPS and the environment

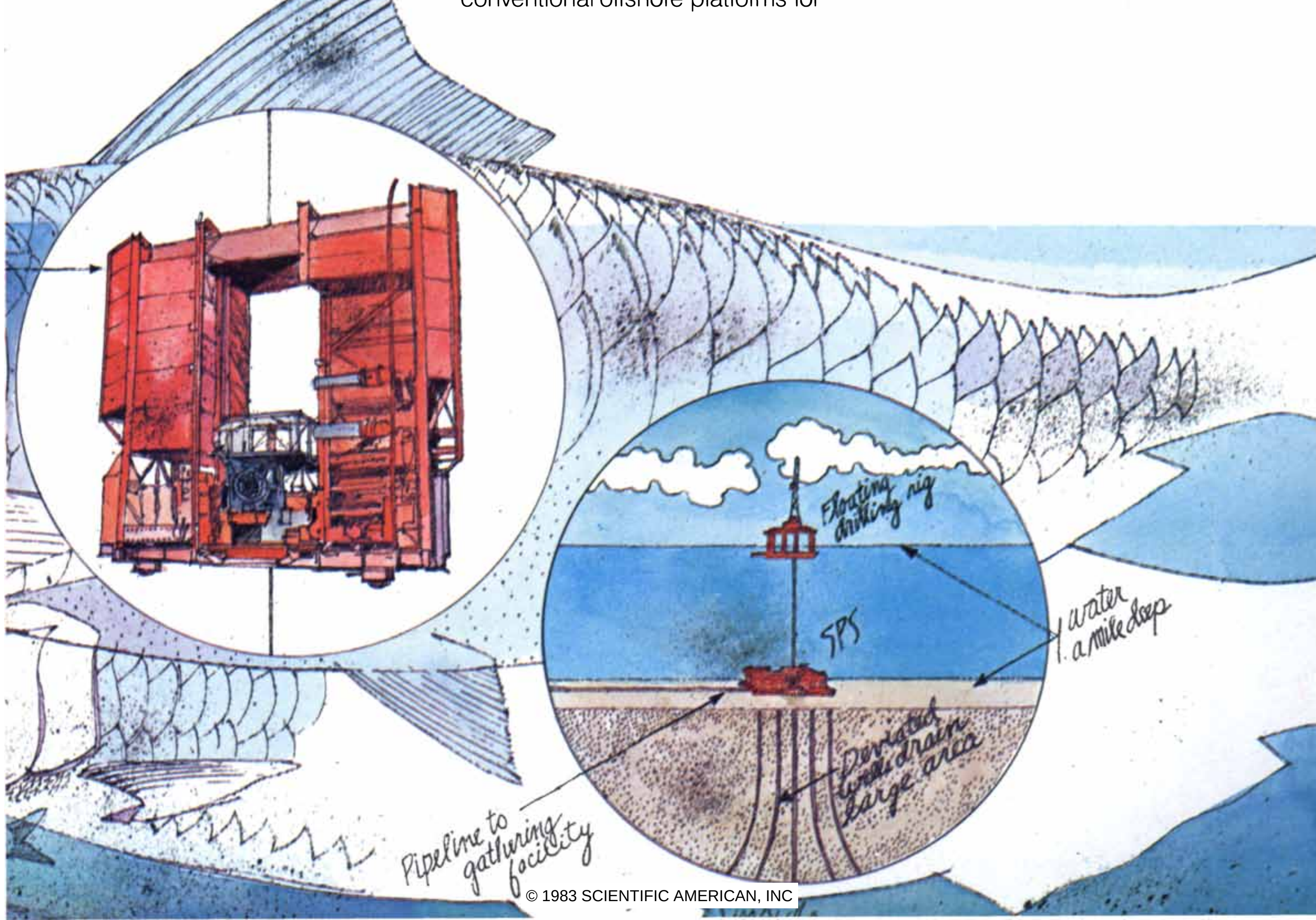
Safeguards incorporated in the SPS reflect Exxon's continuing concern for the environment. A high degree of redundancy is built into many of its subsystems. Fail-safe electrohydraulic control and shut-in systems are used throughout. In the unlikely event that leakage should occur, manifolding on the SPS is covered by inverted "pans" which would collect the leaked oil, trigger an alarm, and stop the leak by closing valves.

The potential of the SPS

While it is impractical to build conventional offshore platforms for water depths much beyond 1,000 feet, SPS technology could be practical to depths of a mile or more. If this full potential is realized, the SPS could be used to open up new areas for oil and gas recovery—areas that cannot be tapped today.

The design and construction of sophisticated deepwater energy systems is just one example of the research and development carried out by Exxon.

EXXON



© 1983 SCIENTIFIC AMERICAN, INC

DISCOVER THE TASTE OF THE RENAISSANCE.

·AMARETTO DI SARONNO·ORIGINALE·1525·

FOR FREE FOOD AND DRINK RECIPE BOOKS, WRITE: BOOKLETS, P.O. BOX 99085, LOUISVILLE, KY. 40299.

LIQUEUR 56 PROOF SOLE U.S. DISTRIBUTOR, FOREIGN VINTAGES, INC., N.Y., N.Y. ©1982 A UNIQUE PRODUCT FROM ILLVA SARONNO, ITALY

© 1983 SCIENTIFIC AMERICAN, INC

ence 6 3)" evaluates to 3 and their product is 27. This notation, in which each operation is placed to the left of its operands, was invented by the Polish logician Jan Łukasiewicz before computers existed. Unfortunately for Lukasiewicz his name was too formidable-looking for most speakers of English, and so this type of notation came to be called "Polish notation." Now, here is a simple problem in this notation for you, in which you are to play the part of the Lisp genie:

```
—> (quotient
      (plus 21 13)
      (difference
        23
        (times
          2
          (difference 7 (plus 2 2)))))
```

Perhaps you have noticed that statements of Lisp involve parentheses. A profusion of parentheses is one of the hallmarks of Lisp. It is not uncommon to see an expression that terminates in a dozen parentheses! This makes many people shudder at first, and yet once you get used to their characteristic appearance Lisp expressions become remarkably intuitive, even charming, to the eye, particularly when they are "pretty-printed," which means that a careful indentation scheme is followed that reveals their logical structure (as in the preceding expression).

The heart of Lisp is its manipulable structures. All programs in Lisp work by creating, modifying and destroying structures. Structures come in two types: atomic and composite, or, as they are usually called, *atoms* and *lists*. Thus every Lisp object is either an atom or a list (but not both). The only exception is the special object called "nil", which is both an atom and a list. More about nil below. What are some other typical Lisp atoms? Here are a few:

hydrogen, helium, j-s-bach, 1729, pi, foo, bar, baz, buttons-&-bows

Lists are the flexible data structures of Lisp. A list is pretty much what it sounds like: a collection of some parts in a specific order. The parts of a list are usually called its *elements* or *members*. What can these members be? Well, not surprisingly, lists can have atoms as members. Just as easily, however, lists can contain lists as members, and those lists can in turn contain other lists as members, and so on recursively. Oops! I jumped the gun with that word, but no harm done. You certainly understood what I meant, and it will prepare you for a more technical definition of the term to come later.

A list printed on your screen is recognizable by its parentheses. In Lisp anything bounded by matching parentheses constitutes a list. For instance, "(zonk blee strill (cronk flonk))" is a four-element list whose last element is itself a two-element list. Another short list is "(plus 2 2)", illustrating the fact that Lisp statements themselves are lists. This is important because it means that the Lisp genie, by manipulating lists and atoms, can actually construct new wishes by itself. Hence the object of a wish can be the construction—and subsequent evaluation—of an entirely new wish!

Then there is the empty list: the list with no elements at all. How is this written down? You might think that "()" would work. Indeed, it will work, but there is a second way of indicating the empty list, and that is by writing "nil". The two notations are synonymous, although "nil" is more commonly written than "()". The empty list, nil, is a key concept of Lisp; in the universe of lists it is what zero is in the universe of numbers. To use another metaphor for nil, it is like the earth in which all structures are rooted. To understand what this means, however, you will have to wait a bit longer.

The most commonly exploited feature of an atom is that it has (or can be given) a value. Some atoms have permanent values; others are variables. As you might expect, the value of the atom "1729" is the integer 1729, and this is permanent. (I am distinguishing here between the atom whose "print name" or "pname" is the four-digit string "1729" and the eternal Platonic essence that happens to be the sum of two cubes in two different ways.) The value of nil is also permanent, and it is—nil! One other atom that has itself as its permanent value is the special atom "t".

Apart from t, nil and atoms whose names are numerals, atoms are generally variables, which means that you can assign values to them and later change their values at will. How is it done? Well, to assign the value 4 to the atom "pie" you can type to the Lisp genie "(setq pie 4)". Or you could just as well type "(setq pie (plus 2 2))", or even "(setq pie (plus 1 1 1 1))". In any of these cases, as soon as you type your carriage return, pie's value will become 4, and so it will remain forevermore, or at least until you perform another setq operation on the atom "pie".

Lisp would not be crisp if the only values atoms could have were numbers. Fortunately an atom's value can be set to any kind of Lisp object—any atom or list whatsoever. For instance, we might want to make the value of the atom "pi" be a list such as "(a b c)" or perhaps "(plus 2 2)" instead of the number 4. The value belonging to an atom is often called the atom's *binding*, and one says the atom is *bound* to that value. To assign new bindings to the atoms "pie" and "pi" we again use the "setq" operation. To illustrate, here follows a brief conversation with the genie:

```
—> (setq pie (plus 2 2))
4
—> (setq pi '(plus 2 2))
(plus 2 2)
—>
```

Notice the vast difference between the values assigned to the atoms "pie" and "pi" as a result of these two wishes asked of the Lisp genie, which differ merely in the presence or absence of the small but critical single-quote mark in front of the inner list "(plus 2 2)". In the first wish, since there is no quote mark, that inner "(plus 2 2)" must be *evaluated*. This returns 4, which is assigned to the variable "pie" as its new value. On the other hand, in the second wish, since the quote mark is there, the list "(plus 2 2)" is never executed as a command but is treated merely as an inert lump of Lispstuff, much like meat on a butcher's shelf. It is ever so close to being "alive," yet it is dead. Therefore the value of pi in this second case is the list "(plus 2 2)", a fragment of Lisp code. The following interchange with the genie confirms the values of these atoms:

```
—> pie
4
—> pi
(plus 2 2)
—> (eval pi)
4
—>
```

What is this last step? I wanted to show how you can ask the genie to *evaluate* the value of an expression rather than simply *printing* the value of that expression. Ordinarily the genie automatically performs just one level of evaluation, but by writing "eval" you can get a second stage of evaluation carried out. (And of course by using eval over and over again you can continue as long as you like.) This feature often proves invaluable, but it is a little too advanced to discuss further at this stage.

Every list except nil has at least one element. This first element is called the list's "car". Thus the car of "(eval pi)" is the atom "eval". The cars of the lists "(plus 2 2)", "(setq x 17)", "(eval pi)" and "(car pi)" are all names of operations or, as they are more commonly called in Lisp, *functions*. The car of a list need not be the name of a function; it need not even be an atom. For instance, "((1) (2 2) (3 3 3))" is a perfectly fine list. Its car is the list "(1)", whose car in turn is not a function name but merely a numeral.

If you were to remove a list's car, what would remain? A shorter list. This is called the list's "cdr", a word that is spoken so that it sounds about halfway be-

© 1983 SCIENTIFIC AMERICAN, INC

IF A CAR IS REALLY AN EXTENSION OF ONE'S PERSONALITY, WHAT KIND OF PERSON WOULD DRIVE A SAAB?

Not long ago, a leading car magazine called Saab owners "the lunatic fringe of the American car-buying public."

Yet according to our statistics, the average Saab owner is male, age 38, college-educated, works in a managerial job, and earns over $40,000 a year. He is married and has 1.2 children.

The fact is, both descriptions are accurate.

The fringe.

Some people call this person a driving enthusiast; others call him a car nut.

Whatever you call him, he buys a car for one reason.

Economy? Who cares. Luggage space? Who needs it.

His attitude is if a car doesn't give you goose bumps when you drive it, what's the point of owning it.

For him, even a drive to the supermarket should be exhilarating.

For that, Saab's front-wheel drive and taut suspension give him the cornering ability of a sports car.

And every time Saab's new APC turbocharger kicks in, he feels like he's just engaged warp drive.

Engineering philosophy doesn't interest him. Results do.

Often, he belongs to a car club.

Not the kind with leather jackets and secret handshakes.

But every month or so, they sponsor an event called an Autocross. Much to the dismay of the local townspeople, club members roar their Saabs against the clock through staid suburban parking lots.

Beyond the fringe.

At the other end of the spectrum is the Saab owner who is largely responsible for the respectable statistics that were cited earlier.

He bristles at Saab's cult car reputation. He thinks of car clubs in the same light as motorcycle gangs.

Nonetheless, he does realize that many of Saab's "radical" innovations like turbocharging, front-wheel drive, and aerodynamic design have broader applications than just blowing your neighbor's BMW off the road.

He sees the safety in high performance every time he merges onto a crowded freeway or passes a truck on a two-lane highway.

And, in a Saab APC Turbo, this performance is attained without sacrificing fuel economy. In fact, the APC system actually improves gas mileage.*

He sees the logic of Saab's front-wheel drive and four-wheel disc brakes, especially after the first snowfall. Or the last rainfall.

Even Saab's hatchback design, which some find unconventional, he finds practical, considering that it gives his Saab the carrying capacity of a station wagon.

And not only does his Saab have plenty of room for luggage, it also has plenty of room for people. More, in fact, than many elitist cars.

For those who insist on luxury for luxury's sake, Saab has made one concession. Some turbo models are now equipped with an *Exclusive Appointments Group* that includes leather-upholstered seats and electric sunroofs. (That's really two concessions, isn't it?)

*1983 SAAB PRICE** LIST*

900 3-door	*$10,750*
900 4-door	*$11,050*
900S 3-door	*$13,550*
900S 4-door	*$13,950*
900 Turbo 3-door	*$16,510*
900 Turbo 4-door	*$16,910*

Automatic transmission $370 additional.

Even with leather upholstery and sunroofs you don't have to open manually, Saabs have not replaced Mercedes and BMW as the standard-bearer at the country club.

But for Saab owners, whatever type they may be, the experience of driving a Saab outweighs the lure of status.

It has to.

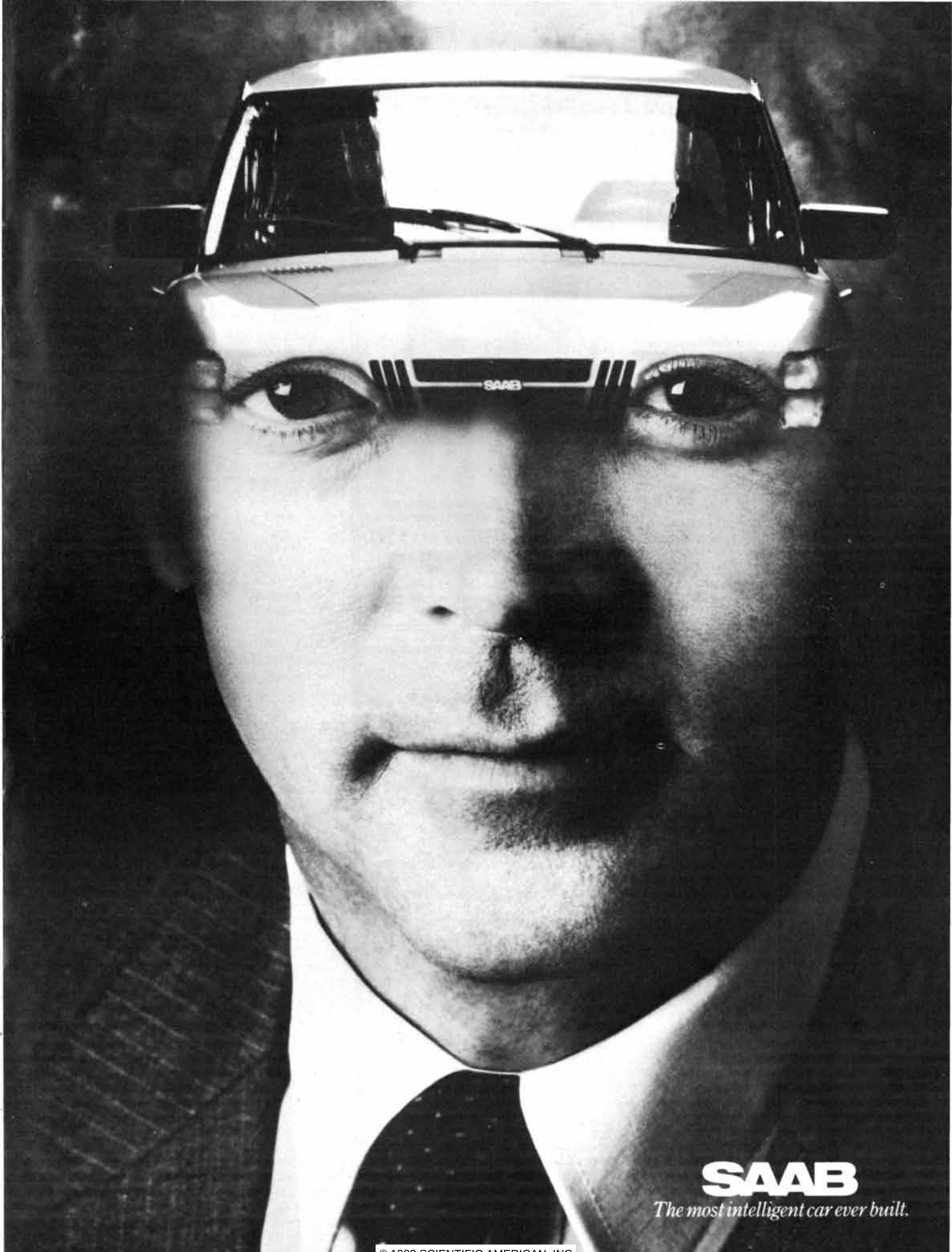
How else could they get a practical car that drives as well as most wildly impractical cars? A car that appeals to their emotions as well as their intellect?

So what kind of person drives a Saab?

A very satisfied one.

**Saab 900 5-speed APC Turbo: (21) EPA estimated mpg, 34 estimated highway mpg. Use estimated mpg for comparison only. Mileage varies with speed, trip length and weather. Actual highway mileage will probably be less. **Manufacturer's suggested retail prices. Not including taxes, license, freight, dealer charges or options.*

© 1983 SCIENTIFIC AMERICAN, INC



© 1983 SCIENTIFIC AMERICAN, INC

PSST!
WANT AN UNLISTED TELEPHONE NUMBER?

We can give you one with the press of a button.
That's right. The press of a button.

No, we're not Ma Bell. We're International Mobile Machines Corporation. And with the press of a button, we make it possible for you to create, not one, but as many as 16 separate, unlisted, private telephone numbers... whenever you want them, as often as you want them.

How? With PriveCode®- the Telephone Access Control Terminal™... an affordable, convenient alternative to Ma Bell's unlisted number system.

Here's how it works:
Just plug PriveCode into any modular jack in your home or apartment, turn it on, and you're in complete control over who can reach you. You simply assign a 3 digit Access Code to people you want to hear from. Whenever anyone dials your regular listed telephone number, PriveCode intercepts their call and **stops all your phones from ringing.** PriveCode then asks the caller to enter their Personal Access Code. Only after a valid code has been entered will PriveCode signal you with a pleasing electronic sound, and display the caller's personal code, so you know who is calling **before** you pick up your phone. PriveCode even tells you who called while you were out. Callers without valid Access Codes never get through. Since PriveCode handles up to 16 different codes at a time it's just like having 16 private lines... on your present Ma Bell line. So say "Goodbye" to wrong numbers and heavy breathers... say "Hello"... only when you want to.

To create the privacy, and the security of an unlisted telephone number with the press of a button, say "Hello" to PriveCode.

PriveCode is available in fine department stores, electronic specialty shops and phone stores.

See a PriveCode demonstration today, call toll-free for the PriveCode dealer nearest you.

800-523-0103, Ext. 110

(In PA call 215-569-1300, ext. 110)

For further information write
International Mobile Machines Corporation
100 North 20th Street, Department No. 111
Philadelphia, Pennsylvania 19103

PRIVECODE
The Telephone Access Control Terminal™

New Management Tool!

It's called the **Executive ScanCard™ System,** and it serves as a portable "control center" for monitoring all your projects on a daily basis, so you can double, even triple, the number of projects you can keep moving simultaneously.

Easy to use
Projects are recorded on ScanCard™ project cards that can be quickly scanned to jog your memory on what's to be done next.

Expandable
Initially holds 64 projects. More panels (40 cards each) and ScanCards can be ordered as needed.

The "Executive Look"
In handsome, rugged leather like material that will last for years. Available in black, brown, tan, burgundy or natural suede.

National Introductory Price
$39.95
Add $4 shipping and handling
Ohio residents add sales tax.

Your name printed here

Important extras
Includes pen holder and space for your tablet. Plus 500 ScanCards and illustrated instruction booklet.

FREE
Handy **Pocket Idea™** file for carrying small supply of ScanCards. Jot your ideas on cards and merge into system later.

Satisfaction Guaranteed
If for any reason you are not satisfied, you may return the system within 30 days for a full refund.

To order phone
800-848-2618
IN OHIO CALL 800-282-2630

or mail check to
Executive ScanCard™ Systems Dept. 748
6480 Busch Blvd. Suite 200
Columbus, Ohio 43229

International Orders — Shipping & Handling Charges: Canada — $7.50, Mexico — $12.00, All other International — $18.00. International payments in U.S. Dollars.

We honor American Express, MasterCard, Visa and Diners Club.

PATENTS PENDING©COPYRIGHT 1981 EXECUTIVE SCANCARD™ SYSTEMS

tween "kidder" and "could 'er." (The words "car" and "cdr" are relics from the first implementation of Lisp on the IBM 704. The letters in "car" and "cdr" stand respectively for "contents of the address part of register" and "contents of the decrement part of register." They refer to specific hardware features of that machine, now long since irrelevant.) The cdr of "(a b c)" is the list "(b c)", whose cdr is "(c)", whose cdr is nil. And nil has no cdr, just as it has no car. Attempting to take the car or cdr of nil causes (or should cause) the Lisp genie to cough out an error message, just as attempting to divide by zero should evoke an error message.

Here is a little table showing the cars and cdrs of a few lists, just to make sure the notion is unambiguous:

list	*car*	*cdr*
((a) b (c))	(a)	(b (c))
(plus 2 2)	plus	(2 2)
((car x) (car y))	(car x)	((car y))
(nil nil nil nil)	nil	(nil nil nil)
(nil)	nil	nil
nil	ERROR	ERROR

Just as car and cdr are called *functions,* so the things they operate on are called their *arguments.* Hence in the command "(plus pie 2)", "plus" is the function name and the arguments are the atoms "pie" and "2". In evaluating this command (and most other commands) the genie figures out the values of the arguments and then applies the function to those values. Since the value of the atom "pie" is 4 and the value of the atom "2" is 2, the genie returns the atom "6".

Suppose you have a list and you would like to see a list just like it, only one element longer. For instance, suppose the value of the atom "x" is "(cake cookie)" and you would like to create a new list called "y" just like x, except with an extra atom—say "pie"—at the front. You can then use the function called "cons" (short for "construct"), whose effect is to make a new list out of an old list and a suggested car. Here is a transcript of such a process:

```
—> (setq x '(cake cookie))
(cake cookie)
—> (setq y (cons 'pie x))
(pie cake cookie)
—> x
(cake cookie)
```

Two things are worth noticing here. I asked for the value of x to be printed out after the cons operation, so that you could see that x itself was not changed by that operation. Cons created a new list and made that list be the value of y, but it left x entirely alone. The other noteworthy fact is that I used that quote mark again, in front of the atom

© 1983 SCIENTIFIC AMERICAN, INC

EXTINCTION. IT'S FOREVER.

One quarter of all species of animals and plants on Earth may disappear in the next 30 years because of man's destruction of their habitat. The rate of extinctions is increasing enormously as forests are destroyed and other wild areas are lost. Organisms that evolved over hundreds of millions of years will be gone forever. The complex interdependence of all creatures, from the largest mammals to the smallest plants, is being shattered. It is a crisis with profound implications for the survival of all life. Unfortunately, little is being done to save our planet's natural heritage. Here are some warnings by leading scientists:

"The worst thing that can happen—will happen—in the 1980's is not energy depletion, economic collapse, limited nuclear war, or conquest by a totalitarian government. As terrible as these catastrophes would be for us, they can be repaired within a few generations. The one process ongoing in the 1980's that will take millions of years to correct is the loss of genetic and species diversity by the destruction of natural habitats. This is the folly our descendants are least likely to forgive us."

DR. EDWARD O. WILSON
Baird Professor of Science • Harvard University

"We are encroaching on nature, in the U.S. and around the world, at an unprecedented rate. A large proportion of the chemicals in use in our present-day civilization were 'invented' by nature, not by the chemist in the laboratory. An estimated 40% of all drug prescriptions in the U.S. contain as their chief ingredients compounds derived from plants. There is no end to the potential for discovery in nature, because we have only begun the chemical exploration of nature. Tragically, we are burning our library of priceless genetic treasures with our reckless destruction of species."

DR. THOMAS EISNER
Schurman Professor of Biology • Cornell University

"Few problems are less recognized but more important than the accelerating disappearance of Earth's biological resources. In pushing other species to extinction, humanity is busily sawing off the limb on which it is perched."

DR. PAUL EHRLICH
Bing Professor of Population Studies • Stanford University

"The extermination of a quarter of the plant species on Earth during the coming few decades could lose forever the medicines and food sources we need to cure the diseases and hunger that plague mankind."

DR. PETER RAVEN
Director • Missouri Botanical Garden

The Ark is sinking. We need the help of every concerned citizen to conserve the diversity of life on Earth. For information about how you can help, please write to us at:

Save Endangered Species

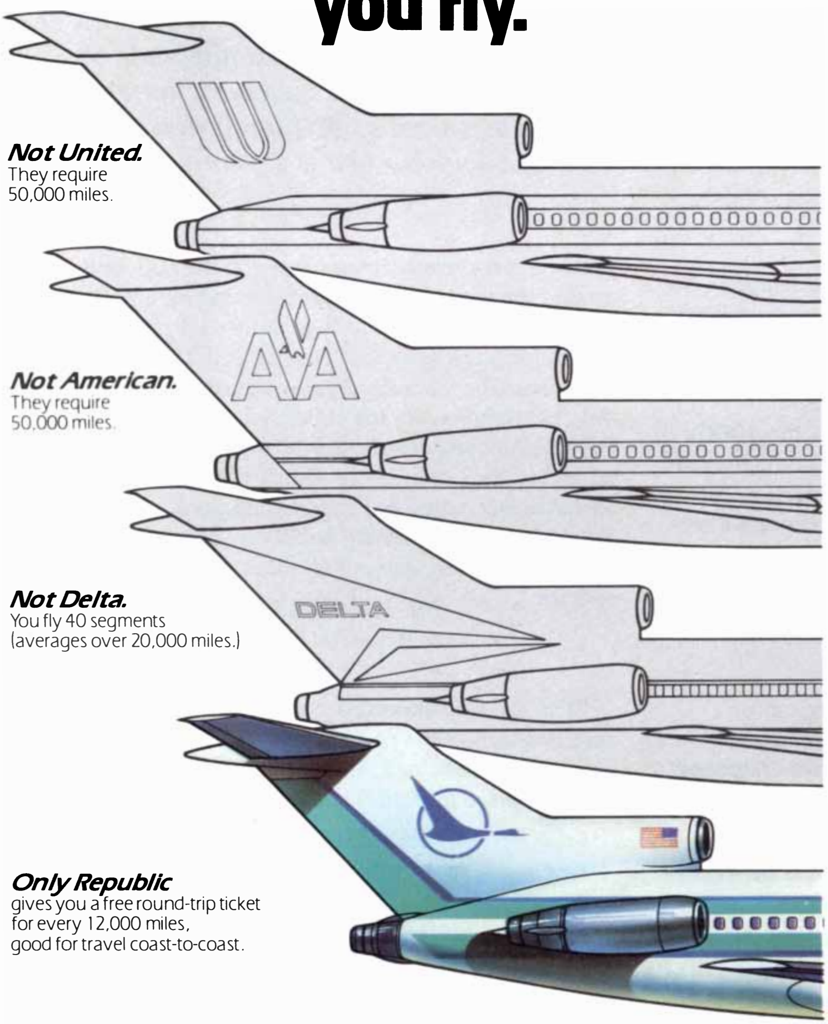
P.O. Box 50771 • Washington, D.C. 20004

Sponsored by:

- American Association of Zoological Parks and Aquariums
- American Cetacean Society
- American Humane Association
- American Institute of Biological Sciences
- Animal Protection Institute of America
- California Native Plant Society
- Center for Action on Endangered Species
- Center for Environmental Education
- Defenders of Wildlife
- Desert Fishes Council
- Environmental Defense Fund
- Florida Audubon Society
- Friends of the Earth
- Friends of the Sea Otter
- Fund for Animals
- Garden Club of America
- Greenpeace U.S.A.
- Humane Society of the United States
- International Council for Bird Preservation
- Massachusetts Audubon Society
- National Audubon Society
- National Parks and Conservation Association
- National Shell Fisheries Association
- Natural Resources Defense Council
- Ocean Research and Education Society
- Sierra Club
- Society for Animal Protective Legislation
- Wilderness Society
- World Wildlife Fund - U.S.

© 1983 SCIENTIFIC AMERICAN, INC

Nobody but Republic gives you free travel for every 12,000 miles you fly.



Not United.
They require 50,000 miles.

Not American.
They require 50,000 miles.

Not Delta.
You fly 40 segments (averages over 20,000 miles.)

Only Republic
gives you a free round-trip ticket for every 12,000 miles, good for travel coast-to-coast.

The new Republic Frequent Flyer™ Program: enroll now and begin earning your free round-trip travel.

Name ______ Company ______

Phone: Home ( ) ______ Business ( ) ______

Do you prefer to receive your Frequent Flyer mail at: ☐ Home ☐ Business

Address ______

City ______ State ______ Zip ______

Mail to Republic Airlines Frequent Flyer Program, Dept. SA , P.O. Box 20452, Atlanta, GA 30325.
We'll forward your personal account number and full information.

NOBODY SERVES OUR REPUBLIC LIKE REPUBLIC™ AIRLINES

"pie". What if I had not used it? This is what would have happened:

```
—> (setq z (cons pie x))
(4 cake cookie)
```

Remember, after all, that the atom "pie" still has the value 4, and whenever the genie sees an unquoted atom inside a wish, it will always use the value belonging to that atom rather than the atom's name. (Always? Well, almost always. I shall explain in a moment. In the meantime look for an exception; you have already encountered it.)

Now here are a few exercises for you, some of them a bit tricky. Watch out for the quote marks! One last thing: I use the function "reverse", which produces a list just like its argument, only with its elements in reverse order. For instance, the genie, on being told "(reverse '((a b) (c d e)))", will write "((c d e) (a b))". The genie's lines in this dialogue are given afterward.

```
—> (setq w (cons pie '(cdr z)))
—> (setq v (cons 'pie (cdr z)))
—> (setq u (reverse v))
—> (cdr (cdr u))
—> (car (cdr u))
—> (cons (car (cdr u)) u)
—> u
—> (reverse
        '(cons (car u) (reverse (cdr u))))
—> (reverse
        (cons (car u) (reverse (cdr u))))
—> u
—> (cons 'cookie
          (cons 'cake
                (cons 'pie nil)))
```

The answers (as printed by the genie) are as follows:

```
(4 cdr z)
(pie cake cookie)
(cookie cake pie)
(pie)
cake
(cake cookie cake pie)
(cookie cake pie)
((reverse (cdr u)) (car u) cons)
(cake pie cookie)
(cookie cake pie)
(cookie cake pie)
```

The last example, featuring the repeated use of cons, is often called, in Lisp slang, "consing up a list." You begin with nil and then do repeated cons operations. It is analogous to building a positive integer by starting at zero and then performing the successor operation over and over again. However, whereas at any stage in the latter process there is a unique way of doing the successor operation, given any list there are infinitely many different items you might cons onto it, thus giving rise to a vast branching tree of lists instead of the unbranching number line. It is because of this im-

© 1983 SCIENTIFIC AMERICAN, INC

age of a tree growing out of the ground of nil and containing all possible lists that I likened nil to the earth in which all structures are rooted.

As I mentioned above, the genie does not *always* replace (unquoted) atoms by their values. There are cases where a function treats its arguments, although unquoted, as if they were quoted. Did you go back and find such a case? It is not hard. The answer is the function "setq". In particular, in a setq command the first atom is taken straight, not evaluated. As a matter of fact, the "q" in "setq" stands for "quote," meaning that the first argument is treated as if it were quoted. Things can get tricky when you learn about "set", a function similar to setq except that it *does* evaluate its first argument. Hence if the value of the atom "x" is the atom "k", then entering "(set x 7)" will not do anything to x (its value will remain the atom "k") but the value of the atom "k" will now become 7. Watch closely:

```
—> (setq a 'b)
—> (setq b 'c)
—> (setq c 'a)
—> (set a c)
—> (set c b)
```

Now tell me: What are the values of the atoms "a," "b" and "c"? Here comes the answer, so try not to peek. They are respectively "a", "a" and "a". This may be confusing. You may be reassured to know that in Lisp, "set" is not very commonly used and such confusions do not arise often.

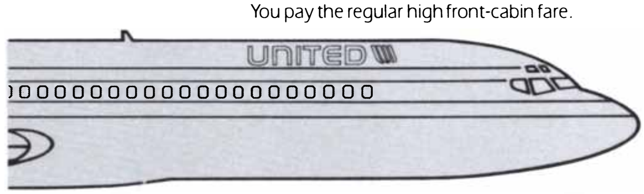
Psychologically, one of the great powers of programming is the ability to define new compound operations in terms of old ones, and to do this over and over again, thus building up a vast repertory of ever more complex operations. It is reminiscent of the early stages of biological evolution, when ever more complex molecules evolved out of less complex ones in an ever upward helix of complexity and creativity. It is also reminiscent of the Industrial Revolution, in which people used simple early machines to help them build more complex machines, then used those in turn to build even more complex machines, again in an ever upward helix of complexity and creativity. At each stage, whether in evolution or Revolution, the products get more flexible and more intricate, more "intelligent" and yet more vulnerable to subtle "bugs" or breakdowns.

So it is with programming in Lisp, except that here the "molecules" or "machines" are Lisp functions defined in terms of previously known Lisp functions. Suppose, for instance, you want to have a function that will always return the last element of a list, just as "car" always returns the first element of a

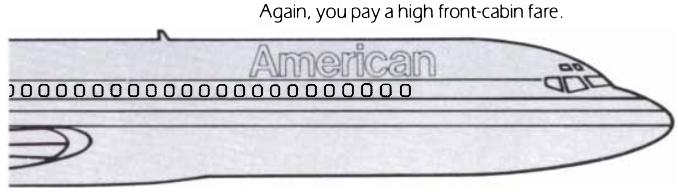
Nobody
but Republic gives you first class comfort for just $10 over Coach.

(No matter how far you're flying.)

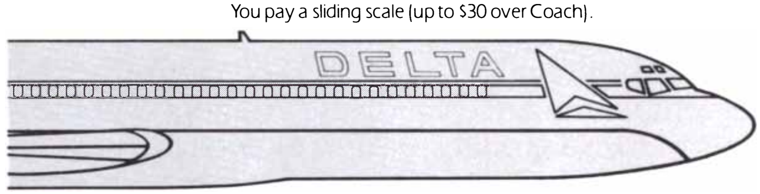
Not United.
You pay the regular high front-cabin fare.



Not American.
Again, you pay a high front-cabin fare.



Not Delta.
You pay a sliding scale (up to $30 over Coach).



Only Republic
flies you in front-cabin comfort for only $10 extra.

Republic Business Coach:™ the fare is as comfortable as the flight.

For a flat $10 more than our one-way nondiscounted Coach fare, you can enjoy all the amenities of Republic Business Coach. The privacy of a separate up-front cabin. A comfortable two-by-two seat. Complimentary beverages. And priority baggage handling. A great place to prepare for the business day. A great place to relax afterward. Available on most flights. Call a travel agent, your corporate travel planner or Republic.

NOBODY SERVES OUR REPUBLIC LIKE REPUBLIC™
AIRLINES

© 1983 SCIENTIFIC AMERICAN, INC

THE CAR THAT TOOK TEN YEARS TO BUILD.

In the past ten years alone, no fewer than 30 "exciting" new cars have been introduced in the U.S., hyped, sold and then quietly discontinued.

Remember the once-famous Monza, Versailles, Pacer, Dasher, Fox, Astre, Monarch, Sunbird, Bobcat, Starfire, Aspen, Volare? Where, oh where, are they now?

In that same ten year period Volvo has been developing a single, remarkable new automobile. The 760 GLE.

As excited as we were about this car, we weren't about to introduce it until it was as close to perfect as we could make it.

So prototypes were driven a total of two million miles through some of the most gruelling conditions on earth. They hauled trailers over the Alps, plowed through snowbanks near the Arctic Circle, pounded through potholes in New York City and ate dust in Australia. They endured five searing summers in the Nevada desert, frequently being driven 24 hours a day.

All this testing led to hundreds of design improvements. And after we'd made those improvements, we improved on the improvements.

What emerged from all this is a car that's astonishing indeed.

The 760 GLE is more aerodynamically efficient than a Porsche 928, yet roomy enough for five adults.

It has traveled through a high-speed slalom course faster than a Jaguar or Audi.

It has a climate control system so advanced it can change the interior air five times a minute while the car is standing still.

The Volvo 760 GLE.

It took a long time to get here. But we wanted to make sure that, unlike so many cars, once it got here it stayed.

THE VOLVO 760 GLE

The closest thing yet to a perfect car.

© 1983 VOLVO OF AMERICA CORPORATION.

© 1983 SCIENTIFIC AMERICAN, INC

list. Lisp does not come equipped with such a function, but you can easily create one. Do you see how? To get the last element of a list called "lyst", you simply reverse lyst and then take the car of that: "(car (reverse lyst))". To dub this operation "rac" ("car" backward) we use the "def" function, as follows:

```
—> (def rac (lambda (lyst)
        (car (reverse lyst))))
```

Using def in this way creates a *function definition.* In it the word "lambda" followed by "(lyst)" indicates that the function we are defining has only one parameter, or dummy variable, to be called "lyst". (It could have been called anything; I just happen to like the atom "lyst".) In general the list of parameters must immediately follow the word "lambda". After this "def wish" has been carried out the rac function is as well understood by the genie as the car function is. Thus "(rac '(your brains))" will yield the atom "brains". And we can use rac itself in definitions of still further functions. The whole thing snowballs rather miraculously, and you can quickly become overwhelmed by the power you wield.

Here is a simple example. Suppose you have a situation where you know you are going to run into many big long lists and you know it will often be useful to form, for each such long list, a short list that includes just its car and rac. We can define a one-parameter function to do it for you:

```
(def readers-digest-condensed-version
   (lambda (biglonglist)
    (cons (car biglonglist)
          (cons (rac biglonglist) nil))))
```

Hence if we apply the function "readers-digest-condensed-version" to the entire text of James Joyce's *Finnegans Wake* (treating it as a big long list of words), we shall obtain the shorter list "(riverrun the)". Unfortunately reapplying the condensation function to this new list will not simplify it any further.

It would be nice as well as useful if we could create an inverse operation to readers-digest-condensed-version called "rejoyce" that, given any two words, would create a James Joyce novel beginning and ending with them respectively. Thus execution of the Lisp statement "(rejoyce 'Stately 'Yes)" would result in the Lisp genie's generating from scratch the entire novel *Ulysses.* Writing this function is left as an exercise for the reader.

One goal that has seemed to some people to be both desirable and feasible with Lisp and related programming languages is (1) to make every single statement return a *value* and (2) to have it be through this returned value and *only* through it that the statement has any effect. The idea of (1) is that values are handed "upward" from the innermost function calls to the outermost ones, until the full statement's value is returned to you. The idea of (2) is that during all these calls no atom has its value changed at all (unless the atom is a dummy variable). In all dialects of Lisp that are known to me (1) is true but (2) is not necessarily true.

Thus if x is bound to "(a b c d e)" and you say "(car (cdr (reverse x)))", the first thing that happens is that the value of "(reverse x)" is calculated; then this value is handed "up" to the cdr function, which calculates the cdr of that list; finally this shorter list is handed to the car function, which extracts one element, the atom "d", and returns it. In the meantime the atom "x" has suffered no damage; it is still bound to "(a b c d e)".

It might seem that an expression such as "(reverse x)" would change the value of x by reversing it, just as carrying out the oral command "Turn your sweater inside out" will affect the sweater. Actually carrying out the wish "(reverse x)" no more changes the value of x than carrying out the wish "(plus 2 2)" changes the value of 2. Instead executing "(reverse x)" causes a *new,* unnamed list to come into being, just like x, only reversed. And that list is the value of the statement; it is what the statement returns. The value of x itself, however, is untouched. Similarly, evaluating "(cons 5 pi)" will not change the value of the list named "pi" in the slightest; it merely *returns* a new list with 5 as its car and whatever pi's value is as its cdr.

Such behavior is to be contrasted with that of functions that leave "side effects" in their wake. Such side effects are usually in the form of changed variable bindings, although there are other possibilities, such as causing input or output to take place. A typical "harmful" command is a setq, and proponents of the "applicative" school of programming—the school that says you should never have any side effects whatever—are profoundly disturbed by the mere mention of setq. For them all results must come about purely by the way functions compute their values and hand them along to other functions.

The only bindings the advocates of the applicative style approve of are transitory "lambda bindings": those that arise when a function is applied to its arguments. Whenever any function is called, that function's dummy variables temporarily assume "lambda bindings." These bindings are just like those resulting from a setq except that they are fleeting. That is, the moment the function is finished computing they go away, vanishing without a trace. For example, in the computation of "(rac '(a b c))" the lambda binding of the dummy variable "lyst" is to the list "(a b c)", but as soon as the answer "c" is passed along to the person or function that requested the rac, the value of the atom "lyst" used in getting that answer is totally forgotten. If you ask for the value of lyst, the Lisp interpreter will tell you that the atom "lyst" is an "unbound atom." Applicative programmers much prefer lambda bindings to ordinary setq bindings.

I personally am not a fanatic about avoiding setq's and other functions that cause side effects. Although I find the applicative style elegant, I think it impractical for the construction of large AI programs. Therefore I shall not advocate the applicative style here, although I shall adhere to it when possible. Strictly speaking, in applicative programming you cannot even define new functions, since a "def" statement causes a permanent change to take place in the genie's memory, namely the permanent storage in memory of the function definition. Thus the ideal applicative approach would have function definitions, like variable bindings, being created only temporarily and then discarded the moment after they had been used. This is extreme "applicativism."

For your edification here are a few more simple function definitions:

```
—> (def rdc
       (lambda (lyst)
        (reverse (cdr (reverse lyst)))))
—> (def snoc
       (lambda (x lyst)
        (reverse
         (cons x (reverse lyst)))))
—> (def twice (lambda (n) (plus n n)))
```

The functions rdc and snoc are analogous to cdr and cons, only backward. Hence the rdc of "(a b c d e)" is "(a b c d)", and if you type "(snoc 5 '(1 2 3 4))", you will get as your answer "(1 2 3 4 5)".

All of this is mildly interesting so far, but if you want to see the genie do anything *truly* surprising, you have to allow it to make some decisions based on things that happen along the way. These are sometimes called *conditional* wishes. A typical example would be the following:

```
—> (cond ((eq x 1) 'land)
          ((eq x 2) 'sea))
```

The value returned by this statement will be the atom "land" if x has the value 1, and the atom "sea" if x has the value 2. Otherwise the value returned will be nil (for instance if x is 5). The atom "eq" (pronounced "eek") is the name of a common Lisp function that returns the atom "t" (standing for "true") if its two arguments have the same value, and nil (for "no" or "false") if they do not.

A cond statement is a list whose car is the function name "cond", followed by any number of *cond clauses,* each of

© 1983 SCIENTIFIC AMERICAN, INC

A MAGAZINE FOR YOUR PIANO

Keyboard Classics Magazine *takes up where* ***The Etude*** *magazine left off. Filled with musical treasures, fancies and delights to be played on your piano, it is truly the one magazine you will enjoy again and again.*

Available only by subscription. Write To: Keyboard Classics, Dept. A1-05-23, 352 Evelyn Street, Paramus, NJ 07652. Enclose $9.97 for a full year's subscription (6 issues). Completely refundable if you are not delighted with your first issue.

SPOTTING SCOPES

Precision Optics From **CELESTRON®**

This 90mm (diameter) Spotting Scope is also a 1000mm mirror lens telephoto! We offer superb quality Spotting Scopes, Binoculars, Telephotos and Astronomical Telescopes in all price ranges. For **FREE** product, price and ordering information send coupon below (or send $3.00 for a detailed catalog/selection guide).

CALL TOLL FREE 1-800-421-1526

CELESTRON INTERNATIONAL
P.O. Box 3578-SA, 2835 Columbia St.
Torrance, CA 90503

Name
Address
City State Zip

which is a two-element list. The first element of each clause is its "condition," the second element is its "result." The clauses' conditions are checked out by the Lisp genie one by one, in order; as soon as the genie finds a clause whose condition is "true" (meaning that the condition returns anything other than nil) it begins calculating that clause's *result,* whose value gets returned as the value of the entire cond statement. None of the further clauses is even so much as glanced at! This may sound more complex than it ought to. The real idea is no more complex than saying that the genie looks for the first condition that is satisfied, then returns the corresponding result.

Often one wants to have a catchall clause at the end whose condition is *sure* to be satisfied, so that if all other conditions fail, at least this one will be true and the accompanying result, rather than nil, will be returned. It is as easy as pie to make a condition whose value is non-nil; just choose it to be "t", for instance, as in the following:

```
—> (cond ((eq x 1) 'land)
          ((eq x 2) 'sea)
          (t 'air))
```

Depending on what the value of x is, we shall get either "land", "sea" or "air" as the value of this cond, but we shall never get nil. Here are a few sample cond statements for you to play genie to:

```
—> (cond ((eq (eval pi) pie)
          (eval (snoc pie pi)))
          (t (eval (snoc (rac pi) pi))))
—> (cond ((eq 2 2) 2) ((eq 3 3) 3))
—> (cond (nil 'no-no-no)
          ((eq '(car nil) '(cdr nil))
           'hmmm)
          (t 'yes-yes-yes))
```

The answers are 8, 2 and "yes-yes-yes". Did you notice that "(car nil)" and (cdr nil)" were quoted?

I shall close this part of the column by displaying a family of function definitions so obvious in their pattern that you would think the Lisp genie would simply "get the hang of it" after seeing the first few. Unfortunately Lisp genies are frustratingly dense (or at least they play at being dense), and they will not jump to any conclusion unless it has been spelled out explicitly for them. Look first at the family:

```
—> (def square
     (lambda (k) (times k k)))
—> (def cube
     (lambda (k) (times k (square k))))
—> (def 4th-power
     (lambda (k) (times k (cube k))))
—> (def 5th-power
     (lambda (k)
      (times k (4th-power k))))
—> (def 6th-power
     (lambda (k)
      (times k (5th-power k))))
 :
 :
 :
```

My question for you is this: Can you invent a definition for a *two*-parameter function that subsumes this entire infinite family in one fell swoop? (Please do not write to give me your thoughts; I know the answer and will provide it next time.) More concretely, the question is: How would one go about defining a two-parameter function called "power" such that, for instance, "(power 9 3)" yields 729 on being evaluated and "(power 7 4)" yields 2401? I have given you in this column all the necessary tools to do it, provided you exercise some ingenuity.

I thought I would end with a newsbreak about a freshly discovered beast: the Glazunkian porpuquine, so called because it is found only on the island of Glazunkia (claimed by Upper Bitbo, although it is just off the coast of Burronymede). And what is a porpuquine, you ask? Why, it is a strange breed of porcupine whose quills—of which, for some reason, there are always exactly nine (in Outer Glazunkia) or seven (in Inner Glazunkia)—are smaller porpuquines. Oho! This would certainly seem to be an infinite regress! But no. It is just that I forgot to mention that there is a smallest size of porpuquine: the zero-inch type, which, amazingly enough, is totally bald of quills. Thus quite luckily (or perhaps unluckily, depending on your point of view) that puts a stop to the threatened infinite regress.

Students of zoology might be interested to learn that the quills on five-inch porpuquines are always four-inch porpuquines, and so on down the line. And students of anthropology might be equally intrigued to know that the residents of Glazunkia (both Outer and Inner) utilize the nose (yes, the nose) of the zero-inch porpuquine as a unit of barter—an odd thing to our minds, but then who are you and I to question the ancient ways of the Outer and Inner Glazunkians? Hence since a largish porpuquine—say a three- or a four-incher—has many, many such tiny noses, it is a most valuable commodity. The value of a porpuquine is sometimes referred to as its "buying power," or just "power" for short. For instance, a two-incher found in Inner Glazunkia is almost twice as powerful as a two-incher found in Outer Glazunkia. Or did I get it backward?

Anyway, why am I telling you all this? Oh, I just thought you would like to hear about it. Anyway, who knows? You just might wind up visiting Glazunkia (Inner or Outer) one of these fine days. And then all of this could come in mighty handy.

© 1983 SCIENTIFIC AMERICAN, INC

CALCULATOR MATH?

Your brain works 3 times as fast as your fingers can write! That's where the Calculator comes in — you can go three times as fast with your machine doing the dirty work.

It's easier and more interesting. And you'll understand it better when it's fun. (If you haven't yet tried out a modern scientific-statistical calculator, you're in for a pleasant surprise — see below!)

WHY MATH? To get a promotion or a new job. Just to hang on to that job you've got. To quicken your mind. To understand this world you live in. To impress people. Just for some fun you'll feel good about. (Read those letters below to see how it has worked for others.)

Do it at home... You can start right away and go as fast as you like! Put it aside while you're busy with something else — in short, *set your own pace!*

Lots of us like to avoid being compared with the other members of a class — we like to set our own standards. Too, we like to save the time and expense of going to a school or seminar. (Best of all, it's a lot cheaper to study at home where the rent is already paid — *read on!*)

Let's be honest. You wouldn't expect to learn to play baseball or chess by watching an expert do it. Math is the same — you'll need to do some work. BUT — with my calculator method, you'll see it's a whole new ball game!

THIS METHOD IS FRIENDLY! You can use it on your own. You follow interesting examples step-by-step with calculator keystrokes — instead of algebra.

Let me introduce myself — I'm Dr. George McCarty and I teach at the University of California. Tens of thousands of people like yourself have already used my guidebook to teach themselves calculus. Recently I have teamed up with another professor. We put those same tested and proven techniques to work for you in a new *Statistics Guidebook.*

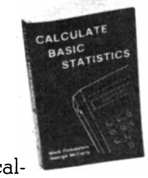
Even if you've had trouble with math, this method will work for you. Choose the one you want to learn, calculus or statistics (or both) and start *today!*

As pennywise Ben Franklin said, "An investment in knowledge pays the best dividends."

CHOOSE STATISTICS

It's used everywhere — to set insurance rates and to assess a drug's effectiveness, to detect a crooked roulette wheel and to measure lightbulb lifetimes, to audit the books and to find your position at sea.

People have always known that their experience was not a perfect guide to the future. Statistics shows us that uncertainty is not only inevitable but it itself is measurable.



You'll find out about this in my guidebook. You'll read informal explanations with detailed calculations for topics like "confidence intervals" and "contingency tables."

YOU'LL BE DOING Inference and Hypothesis Testing, the Chi-Square Distribution, ANOVA, Regression and Correlation, Non-Parametric Methods — they're all here.

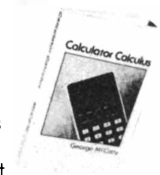
SUPPOSE YOU HAVE ALREADY STUDIED STATISTICS. You'll use Quick-Reference Guides to go straight to a topic you want to refresh yourself on — then get *right-now* answers in your own field, business, biology, science, psychology, education.

The Guidebook is guaranteed to work with any calculator or computer. If you would like to start out with a fancy new machine, order my Statistics Kit which includes the guidebook and a TI-35 calculator which finds means and standard deviations at the touch of a button. It has an LCD display, 1000 hour batteries — the works.

PHONE MC or VISA ORDERS TO
(714) 831-2637
Ask for our Catalog of books and calculators.

CHOOSE CALCULUS

The math of science and engineering — satellite orbits and compound interest, bacterial growth and radioactive decay, marginal profits and reaction rates.



You'll find out about these concepts when you use the Calculus Kit. This Kit contains my guidebook and a TI-35 calculator which finds log and trig values at the touch of a button. (It's modern, too, with LCD display and 1000 hour batteries.)

In the guidebook you'll learn both differential calculus and integral calculus — *even differential equations!*

NEED TO EVALUATE functions, areas, volumes — solve equations — use curves, trig, polar coordinates — find limits for sequences and series? It's all here!

If you're in the biological, social or physical sciences, you'll be doing Bessel functions, carbon dating, Gompertz' growth curves, half-life, future value, marginal costs, motion, cooling, probability, pressure — and plenty more.

Important numerical algorithms are here, too: rational and Padé approximation, bracketing, continued fractions, Euler's method, Heun's method, iteration functions, Newton's method, predictor-corrector, successive substitutions, Simpson's method and synthetic division.

If you already have a scientific calculator, you can get just the *Calculus Guidebook.*

George McCarty

P.S. — YOU *CAN* DO IT! I'm certain I can show you this easy way to success with math, so ***I guarantee it to you unconditionally*** — if you are not satisfied just send it back for a prompt refund!

LOOK AT WHAT THEY SAY...

• *"...excellent diagrams and clear explanations."* —POPULAR SCIENCE

•Tektronix Engineer Bill Templeton says, *"The best, most clearly written book I have seen for improving your math skills."*

• *"...straight-to-the-point explanations..."* —HP Key Notes

•Professor H.I. Freedman of the U. of Alberta, states: *"There can be no question as to the usefulness of this book...lots of exercises...very clearly written...easy reading."*

•S.C. McCluney, Jr. of Philadelphia writes: *"With your book and a calculator the whole idea becomes clear in a monent, and is a MOST REFRESHING EXPERIENCE."*

TAX DEDUCTIBLE FOR PROFESSIONALS

YES, I want to learn math **fast** and **easy** with your *new method.* If I am not more than satisfied I can return it for a quick refund. RUSH me the guidebook or complete kit I select below (kit contains the guidebook and a modern TI-35 calculator).

☐ Send **FREE** Catalog of books & calculators

STATISTICS GUIDEBOOK		**KIT**
How many?	☐ $14.95	☐ $39.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	.90	2.40
TOTAL	______	______

CALCULUS GUIDEBOOK		**KIT**
How many?	☐ $19.95	☐ $44.95
Handling/Shipping	1.00	2.00
Tax (within Calif.)	1.20	2.70
TOTAL	______	______

For AIR SHIPMENT add $3.00 per item
Phone (714)831-2637 (MC or VISA only)

Enclosed is ☐ Check, ☐ M.O.; or charge my MC/VISA... exp. ______

Card # ______

Name ______

Address ______

City ______

State, Zip ______

***OUTSIDE USA,* surface shipping is $2.00 per book or $5.00 per kit. Air shipment is $5.00 per book or $10.00 per kit. U.S. FUNDS PLEASE.**

EduCALC Publications, Dept. A1
27953 Cabot Road
South Laguna, CA 92677

©1982

© 1983 SCIENTIFIC AMERICAN, INC

BOOKS

The real Albert Einstein, human clocks, amazing ammonites and megalith mania

by Philip Morrison

SUBTLE IS THE LORD . . . : THE SCIENCE AND THE LIFE OF ALBERT EINSTEIN, by Abraham Pais. Oxford University Press ($25). During Einstein's first visit to the U.S. in 1921 word reached him in Princeton that out at Mount Wilson some new measurements were showing a nonzero ether drift. His comment was: "Raffiniert ist der Herr Gott, aber boshaft ist er nicht." Subtle is the Lord, but malicious he is not. The remark is chiseled into stone at Princeton University still; Einstein interpreted his own saying when he consented to the inscription in 1930: "Nature hides its secret because of its essential loftiness, but not by means of ruse." Nor did the drift persist. Our current view of the vacuum, a complex virtual plenum with effects on particle energies and on cosmic expansion, "has nothing to do with the aether . . . , which is gone for good."

Einstein was an artist in exposition, as he himself said of Newton. His gift for vivid and apt expression was great; he was a "master of nuances" and a perceptive appraiser of the men and women of his time. "His portraits serve as the best foil for the opinion that he was a naive man." At ease only in German, he wrote every one of his scientific papers in that language. In this book we are given a sympathetic but clear-eyed view of his life and work and the physics he pursued from youth to death, examined mainly in his own plentiful papers, supplemented by a remarkably wide list of unusual sources.

Those include an unpublished "contribution to her brother's biography" written by his younger sister Maja in 1924, recent books from Germany and Switzerland, even Japan, an unpublished manuscript in the Morgan Library, reference letters from files in Zurich, Munich and Berlin, and of course the bulging Einstein Archives. The best of trails into the archives were opened by Einstein's secretary Helen Dukas. ("Dear Helen, thank you; it was wonderful." Miss Dukas died early in 1982, the most experienced of guides and witnesses to that life.) It is a rare treatment of a physicist dead for nearly three decades that can include entirely apropos references to current papers in *Physical Review Letters;* both the man studied and the man who studies him must be counted as unusual.

Professor Pais is a senior theorist, distinguished in his own right, who has in the past few years disclosed a flair for the history of ideas in physics. To this book he brings that power, and much more. He was lucky enough to walk and talk with Einstein at the Institute for Advanced Study in Princeton every few weeks for all the postwar years until a few months before the end. Vignettes of the old Einstein with the young quantum field theorist fresh from Bohr's institute in Copenhagen are poignant: once "he suddenly stopped, turned to me, and asked me if I really believed that the moon exists only if I look at it."

The last time, Pais was about to leave Princeton for a term's leave. Einstein, 75 years old, had been ill and away from his office for some weeks. It was natural to pay him a brief visit at home. "He was working. He put his pad aside at once.... We spent a pleasant half hour or so.... I told him I should not stay any longer.... I turned around as I opened the door. I saw him in his chair, his pad back on his lap, a pencil in his hand.... He was back at work."

It is the strength of this fine book that it displays and assesses that beloved work, in a thorough sampling of the *ix-erei* that is the terse language of the theorist. Always it appears in a prose frame of context and meaning; readers with some knowledge of this form of expression will be enriched by the brief treatments. Those who cannot grasp the *x*'s and their subscripts are given much aid in learning the arguments and in sharing the life story that surrounds the equations. Among the many books on Einstein this one more than any centers on the work, its springs and its meanders, and of what came of it in the end in other hands—or has still to come.

The events, public and private, of Einstein's long life are by now canonical, and yet both in facts and in understandings a good deal here gleams new and rings true. The baby was born with a strange skull shape (said grandmother, "Viel zu dick": Much too thick) and was slow in learning to speak. The grade school student brought home a fine report card; his mother wrote, her Albert seven and a half, "He was again number one, his report card was brilliant." Throughout the years in Luitpold Gymnasium the boy earned either the highest or the next-highest mark both in mathematics and in Latin. The poor-student myth is unfounded, although the independent Einstein himself recounted that a high school teacher objected to his presence. "You sit there in the back row and smile, and that violates the feeling of respect which a teacher needs from his class." Einstein was fascinated by experiments, and from his first Zurich years he planned big ones, for which his professor was not enthusiastic.

Einstein's experiments and patents, sought with able partners, often on their initiative, make by no means a petty list, from the first successful demonstration that magnetized currents implied rotation to the electromagnetic pump for home refrigerators, now found in nuclear reactors. Twice married, first to a fellow student at Zurich, later to a devoted and motherly cousin, he was no model husband and father, even in his own eyes. When his friend Besso died, old Einstein wrote: "What I most admired in him as a human being is the fact that he managed to live for many years not only in peace but also in lasting harmony with a woman—an undertaking in which I twice failed rather disgracefully."

It is fascinating to follow the progression of his researches. Pais remarks that his early career reminds one of Mozart: effortless, fresh, a flood of successes. Then the chronicle rather passes into the mode of a Beethoven, persisting, learning, maturing at last to peerless mastery: general relativity and its cosmology stated with clarity by 1918. If one inference is to be drawn from Pais's incisive study, it is that the lonely excursion Einstein took, a man apart, into the unified field theory after 1925 to his last years was not chiefly a search for a single theory of forces. That is a current quest of physics. His was a grand effort to conjure the quantum graininess out of the flowing field world, to make particles and probability necessary somehow by the demands of invariance on the nonlinear but strictly causal fields gravity required. We do not yet know how that drama will end; the unification of today, brilliant as it is, remains incomplete. Theorists do not at all share Einstein's deepest goal of restoring Newtonian causality to a world reliably ruled to a good many decimal places by the subtle prescriptions of quantum fields.

It is tempting to imagine for each of

© 1983 SCIENTIFIC AMERICAN, INC

FOR THE WET


FOR THE WHITE


FOR THE WINDING


FOR THE WILD


FOR THE WARM


FOR THE WEEKEND


FOR THE WALLET


W is for Eagle Wagon. And for the wherever. Because Eagle is the only automobile in the world that lets you go from 2-wheel drive to full-time 4-wheel drive at the flick of a switch. 2-wheel drive for high mileage on the highway. 4-wheel drive for worry-free traction and security when the roads weave, the weather worsens, or when you just want to wander where there is no road.

And Eagle's well-appointed interior makes the going all the more wonderful. Why not see the 1983 Eagle Wagon today. Wherever you're going, whatever the weather, it'll fit you to a **W.**

SAFETY IS A SNAP

The 2-wheel/4-wheel drive EAGLE

FROM AMERICAN MOTORS

*Optional 5-speed stick. Use these figures for comparison. Your results may differ due to driving speed, weather conditions and trip length. Highway and California figures lower.

© 1983 SCIENTIFIC AMERICAN, INC

THE FIRST PERSONAL COMPUTER

When you sit down with the Professional from Digital Equipment Corporation, you discover something very quickly: this is not just another personal computer.

It is a thoroughly professional system designed for the sole purpose of working in a professional, business environment. Which is why it has capabilities superior to other personal computers you might have seen before.

EASIER TO WORK WITH.

Through Computer-Based Instruction, you start using the Professional™ computer the same day you get it. Further, with software that's certified by Digital, the Professional computer gives you one consistent menu for all your applications. A first in personal computers. There's also a HELP command to answer questions about operation. And if something goes wrong with the system, it will actually draw you a picture of the problem and tell you what to do about it.

BETTER COMMUNICATIONS.

The pictures and graphics you see on the Professional are four times sharper than conventional personal computers. The text can be displayed in bold, underlined, double-sized, or practically any other form that will make communications clear. And the unique Telephone Management System opens the way to

*Dow Jones News/Retrieval is a registered trademark of Dow Jones & Co., Inc. Professional is a registered trademark of Digital Equipment Corporation.

© 1983 SCIENTIFIC AMERICAN, INC

AS PROFESSIONAL AS YOU ARE.

voice communications in personal computers for the first time.

MORE INFORMATION.

The Professional computer can hold a staggering 5.8 million bytes of information right at your desk. When you need information from other sources, you can be networked with other Professionals, larger systems from Digital, IBM mainframes, and data services like Dow Jones News/Retrieval.*

And the Professional computer can do *more* with that information. Its "multitasking" capability means it can work on one or more jobs while you work on yet another.

MORE COMPUTER.

Where other personal computers have reached the limit of what they can do, the Professional computer is only beginning. It has true minicomputer power, and thus offers extraordinary growth potential. It has many capabilities that are only starting to be explored.

We'd like you to meet the Professional computer. In person.

To find out where you can see Digital's Professional™ 300 Series computer, call 1-800-DIGITAL. Or write: Digital Equipment Corporation, 129 Parker Street, Media Response Manager PK32/M94, Maynard, MA 01754.

digital

MORE PERSONAL.
MORE COMPUTER.

© 1983 SCIENTIFIC AMERICAN, INC

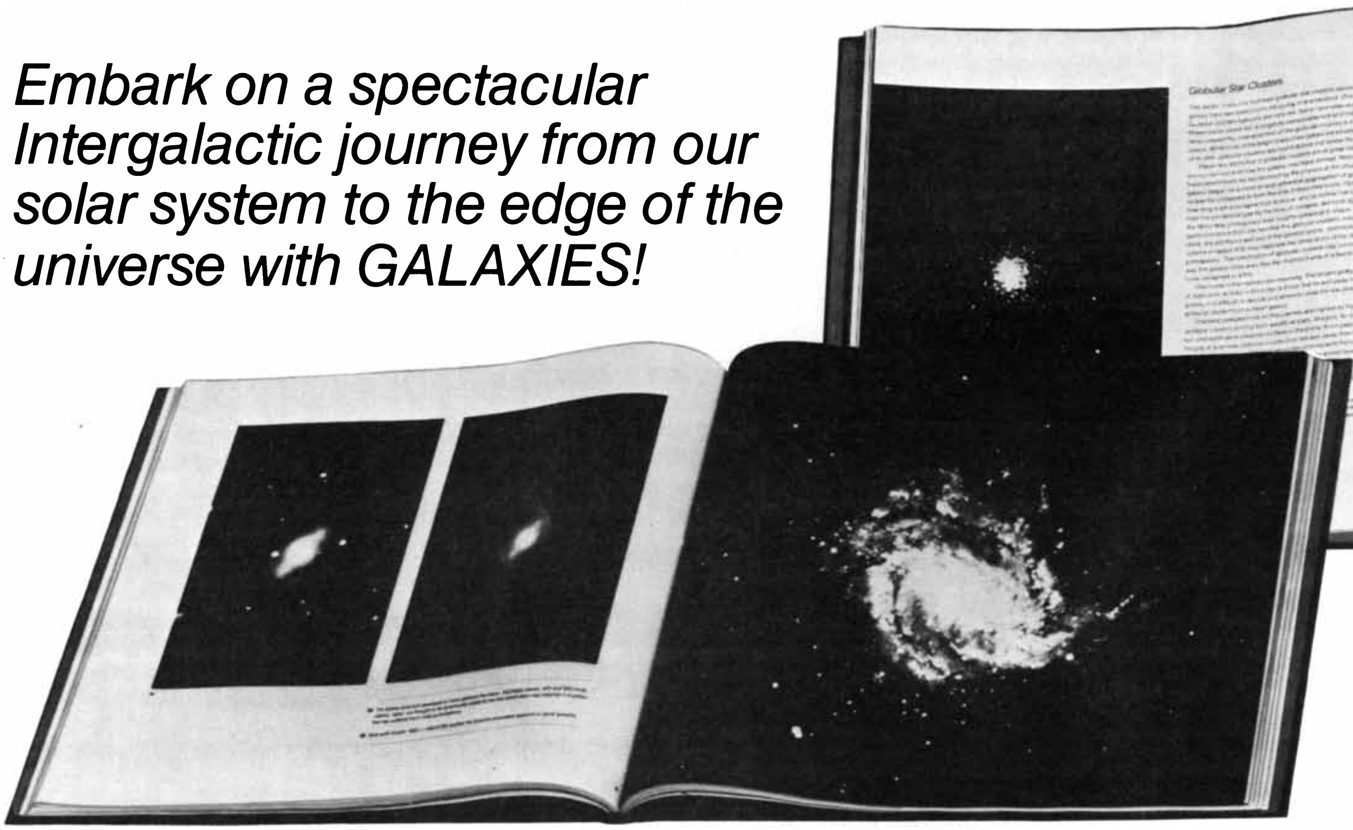
Now, you are invited to take this magnificent reference and

GALAXIES for only

as your introductory selection to the ASTRONOMY

You simply agree to buy 3 more books—at handsome discounts—over the next 12 months.

Embark on a spectacular Intergalactic journey from our solar system to the edge of the universe with GALAXIES!



Shooting stars... novas... star clusters... black holes and much more await you in the inspiring and beautiful pages of GALAXIES.

With this exciting offer, you can take GALAXIES, a $75.00 value, for only $2.95 as your remarkable introduction to the Astronomy Book Club. Handsomely bound in the toughest, most durable materials and printed on high quality paper, your impressive volume has been constructed to last a lifetime.

And GALAXIES is visually stunning. It includes a virtual treasury of outstanding photography—each a work of art in itself—all collected by astronomers from observatories throughout the world.

In GALAXIES, 145 spectacular photographs (39 in blazing full color) take you on an intergalactic journey beyond the realm of the ordinary... a fantastic journey to the outer reaches of the universe.

GALAXIES is more, too. It includes 20 magnificent line drawings by noted illustrator Sarah Landry and a huge 13" x 56" gatefold photographic mosaic of the Milky Way.

GALAXIES is the most beautiful and significant book ever produced on the study of the worlds beyond our own... a book you will be justifiably proud to own.

In fact, Sky and Telescope magazine had this to say about GALAXIES:

"The artwork that accompanies the text and photographs is superb. The paper and binding are of high quality... a marvelous gift for someone who wants to know about the universe."

© 1983 SCIENTIFIC AMERICAN, INC

adventure volume...

$2.95

BOOK CLUB.

ORIGINAL $75.00
HARDCOVER EDITION

GALAXIES

the most evocatively beautiful volume on the nature of deep-sky splendors ever published.

- An authoritative reference work... and an exciting literary adventure
- Spans an extraordinary 13" x 14½"... 182 pages of vivid text and clear, concise captions
- Tough, durable binding and lustrous, coated stock, smudge-proof paper
- 106 dramatic duotone photos... 39 full color photos—many full page size
- Original hardcover edition
- Written by award-winning author, Timothy Ferris
- 20 line drawings by noted illustrator, Sarah Landry
- Includes extensive bibliography, glossary and index
- A $75.00 value—yours for only $2.95 as an introduction to the Astronomy Book Club

"...a very good candidate for the most beautiful book in the world... I was absolutely overwhelmed by its excellence —text *as well as* photographs."

—Isaac Asimov

According to the New York Times, "The book is at once a linear, visual, historical, futurist, cultural and philosophical journey to forever."

And now you may have the original hardcover edition of GALAXIES for the exceptional bargain price of only $2.95 (you save over $72.00 off the publisher's price) as your introduction to the additional delights that await you as a member of the Astronomy Book Club.

Astronomy Book Club is a very special club—one that is operated exclusively for astronomy enthusiasts. Recent offerings include titles by such eminent authorities as Carl Sagan, Robert Jastrow, John Gribbin and Paolo Maffei. The Club is devoted to making the finest astronomy books available conveniently and inexpensively!

4 Good Reasons to Join

1. The Finest Books. Of the hundreds and hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, **never** economy editions.

2. Big Savings. In addition to getting GALAXIES for only $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3–4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to:
Astronomy Book Club, Dept. L-AW4, Riverside, N.J. 08075
to obtain membership information and an application.

© 1983 SCIENTIFIC AMERICAN, INC

SCIENCE/SCOPE

High efficiency solar cells now being developed could cut substantially the weight and area of some satellite solar panels. These cells, made with gallium arsenide by a special liquid-phase epitaxial growth process developed by Hughes Aircraft Company scientists, have shown efficiencies higher than 19% in converting sunlight into electricity. This compares with efficiencies between 15% and 16% for the best available silicon solar cells. In addition, the gallium arsenide cells can operate at higher temperatures and can tolerate more high-energy proton irradiation in space than silicon solar cells.

A new wireless entertainment system will use infrared light to carry music and movie soundtracks to passengers on commercial and corporate jet aircraft. The system, under development at Hughes, transmits a digital infrared signal that is received and decoded by a passenger's headset. The signal is completely harmless to people and won't interfere with other aircraft equipment. The infrared system transmits 16 channels. It would weigh less than half of a conventional wired system and would cost a third less. Since there are no wires with this system, it is particularly suited for aircraft with changeable seating.

Communications among Brazil's cities and remote villages will be improved when the first domestic communications satellite system in Latin America goes into operation in 1985. The primary function of the two Brazilian satellites will be to improve telephone and television services. The spacecraft are based on a Hughes model that is the world's most widely purchased commercial communications satellite. Now 27 are under contract. Hughes will serve as major subcontractor to Spar Aerospace of Canada. Embratel, Brazil's government-owned telecommunications company, will buy and operate the satellite system.

Now in its 17th year, Science/Scope remains among the best-read corporate advertising campaigns in history. The campaign was created to explain what Hughes does and how. About 160 different ads are produced each year, with a mix of over 175 paragraphs for use in over 80 publications serving over 60 nations. It is translated into Arabic, Chinese, French, German, Italian, Japanese, Korean, Portuguese, and Spanish -- 10 languages in all, including English. It consistently scores in the top 5% to 10% in readership surveys.

Career growth opportunities exist at all levels at Hughes Support Systems for a variety of engineers qualified by degree or extensive work experience. They include systems engineers, applications engineers, software and hardware design engineers for major simulation and test equipment programs, and automatic test equipment engineers. Also, field engineering posts throughout the U.S. and the world offer travel, autonomy, and responsibility. Call collect (213) 513-5238. Or send your resume to Professional Employment, Dept. SE, Hughes Aircraft Company, P.O. Box 9399, Long Beach, CA 90810-0463. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

(213) 670-1515 EXTENSION 5964

© 1983 SCIENTIFIC AMERICAN, INC

Einstein's individual victories the effect that removing that single most productive theorist of this century might have had on the history of physics. His first fine work of all, on the reality of molecules, was foreshadowed, in part repeated independently, by Willard Gibbs and Marian Smoluchowski. The theory of special relativity lay a little beyond the reach of Henri Poincaré, who had recognized both the technical transformations needed and the requirement for a rethinking of simultaneity but could not fuse them. Would another young physicist have come along to do that? It is a good bet either way.

The theory of gravitation required still more insight. There were signs, however, that among the mathematicians of Göttingen around David Hilbert that theory too might have emerged. (Given, to be sure, the hints of Einstein's earlier approaches, Hilbert did write the equations correctly quite on his own.) Perhaps it was the photon, the particlelike bearer of Maxwellian waves, that most needed Einstein himself. From his 1905 photoelectric effect straight through a series of permanently valuable and brilliant inferences about radiation and matter, Einstein remained loyal to the reality of the photon, which he perhaps more than anyone established. Even Bohr once abandoned the photon, when in 1924 he despaired of individual atomic acts of conservation of energy and momentum.

Einstein remained steadfast; much earlier, in 1909, he had written: "It is my opinion that the next phase... will bring us a theory of light that can be interpreted as a kind of fusion of the wave and the emission theory... [which] are not to be considered as mutually incompatible." When that theory appeared in full in about 1930, Einstein found it repugnant. He maintained it was a deceit. No final truth was in it, for all its predictive success. Pais senses a reason for this stubborn philosophical set on the part of the very man who argued the world into accepting not only a new structure for space and time but also a new entity at once a wave and a particle. Einstein almost said it himself: "To the discoverer the creations of his imagination appear so necessary and so natural that he is apt to treat them not as the creations of his thoughts but as given reality." Twice he had found reality in depth by insisting on the constraints imposed by certain freely chosen simplicities. A third time has not come.

"Einstein was the freest man I have known." He was "the master of his own destiny," free to ask, fearless of time and death. No tragedy lies in his attitude toward the quantum theory or in his unsuccessful search for the unified field. So writes Pais, physicist and friend. The tragedy of Einstein is not personal but universally shared; we are linked through the death camps, the shadow of the distant mushroom cloud, the endemic racist plagues. Einstein free is symbol of this ironic century of ours, its sands not yet run out.

The Clocks That Time Us: Physiology of the Circadian Timing System, by Martin C. Moore-Ede, Frank M. Sulzman and Charles A. Fuller. Harvard University Press ($25). The storyteller vividly reminded us that to "the cat that walks by itself" all places are alike. Here a striking factual diagram documents even more: for that cat all *times* are alike as well. A month of plotted activity hour by hour shows little temporal pattern. Bursts of activity and rest follow each other irregularly for a couple of hours, giving little sign of any self-sustained longer rhythm, independent of even a shifting cycle of light and dark. Mice and men are not like cats. Both species show strong activity rhythms. Living without time cues in a carefully isolated laboratory, house mice and human beings both show a regular free-running circadian pattern, close to but not matching the 24-hour rhythm of earth spin. (Cats and dogs and bobcats are exceptional.) Most mammals—and living forms in general—have a marked built-in circadian rhythm, with a period that differs among individuals and from species to species. House mice tend to run fast: their stable free-running period centers near 23.5 hours. Human beings run a little slow: their internal-activity clocks average about 25 hours around the dial.

This carefully documented book is a comprehensive yet nontechnical introduction to a fascinating new system in human physiology: the internal timing system, tightly bound to sleeping and waking, plainly part of the still unexplored richness of the unconscious brain-body connection. The volume develops the ideas clearly and logically, with a minimum of mathematics and many graphs from the big body of experimental results. Its clear aim is to draw everyday and medical implications from the human case of this subtle outcome of life's rotating environment.

We live by the clock, either by manmade dials or by the older one of dawn and dusk. Many properties of the body, not only our most conspicuous rhythm of sleep and wakefulness, follow the day closely. Both internal processes and the response of the system to all kinds of outside effects show strong diurnal patterns. In the laboratory mouse, for example, clear cyclical effects of the time of day have been demonstrated among 50 or 60 distinct physiological variables.

One plot shows a sample of human daily patterns, varying sharply in phase and form among the quantities measured. Sleep comes by night, and with it comes the maximum blood content of growth hormone but the minimum of internal temperature and urinary potassium excretion. Cortisol levels are at a minimum in the blood as sleep encroaches but wax to a maximum shortly before waking. Evidently the conscientious intern who watches all night over her patient's measured states must take account of such natural waves of activity. These patterns mature; infants sleep and wake as haphazardly as cats for the first month or so (on a demand feeding schedule). Then a marked rhythm sets in that runs a little slow of the dawn-dusk cycle, until after four months of life the external world of day-night and of human activity in general has entrained and synchronized the biddable inner clock.

It is the controlled study of that entrainment by external change that best reveals novelty amidst all these harmonies. The outside pacemaker can cue time commands at will up to a point, usually by a few hours, plus or minus, away from the inbuilt period. The cue need only be small; a single brief light pulse once every 24 hours is sufficient to synchronize the squirrel monkey, our little diurnal primate cousin. The spontaneous loss of that synchronization, once all external pacemaking is removed, is the most remarkable recent result, particularly in human beings.

One typical subject showed a steady 24-hour day-night cycle, still entrained to the stars, for the first five days of total laboratory isolation in time. Then an inner clock took over that kept a 25-hour day. That clock was steady, and both the rest-activity cycle and the body-temperature cycle followed it closely. At about five weeks something inside changed. Thereafter the body-temperature cycle shortened a little, although it was still a bit slow on the 24-hour day. The sleep cycle, however, lengthened markedly. It kept its own time, with a regular if intricate structure of long and short sleeps, on an inner day that was now more than 30 hours long.

We must have two chief timekeepers, not one. Their mutual interaction is well described by a simple mathematical model. One clock controls the skin temperature and many related rhythms, the other controls the sleep-wake cycle and the many other functions that derive from sleep. There is even strong evidence, not yet certain in human beings, that serves to place the two clocks. One of them, an organ we share with most other mammals, is seated in a pair of millimeter-size clusters of special neurons in the hypothalamus deep within the brain. A long history of surgical

© 1983 SCIENTIFIC AMERICAN, INC

excision in monkeys has made this site pretty sure. Clinicians have for decades realized that brain tumors near the corresponding structures in human beings can bring profound sleep-wake disorders. More remarkable still, in monkeys (not, of course, in human patients) quick autopsy after the injection of a radioactive tracer into the retina has disclosed an active tract of neural fibers leading directly from the retina to that clock center. We are here close to an understanding of how we wake by dawn.

The other timekeeper is less well known; there is evidence that it too lies in the hypothalamus, near but distinct from the other one. The retinal-linked clock times the skin temperature, the growth hormone of the blood and our dreamless form of sleep. It runs freely at a very slow rate, although it normally senses night and day. The 25-hour clock controls the rapid-eye-movement sleep of dream, potassium excretion in the urine and the generation of heat deep within the body. Its rhythm carries over to the other slower oscillator, which in return feeds back strong signals of its own informed vision of light and dark. Together they fix our usual coordinated 24-hour rhythm. Of course, this account is much elaborated in the full text by these three experienced research workers. They—a physiologist, a biologist and a neuroscientist—have worked together fruitfully for seven years on these inward beats of night and day.

We live now, as we have ever since the discovery of fire, increasingly cut off from earth spin. Daylight-saving time, shift work and swift jet travel across time zones are clear examples of acute stress on the ancient timing system. Consider too the inhabitants of the Arctic who have chronically endured winter months without the unfailing dawn cues of the ancestral tropical latitudes of our hominid kind. There are gloomy tales: once an airline pilot fell asleep in the very act of touchdown, 200 feet above the runway. (The copilot was awake.) In 1978 a jet overflew Los Angeles on autopilot by 100 miles; the entire flight crew was asleep. The Three Mile Island power reactor was operated by a crew that rotated shifts around the clock on a weekly basis, conventional enough but much too fast for good adjustment. The Navy judges that eight-hour shifts are too arduous for the crew of a missile submarine. Since the boat can hold only three shifts of men, however, it operates on a stressful 18-hour day, quite unphysiologically.

The trick for time-estranged travelers is probably to avoid shifting phase at all when the visit is brief enough. If a shift is necessary, experience and the beginnings of understanding favor a delay of phase over an advance. It is easier to entrain our slow double inner clock in days that are longer by a couple of hours than in days that are shorter; a westbound time gain is more quickly compensated than an eastbound loss; certain out-of-phase insomnias have responded to a careful regimen of 27-hour days. "The study of the circadian timing system is thus at a fascinating stage"; here is human physiology entering a completely new dimension.

THE AMMONITES: THEIR LIFE AND THEIR WORLD, by Ulrich Lehmann. Cambridge University Press ($19.95). THE AUDUBON SOCIETY FIELD GUIDE TO NORTH AMERICAN FOSSILS, by Ida Thompson. Alfred A. Knopf, Inc. ($12.50). Three or four hundred meters deep in the dark sea from Fiji to the Philippines the pearly chambered nautilus hunts by night. Slowly swimming backward, the creature's propelling jet spurts from its head past 90-odd little tentacle arms. The nautilus is an antique like the coelacanth, not much changed over an ancestry going back proudly to the Ordovician. Before ever fish swam these animals and their earlier nautilid kin ruled the seas, their chambered shells often not yet coiled.

The upland limestones of southwestern Germany yield up many fossil coils that resemble the beautiful planar spiral chambers of the nautilus. So widespread is one early form that it has been proposed as the German heraldic figure. Sometimes the calcareous shell material is itself preserved, with its pearly sheen and elegant form plainly "among the most beautiful fossils in existence." Commoner are clear interior molds, all complex detail visible.

These remains have been collected and studied for a long time in the classic localities. There is such a fossil engraved by a Paleolithic artist, from a cave near Ulm. Many thousands of forms are now recorded worldwide, from submillimeter size to nearly two meters across, spanning a time of more than 100 million years, until they end along with the dinosaurs at the Cretaceous-Tertiary time boundary. For a century the intricate morphology of these abundant remains supplied in all their variety the basis of a stratigraphic dating procedure, but the biology of the animals themselves was neglected in the "somewhat schematic concept of index fossils." Indeed, the living nautilus is itself little known, only recently observed in captivity.

Professor Lehmann's book, a broad and well-illustrated introduction to the life that secreted the spirals, is the work of a paleobiologist who has made a major contribution to the newer understanding. The intricacy of the shapes that must be understood in three dimensions and the general strangeness of molluscan life in the sea do not allow easy grasp by the general reader, even though no previous special study is assumed. The sense of excitement in the new inferences drawn from recent finds nonetheless comes across sharply; the brief book will repay the study it needs to anyone who enjoys the paleontologist's hunt for meaning.

The key to the life of the ammonite animal was the finding of its jaws. For a long time it was thought ammonites had none, no analogue to the strong, sharp parrot beak of the nautilus, made of resistant chitin-reinforced calcite. Modern cephalopod jaws are abundant on today's ocean floor, here and there even densely covering the bottom. Only in 1965, however, were the jaws of an ammonite clearly recognized, in a well-preserved specimen collected by students at the University of Montevideo. The find was unique in its preservation; the phosphate nodule had only to be cracked open to reveal to the knowing paleontologist (Professor Closs of Puerto Alegre in Brazil had his Ph.D. from Tübingen for work on certain curious double calcite structures found isolated or near ammonites) the entire relevant apparatus of a mollusk, a rasplike radula between two slightly flattened jaws.

Within 10 years the riddle was pretty well answered, although there remains a question or two. Professor Lehmann, the author of *The Ammonites,* ground away the calcite tenth-millimeter by tenth-millimeter to follow and map the darker contours within the white material that filled the body chamber of one suitable ammonite specimen. His first hint was a structure closing the aperture of the body chamber. Within the familiar but long uninterpretable structure there was at last disclosed a second pair of fitted lobes; it was unmistakably a pair of jaws made of a "horny-coaly substance." A three-dimensional model was patiently prepared.

Teeth and jaws offered new meaning. The ammonite's blunt-edged lower jaw was not a cutting tool or even a sharp beak. The jaw muscles were not strong. The creature held its jaw like a shovel just above the sea bottom as it floated slowly along, feeding on bottom organisms. There is much corroboration by now; Lehmann found "a small insignificant-looking black fragment" in a London collection that yielded a few half-inch spirals. In them pairs of jaws were to be found; in one even a stomach holding a last meal of small foraminifera and ostracods.

Those ornate spirals and the many aberrant forms that look more like saxophones than like French horns are now newly plausible; shell form was not significant ecologically. Water resistance

© 1983 SCIENTIFIC AMERICAN, INC

Home sweet Honda.

The 1983 Accord Hatchback offers all the comforts of home. Plus a few more of its own.

The Accord's spacious interior and plush trim are wonderfully inviting. And once on the road, its excellent handling and roadability are proven at every turn.

Quality Honda engineering features include front-wheel drive, 4-wheel independent suspension and power-assisted self-adjusting ventilated front disc brakes. Steel-belted radial tires, a maintenance reminder and electronic warning system all come standard.

A new 4-speed automatic transmission with variable-assist power steering is also available. It not only makes driving simpler, but offers improved automatic fuel economy as well.*

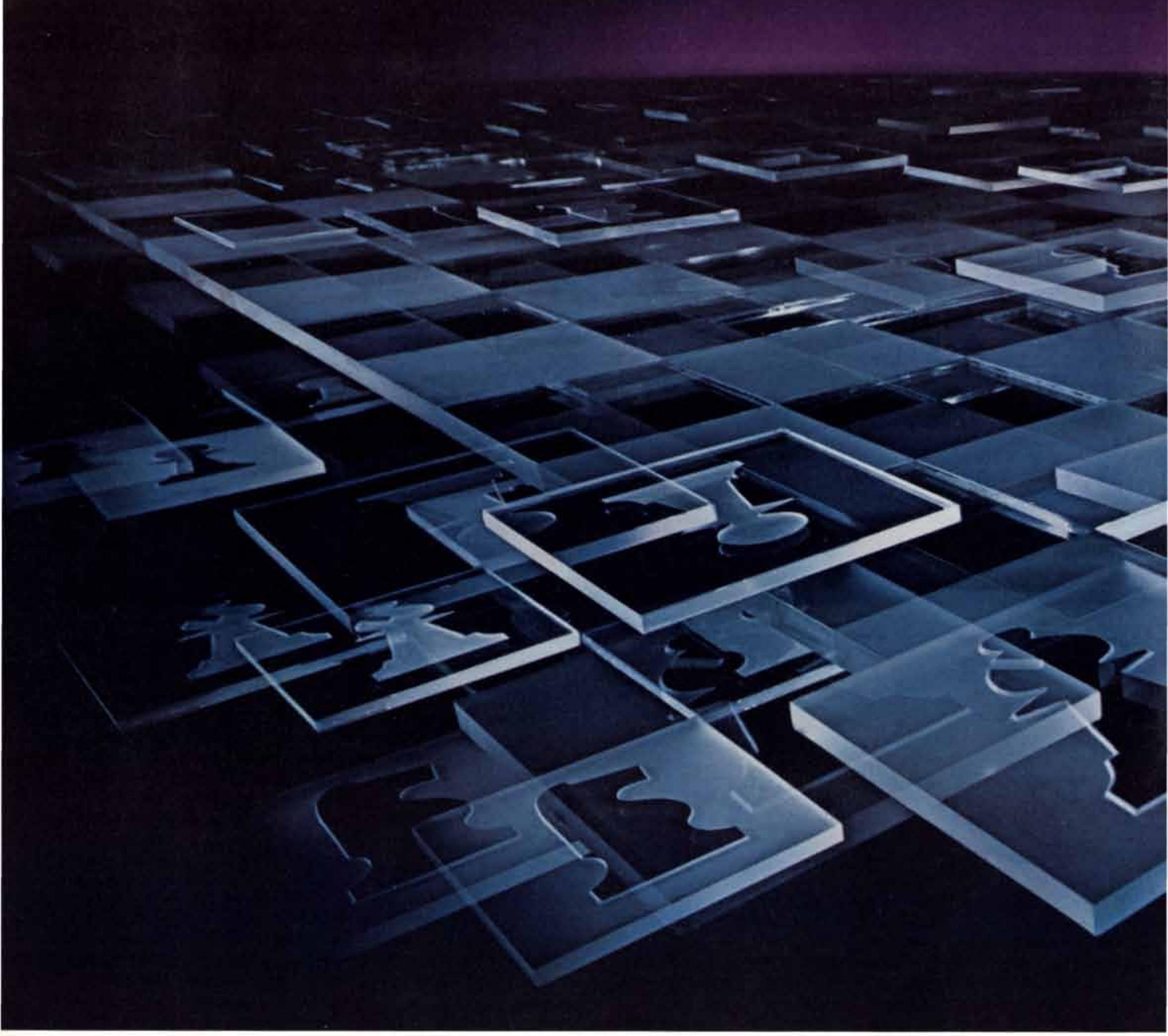
The Accord Hatchback can improve your personal economy, too. Because in spite of all it has to offer, this hatchback is still our lowest priced Accord. Which is one more reason so many people find it so nice to come home to.

*For 5-speed transmission, [32] EPA estimated mpg, 45 estimated highway. For automatic transmission, [29] EPA estimated mpg, 40 estimated highway. Use estimated mpg for comparison. Your mileage may vary according to weather, speed or length of trip. And you can expect actual highway mileage to be less. California mileage will be lower.

©1982 American Honda Motor Co., Inc.

HONDA
We make it simple.

© 1983 SCIENTIFIC AMERICAN, INC



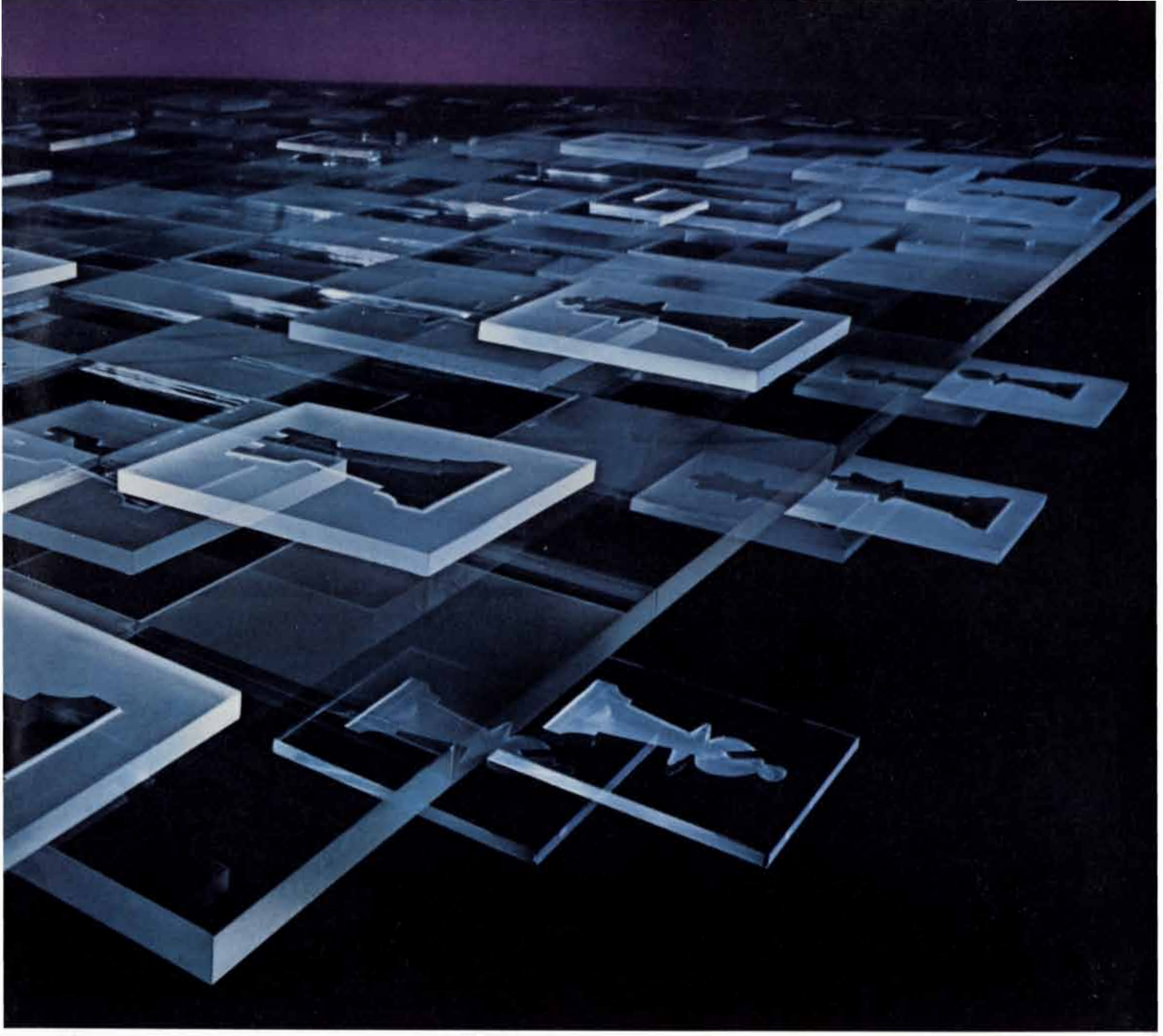
An IBM 4341 super-mini: chip design at Hughes

The board's half the size of a thumbnail, with 100,000 pieces. And it's your move.

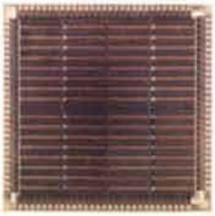
Finding paths for new interconnections on an integrated circuit is like no chess game you have ever seen. The first few moves are easy, but with up to 100,000 electronic devices crammed onto a minuscule silicon chip, the possibilities soon become astronomical. That's why engineers at Hughes Solid State Products in Newport Beach, California use an IBM 4341 super-mini.

Brian Tien, head of design automation, says, "Without the IBM 4341 we couldn't get this much function on a chip. With it, we can finish a circuit in a few weeks instead of many months."

© 1983 SCIENTIFIC AMERICAN, INC



The engineers work interactively with the system, assigning logical functions to devices on the chip. Then, using software, they route conductors—deposited strips of metal—to connect the logic gates. If a pathway becomes too crowded, another layout is automatically provided. And another. Until they find the series of moves that works.



This Hughes integrated circuit is a high-density gate array with two-level metal interconnections. Actual size is only 0.3 inches square.

They can think about the problem, and not worry about the computer, with microcode-assisted IBM software called VM/CMS. "Our people find VM ideal for interactive computing," Tien says. "As many as 40 design engineers make demanding use of the 4341 at once. Response time is excellent. The full-screen editor speeds up programming. And the executive language is simple, yet powerful."

The 4341 is an excellent example of IBM technological leadership. Internal data paths, and the arithmetic and logic unit, are 64-bits wide, built of 64K-bit chips that IBM has been mass producing since 1978. Real memory goes to 16 megabytes. The multiple, semi-autonomous processing units use high-density, large-scale-integration technology.

IBM offers engineering and scientific users extensive support: consultants, education and access to professionals. Tap into our many years of experience.

For an informative brochure, write Dr. Jack W. Hugus, IBM Engineering and Scientific Marketing, 1133 Westchester Avenue, White Plains, NY 10604.

IBM®

© 1983 SCIENTIFIC AMERICAN, INC

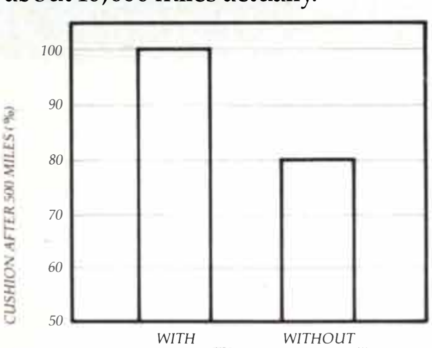
THIS NEW AIR-WEDGE IS OVER 10,000 MILES LONG.

It's hard to believe. But we just made something shorter without reducing its length.

Remember the NIKE-Air™ sole? Well, we just came out with a condensed version. The Air-Wedge.™

We did it the minute we learned from our survey that three out of four runners were heel strikers. With the Air-Wedge, we can now give those runners the protection of Air, but only where they need it most. Right under the heel.

True, the Air-Wedge is a lot shorter. But it reaches just as far—about 10,000 miles actually.



After 500 miles, shoes with the Air-Wedge lose none of their cushioning properties. Those with EVA foams, however, will suffer a 15% to 20% loss.

And probably even further. We don't know how much further because our lab technicians threw in the towel after watching the NIKE-Air sole successfully handle over 6,000,000 impacts.

That's the funny thing about Air. You'd think it would be fragile, susceptible to blow-outs, leaks, etc. When in reality it will outlast virtually every other part of the shoe, from the laces to the outsole.

That's not even the good news.

What's truly phenomenal about the Air-Wedge is that its cushion won't break down. It will absorb just as much shock on the first step as it will on the 5,999,999th.

Pretty remarkable. Especially considering the typical shoe with an EVA wedge will lose about 15 to 20 percent of its cushion after just 500 miles.

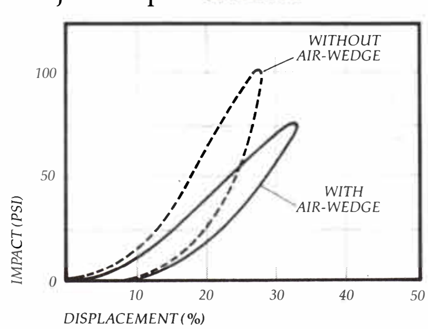
Such rapid compaction isn't just sad, it could be dangerous.

Because, if you're a runner, you might as well hit your heel with a five pound hammer. That's how much shock is generated with every step.

Not so funny.

And, unless you like buying a new pair of shoes every 300 miles, not so cheap.

But the Air-Wedge does more than just keep its cushion. According to tests carried out in our sport research lab in Exeter, New Hampshire, it gives better cushion the first day out of the box.



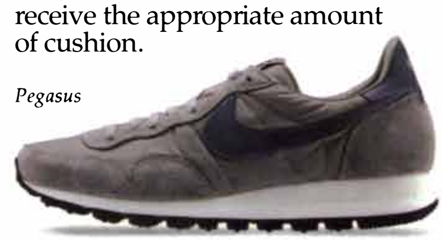
When subjected to impact testing, a new shoe with the Air-Wedge doesn't bottom out like a new shoe with EVA foam. Instead, it responds in a more linear fashion, delivering 12% better cushion. The steeper the slope on the curve, the harder the material becomes when it is compressed under your heel.

Whenever we stuck an Air-Wedge in a developmental shoe, shock absorption immediately jumped 12 percent. The reason being, the Air-Wedge simply doesn't bottom out, or get stiffer as it is compressed.

Okay, but how stable is it? Answer: surprisingly good. We also ran some direct comparison tests between two similar shoes, one with the Air-Wedge, the other with an EVA foam wedge. As far as controlling rearfoot motion, they both did the job equally well.

One last thing. Not all runners will run on the same amount of Air. We have scaled the pressure in the Air-Wedge according to shoe size, so that runners of all sizes will receive the appropriate amount of cushion.

Pegasus



Now where can you find this amazing device? In two new Nike models. The Pegasus. And the Odyssey.

We've made many claims about our NIKE-Air shoes. And with the introduction of the Air-Wedge, we're making even more.

But, then, we know it's going to let you down.

A lot easier.

And a lot longer.

Beaverton, Oregon

© 1983 SCIENTIFIC AMERICAN, INC

meant little to the slowly drifting animals, mere living dustpans over the bottom. Ammonites dwelt on the continental shelf, acting as health police of the sea floor. They seem to have filled the role of marine snails today, consumers of all small morsels, dead or alive. There is a South Dakota ammonite with 16 holes in its shell in a pattern neatly fitting the teeth of a marine reptile of the period. Other coils are cut away as though with scissors by some large crab. But they came to an end, those neutrally buoyant, underpowered subs, once the powerful bony fishes and newly instrumented unshelled cephalopods arose, probably less from direct predation on the armored adult coils than from the loss of tiny larval forms and the tough competition overall.

The jacket strip placed around the second of these books, a chunky little field guide, the right size to stuff determinedly into an outside pocket, is adorned by a pearly four-inch ammonite. That gleaming coil was found long ago in South Dakota and is held by the Peabody Museum at Yale, where lately it served as model for one of the striking photographs that are the heart of the work. From the Japanese printers through the designer and the photographer to the author, a Rutgers paleontologist, the entire team merits readers and praise.

This is a first-rate entry to the shelf of field guides for the American amateur naturalist, young and old. Its apparatus is as good as its images: the way to identify a find or cherish a hope is to seek its general form in the silhouettes, turn to the indicated color plates to find similar examples, narrow in by using the geologic period of your location judged from the detailed maps given here region by region and choose the best match. The text paragraphs help you to decide and enrich your understanding of the ancient plant or animal. Glacial, granite New England and Precambrian Quebec are held not sufficiently fossiliferous to appear on these maps, but people there will find the book an incentive to travel, if perhaps only to a museum. The specimens photographed come from one of four distinguished Eastern museum collections, and they are well identified and located. The book shines with care and forethought.

MEGALITHOMANIA: ARTISTS, ANTIQUARIANS AND ARCHAEOLOGISTS AT THE OLD STONE MONUMENTS, by John Michell. Cornell University Press ($16.95). Fierce Farmer Robinson, the churlish "stone-killer," wantonly broke up Avebury stones by pouring water over each rock as it lay across a pit filled with a blazing wood fire. Robinson's face and name survive because he was made the subject of the tailpiece of the Avebury book of the Lincolnshire physician William Stukeley. Topographic engraver of genius, devotee of the English countryside, careful recorder of its monuments, Stukeley was caught up in the 1740's by an inner religious conversion, out of which flowed his two volumes on Stonehenge and Avebury, celebrating his conviction that it was the British Druid priests whose noble serpentine temples are now the standing stones and artificial hills of Wiltshire. The poet-prophet William Blake followed but inverted Stukeley's generous vision: "All things Begin and End in Albion's Ancient Druid Rocky Shore." You can see Blake's own trilithon engraved amid moon and stars against these words, from his *Milton;* cruel priests at Stonehenge had for Blake usurped the ancient humane cults.

John Michell has here assembled a marvelous sampling, mostly in period drawings and old photographs, to give an account of the vicissitudes the megaliths of Britain and Atlantic France have undergone. That damage came physically less from lithoclasts such as Robinson than from zealous excavators, and in symbol less from evocative poet and artist than from those who over the centuries have sought to draw meaning from the structures in a more argumentative style. Once General Pitt-Rivers excavated a barrow in style nothing was left of it. There is little doubt that the projection of attitude and theory on mute stones, the elevation of modestly supported conjecture to high dogmatic truth, remains a vivid part of the encounter of each generation with the enigma of purpose that is still dim behind circles and dolmens as plain as a pikestaff.

Stukeley's images were pretty faithful to the views he saw. The iconography of Stonehenge is rich, from a 14th-century manuscript rendering of the stones as a *rectangular* temple to the painterly visions of Constable and Turner. Michell clearly sympathizes with the visionaries more than with the meticulous archaeological excavators; their century of the spade ended the Druidic mythmaking and the Roman temples, but it carried with it a certain disdain for the great circles themselves, a disdain that extended to the denigration of all efforts to seek their meaning.

Stonehenge as a whole received little scholarly attention, even though from old Stukeley to the careful measurements by the astronomer Norman Lockyer there were plenty of hard data to link the circles in some part with orientation to sun and stars. It is less than 20 years since the arguments of Gerald Hawkins and the measurements of Alexander Thom—by now meeting their newly skeptical reaction—ignited anew the fires of learned imagination. They began a lasting rehabilitation of the idea that astronomical elements at least in part inform the circles and the alignments.

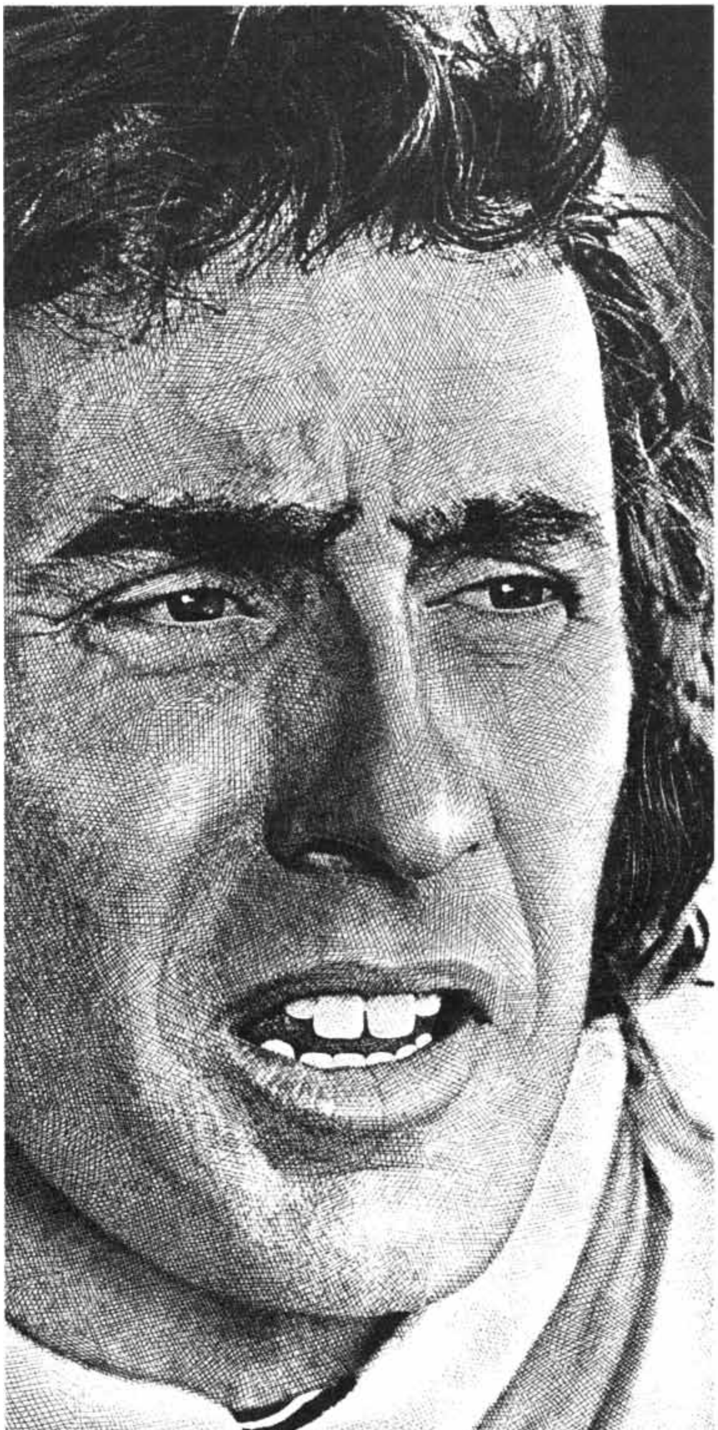
Holes and spirals and lines on the rocks are nowadays read as diagrams, even as inscriptions. There are a delightful few pages on the pitfalls of decipherment, reaching sometimes "the very heights of inspired lunacy." In 1833 an Icelandic scholar construed five heroic poems in an ancient meter out of the marks on a long-celebrated Swedish rock. It has turned out, alas, that the marks so cunningly read are natural lines and cracks, although his poem sequence remains a masterpiece of the old form. "No one has ever disputed its fine quality."

Michell is more willing than most scientific readers can ever be to consider the suggestion that the sites of these antique wonders are in fact special geophysical locales, magnetic or healthful in some way divined by ancient and modern dowsers and geomancers, with or without old visions or splendid new instruments. Surely we see here, as Michell himself wisely points out, the rise of a new myth fit for our ecologically concerned times, the search for lore that once fitted our yearning species more harmoniously into our lands. Pierced stones natural and manmade from Britain to India hold folk promise for health and fertility; infants are still passed through them.

The symbol of birth is clear enough without evoking some unlikely energy fields; one doubts the need for epidemiological studies of the neighborhood. The book nonetheless does much in a gentle way to document the perils of grand conclusions from slight premises, and the enduring danger of the very same preconceptions that are indispensable to progress. Let a hundred schools contend, says the author; he mostly wants to infect the reader who has not seen those "battered old rocks and mounds of earth" with the appealing half obsession he calls megalithomania. Magical cairns are not unknown on the New World plains, either; a wonderful drawing of a buffalo skull placed on a pointed rock as it was found by Prince Maximilian zu Wied on his celebrated expedition to the Yellowstone in 1841 and another of the winding Great Serpent Mound of Ohio help to Americanize this delightful book, if with only a few images. The epigraph is a pleasure; the hard-boiled editor of *Antiquity* remarked in that journal in 1961: "The problem in archaeology is when to stop laughing." This reader still laughs at any claims of efficacious magnetic structures, but the notion of antique sight lines across country is slowly becoming less risible.

© 1983 SCIENTIFIC AMERICAN, INC

Who needs criticism?



Jackie Stewart Three-time world champion driver, now consultant to Ford Motor Company.
Get it together—buckle up.

I've made my living over the years driving cars. Grand Prix cars, sports cars, rally cars, the most honest cars in the world.

You name it, I've driven it.

But in doing so, I've always told a builder just what I thought of his pride and joy.

Blame my skeptical nature.

That's part of it. But the rest of it is the fact that to earn my pay I had to drive the car.

And, at some point, drive it very hard.

And that's what the people at Ford have invited me to do. Drive hard, and tell them what I think...good or bad.

Apparently, they believe the unbiased opinion of a qualified outsider will help build better cars.

They want me to be an honest critic.

It's an unusual position.

But at today's Ford Motor Company, it's not business as usual.

I've seen proof of this in the 1983 Ford, Lincoln, and Mercury cars.

They're full of new ideas.

Microprocessors, electronic engine controls, space-age materials, robotic assembly, employee commitment to quality, are just a few of the things I've had to take in while learning about how things are done today at Ford.

Things have changed.

I'm convinced things have changed at Ford. And I'm glad to be part of it. Even though my part may be that of a critic.

To my way of thinking that's all to the good. Most car builders will listen to constructive criticism. But I know one company that goes even further, by actively seeking it.

Today, having a resident car critic is certainly a new way of doing business.

Ford has it. Now.

Ford

FORD · MERCURY · LINCOLN · FORD TRUCKS

© 1983 SCIENTIFIC AMERICAN, INC

Established 1845 SCIENTIFIC AMERICAN February 1983 Volume 248 Number 2

The Future of American Agriculture

In the coming decades the price of farm products in the U.S. will be influenced not only by constraints on the supply of essential inputs such as land, water and energy but also by the demand for exports

by Sandra S. Batie and Robert G. Healy

Agriculture has long been one of the most successful sectors of the U.S. economy. The decline in the proportion of the labor force engaged in farming, a trend attributable largely to the increasing mechanization of agricultural work, has culminated in an exceptionally high level of farm-labor productivity. A mere 3.1 percent of American workers produce not only an abundance of comparatively low-cost food and fiber for domestic consumption but also a substantial (and in some ways troublesome) surplus. Over the past three years U.S. exports of agricultural commodities have averaged more than $40 billion annually, up from less than $10 billion a decade ago. The foreign exchange generated by these exports has been particularly valuable in cushioning the economy against external shocks; for example, since 1973 payments for agricultural exports have offset more than 62 percent of the country's expenditure for imported oil.

In spite of this impressive record, concerns have been expressed about the ability of American agriculture to maintain, let alone expand, its output at reasonably constant prices. For one thing, it is argued, the U.S. no longer has the excess capacity it once enjoyed in such basic agricultural resources as land, water and energy. For another, chronic problems such as soil erosion appear likely to reduce fertility in some areas. There is no guarantee that it will be possible in the future, as it was in the past, to compensate for an ensuing loss of productivity by the genetic improvement of crops, the application of more fertilizer or other technological remedies. Furthermore, disproportionate increases in the price of certain agricultural inputs, particularly energy, will tend to reduce the profitability of many current agricultural practices. These developments, when combined with other factors, such as the widespread adoption of monoculture (the raising of a single type of crop) or the possibility of a major change in climate, may make American agriculture increasingly vulnerable to disruption.

As economists with a special interest in this topic we took part in a recent interdisciplinary study, sponsored by the Federal Emergency Management Agency and the Conservation Foundation, of the factors likely to affect the long-term productivity of American agriculture. Although there was by no means a consensus among the experts, many shared an outlook we would characterize as guarded optimism. We shall explain the grounds for this judgment in what follows, as we review the factors that enable the U.S. to produce food abundantly and cheaply: the quantity and quality of the basic agricultural inputs, the pace of technological innovation and the advantage of a favorable climate.

In a sense there is nothing new about the present apprehension concerning the long-term adequacy of the food supply. Throughout history the question of whether there would be enough food to feed the growing world population has been raised repeatedly. In 1798 Thomas Malthus advanced his famous thesis that the world population was increasing faster than the means of subsistence. Only war, famine, disease or some form of "moral restraint," Malthus argued, could bring the population into equilibrium with the food supply, and then only at a bare subsistence level.

Although there are still millions of malnourished people, for the world as a whole the Malthusian trap has been avoided. The apparent limits to the expansion of agriculture have been circumvented by an assortment of technological stratagems: mechanization, improved plant and animal species, more fertilizer, better control of pests and disease and the development of new cropland through irrigation, deforestation and drainage. As a result higher productivity has been obtained from scarce resources, and comparatively abundant resources have been substituted for scarce ones. In the U.S., for instance, the relative scarcity of labor stimulated first the exploitation of more-fertile cropland and later the introduction of farm machinery. In the early 1800's, before the widespread mechanization of agricultural work, it took more than 250 hours of labor and five acres of cropland to produce 100 bushels of wheat. Today it takes only three or four hours and less than three acres to produce the same amount.

The first major concern about the future of American agriculture we shall address here has to do with the amount of arable land. Considerable attention has been paid recently to the conversion of agricultural land to other purposes, particularly in the course of urban development. According to the best available estimate, made by Michael Brewer and Robert Boxley for the Federally sponsored National Agricultural Lands Study, some 875,000 acres of actual or potential cropland were converted to urban uses annually between 1967 and 1975. This corresponds to an annual conversion rate of .16 percent of the country's total inventory of about 540 million acres of arable land. If the conversion continues at the same rate, 21.9 million acres will be removed from the agricultural inventory by the year 2000.

Is the rate too high? Is the future food supply of the nation therefore in jeopar-

© 1983 SCIENTIFIC AMERICAN, INC

dy? In exploring questions of this kind it is helpful to keep in mind that the land actually used for crops in 1981 was estimated by the Department of Agriculture to be 391 million acres. This exceeds the previous high of 387 million acres, reached in 1949; a modern low of 331 million acres was reached in 1962.

Projections of future cropland requirements are generally based on a worst-case analysis: they begin with a conservative estimate of the prospects for improved yields and a liberal estimate of the growth of food demand. In making both estimates it is assumed that food prices will not increase in relation to other prices. (Rising food prices would limit the need for cropland not so much by reducing caloric consumption as by inducing consumers to reduce their consumption of meat; much of the cropland that is now devoted to growing feed for cattle, hogs and poultry could then be converted to more efficient forms of food production.)

Estimates of the additional cropland that will be needed by the year 2000 range from 26 to 113 million acres. Assuming that potential cropland can be brought into production without undue economic or environmental costs and that some of the land now used for pasture can be converted to crops, there is clearly enough available new cropland (roughly 150 million acres) to rule out a shortage of arable land by the turn of the century, even if the conversion of agricultural land to other purposes reaches a total of 22 million acres by that time. The conclusion remains valid even if one accepts the highest estimates of the need for additional cropland, estimates based on a high demand for agricultural exports, comparatively low yields and no relative price increases for agricultural products.

The foregoing analysis does not take into account the fact that much of what is considered potential cropland is already devoted to raising trees and livestock. The conversion of this land to cropland could entail a significant reduction in the supply of forest products and meat. Little research has been done on such possibly important side effects.

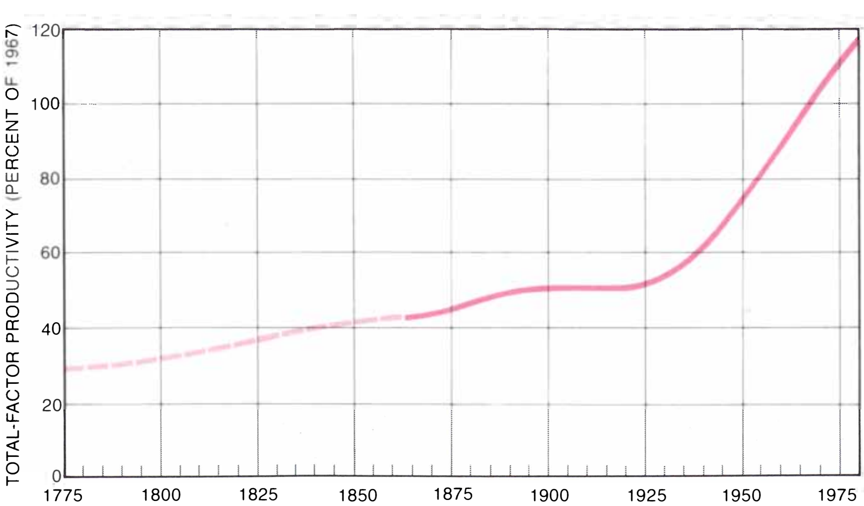
Moreover, aggregate statistics on U.S. cropland can be misleading, since they may mask important regional changes in land use. Between 1939 and 1978 roughly a third of the cropland in the Northeastern and Southeastern states was converted to other uses, whereas all other parts of the country gained cropland. Much of the cropland removed from cultivation in the East has gone not to houses, shopping centers and the like but to forests. According to Philip M. Raup, an agricultural economist and land-use expert at the University of Minnesota, "industry has been the chief competitor for farmland, measured not in acres used but in labor force withdrawn. The associated urbanization has generated demands for open space, for recreation and for residential land uses that are inextricably combined with the largely unplanned expansion of private, noncommercial forest land. Our most urban and industrial regions have become the most heavily forested. In the competition for cropland the message to date is clear: local trees have been preferred over local food." Furthermore, in the Southeast much of the former cropland is now pasture, as the South has become an important source of cattle for the feedlots of the Middle West.

Another major competitor for cropland has been the Federal highway program. Much of the interstate highway system was built in the 20 years from 1956 through 1975. It would be difficult to overemphasize the effect of such a vast construction effort on the nation's land-use patterns. The new highways not only took almost two million acres for right-of-way but also brought the possibility of urban and industrial development to additional millions of acres along the interstate routes. Most of the available data on changing patterns of land use in the U.S. were gathered during the period when the effects of the highway-construction boom were at a peak. These statistics form a poor base for projecting future cropland-conversion rates, since the U.S. is unlikely ever again to embark on such an ambitious expansion of the national highway network.

At about the same time that the highways were being built a massive dam-construction program was also under way. The reservoirs flooded large areas of cropland and made much of the adjacent acreage useless for farming. In the Missouri River Basin, for example, low-lying grassland now under water once supplied the winter feed that made possible the summer grazing of the drier, upland areas. As with highways, however, it seems likely that the dam-building era is for the most part ended.

Unlike the effects of highway and dam construction, the full impact of urbanization has probably not yet been felt. Indeed, the distinctions between the traditional patterns of land use are becoming blurred. New mixed-use patterns blend farms and rural residential zones, as migrants from urban and suburban areas repopulate the countryside. Most rural residents are no longer farmers; people living on farms now account for only 10 percent of the rural population. The frictions that arise in such a heterogeneous population may further reduce the profitability of farming and take additional land out of production, beyond what would otherwise be converted to nonagricultural uses.

The demand for residential development of rural areas promises to grow as the members of the "baby boom" generation of the 1950's and 1960's become home buyers in the 1980's and 1990's. Some 40 million Americans will reach the age of 25 between 1980 and 1990; the resulting demand for housing may be greater than the U.S. has registered in any previous 10-year period. Land required for recreational purposes, vacation cottages and retirement homes is also expected to increase. Probably as important as direct competition for present cropland is the impact of these residential and recreational demands on the supply of potential cropland. The conversion of such land to nonfarm uses is only part of the process; there are also subtler effects resulting from the division of the land into parcels of uneconomic size or the interspersing of developments that would inhibit the ex-



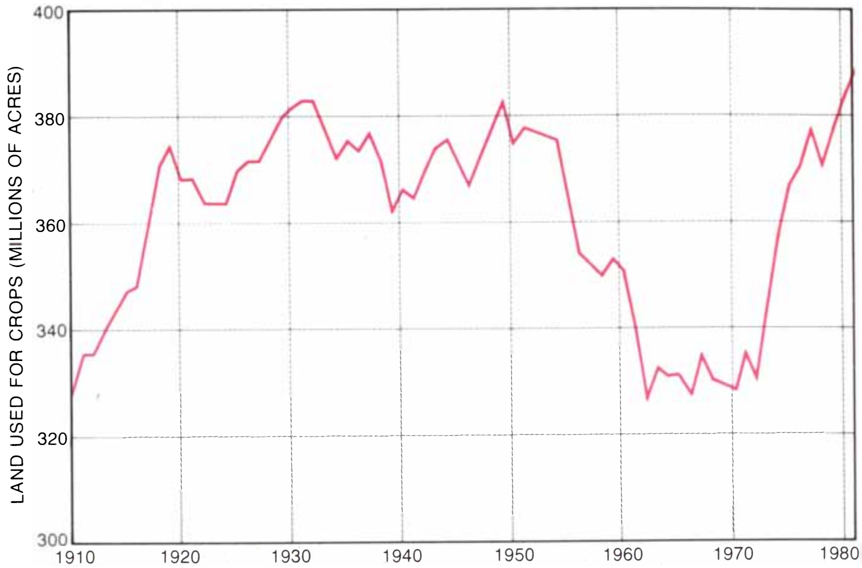
EXTRAORDINARY GROWTH of agricultural productivity in the U.S. over the past two centuries is traced in relative terms. Total-factor productivity is calculated from estimates of the major inputs and outputs of farm production. Data are from Department of Agriculture.

© 1983 SCIENTIFIC AMERICAN, INC

tension of farming to parcels adjoining present-day farms.

Cropland is also being claimed increasingly for the production of energy. This category of land use includes everything from coal mining to the production of fuels from crops or crop residues. The strip-mining of coal is clearly on the increase, particularly in the Great Plains states, where coal seams accessible from the surface are up to 100 feet thick. It is estimated that coal-mining operations will occupy an annual average of 568,000 acres in the U.S. between now and the end of the century. Of course many of these acres are scheduled for restoration. Although restoration takes time, is expensive and may not perfectly replicate the previous condition of the land, the loss of food production does not appear to be substantial. On the other hand, there is some concern that coal mining and the associated power plants and synthetic-fuel operations could be a major competitor for water in the Great Plains states, thereby creating a new constraint on food production in the region.

More serious in terms of the country's capacity to produce food is the proposed cultivation of crops for fuels such as gasohol. It seems doubtful at present that alcohol made from grain could compete with other fuel sources unless its production were subsidized. Moreover, achieving a 10-percent-alcohol blend in the more than 100 billion gallons of gasoline currently burned per year in the U.S. would require 60 percent of the nation's corn crop. The obvious repercussions on food prices are sobering. The Reagan Administration has so far been reluctant to provide subsidies for biomass conversion, however, and so this fuel alternative is not likely to become a serious competitor for cropland in the near future.



CROPLAND IN ACTUAL PRODUCTION in the U.S. reached a new high of 391 million acres in 1981, up sharply from a modern low period that ended about a decade ago. Potential cropland, including land now in pasture that could be converted into cropland, amounts to another 150 million acres or so, giving the country a total inventory of about 540 million acres of arable land. Estimates of acreage were made by workers at the Department of Agriculture.

In addition to cropland two other basic resources in food production are water and energy. In much of the western U.S. irrigation is essential for high crop yields. In more humid areas irrigation can improve yield, facilitate double-cropping and enable farmers to plant more valuable crops. In the country as a whole irrigated acreage has tripled since 1940, and it now accounts for nearly a fourth of the value of the nation's crops. The large increase in irrigated land can be traced to the fact that in many areas water was long considered a free commodity, available to the farmer for no more than the price of transporting it. Low energy prices and Federal subsidies kept the cost of transportation low.

The era of low-cost water is past. In areas where farmers rely on ground water a falling water table and increased pumping costs are making the recovery of the water more expensive. High energy costs also make it less attractive to pump surface water over long distances. In the West most surface water is already fully allocated, and so new consumers will increasingly have to bid water rights away from present consumers.

Much of the recent growth in irrigated acreage has been concentrated in four states: Nebraska, Kansas, Oklahoma and Texas. Here irrigation is used mainly to grow sorghum, corn and alfalfa. These crops are commonly raised to feed cattle, and many of the nation's largest feedlots are in the same states. The beef production has been made possible by the "mining" of the Ogallala aquifer, a vast underground lake that spans eight states [*see illustration on next page*]. Recharge to the aquifer is exceedingly slow, with the result that the water table has been steadily falling and pumping has become more costly.

Eventually the combination of a falling water table and rising energy costs will probably make ground-water irrigation less popular. Farmers are adjusting to higher water costs by adopting water-conservation measures; still, the difference in profitability between irrigated farming and dryland farming has narrowed. If dryland farming returns to these areas, concentrated cattle feeding will no longer have a competitive advantage in the states overlying the Ogallala aquifer. As Raup has stated, "unpriced water... has been capitalized into a national level of beef consumption that cannot be sustained in the long run without a return to the feed-grain supplies of the Corn Belt. We have a fed-beef economy that has become dangerously dependent on an exhaustible resource base."

Whatever happens to the feedlots, it is clear that unless grain prices increase to offset the increasing cost of water, irrigation cannot continue indefinitely to contribute to agricultural expansion as it has in the past three or four decades. Kenneth D. Frederick of Resources for the Future points out that this is not necessarily a bad thing. He comments: "If the transition from water abundance to scarcity allows for an efficient use of resources over time, irrigation will contribute to agricultural production and growth for many more decades. The serious problem will emerge if we attempt to keep water cheap when it is not. Such a policy will ensure its inefficient use... thereby limiting the long-term role of irrigation."

Energy is consumed on the farm for many purposes. Irrigation has already been mentioned, but agricultural chemicals and vehicular fuels make even larger demands. The historical trend has clearly been toward greater energy consumption. Herbicides are now applied to more than 90 percent of the acreage planted in corn, whereas a decade ago only 57 percent of the corn acreage was treated with such chemicals. Similarly, the horsepower of tractors increased by 31 percent between 1966 and 1978.

The economic reasons for this trend are not hard to find. Between 1950 and 1978 labor costs on the farm went up twice as fast as energy prices, and land prices rose almost four times as fast. The incentive was to substitute energy—in the form of chemicals, fuel and fertilizers—for both land and labor. The price

© 1983 SCIENTIFIC AMERICAN, INC

relations were accentuated by government policies that tended to keep the prices of fuel and fertilizer as low as possible. Natural gas, for example, has been kept at an artificially low price by government regulation, and electricity is supplied to most farms on a highly favorable basis.

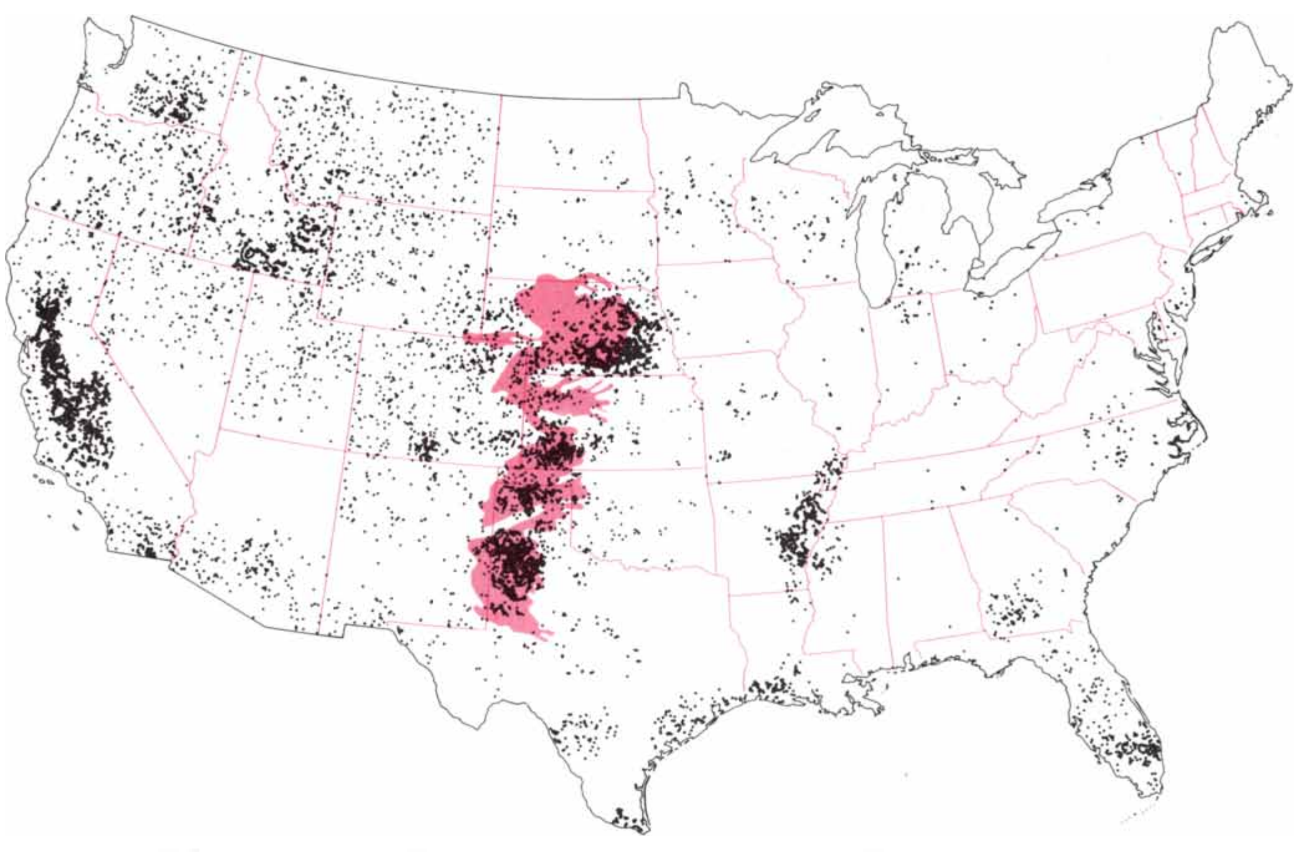
Low-cost energy has been exploited in agricultural production not only to reduce the need for farm labor but also to increase the productivity of all other inputs and to reduce the risk of crop failure or spoilage. The reduction of risk can take many forms: limiting the variability of the harvest through the application of pesticides, reducing the adverse effects of climate through irrigation and reducing the risk of spoilage by drying or refrigeration.

Because energy performs so many vital functions for the farmer, agricultural energy consumption is unlikely to be reduced by much, in spite of rising energy prices. Relative energy prices do not appear to be increasing fast enough to disrupt production or to exert much pressure on food prices. There may be some shifts in the location of farm enterprises with rising energy prices, however. Higher transportation costs may make it feasible to fatten cattle on feedlots in the South rather than shipping them to the Middle West. Local crops such as tomatoes may be better able to compete with nonlocal crops that are cheaper to produce but more expensive to transport. Technological and institutional changes are also likely to facilitate energy conservation; better irrigation scheduling and integrated pest-management planning are two examples. Some farms, such as dairies, may even resort to meeting part of their own energy demands through the recycling of wastes.

The bounty produced by American farmers currently accounts for 13 percent of the world's wheat, 29 percent of the coarse grains (mainly corn), 17 percent of the meat and 62 percent of the soybeans. Because the U.S. has less than 5 percent of the world population, the high level of domestic food consumption still leaves a large surplus for export. Nevertheless, there is concern that further increases in output may not be sustainable unless the quality of the land is maintained. The chief problems are the erosion of soil by wind or water, the compaction of soil by heavy farm equipment, the increasing salinity of irrigated soil and certain side effects of schemes for the production of energy from crop residues.

Erosion can remove large amounts of topsoil and still not be obvious to the casual observer. A farm losing 40 tons of soil per acre per year loses less than a third of an inch of soil per year, which is not a readily visible amount. Water-caused soil erosion is most serious in the Southeastern states, where a single rainstorm can sweep away as much as an inch of topsoil from a cleared field. Significant erosion, however, is recorded on only a small fraction of the nation's cropland. Water-caused erosion associated with soybean production, for example, is concentrated in several major agricultural regions in the South and Middle West. Serious wind erosion is observed mainly in the High Plains region of Texas, southeastern Colorado and southwestern Kansas, where wheat is a major crop [*see illustration on pages 50 and 51*].

Are soil-erosion rates today greater than they were in the 1930's? In spite of many statements to this effect the question is probably unanswerable. There is some evidence, however, to suggest a recent increase in soil erosion. John F. Timmons of Iowa State University has shown that in Iowa annual soil losses decreased from 21.1 tons per acre in 1949 to 14.1 tons in 1957, but the rate rose again to 17.2 tons per acre in 1974. The reversal of the trend can be ascribed for the most part to the rapid rise in the prices of wheat, corn and soybeans in the early 1970's, a development that



IRRIGATED LAND IN THE U.S. totals about 60 million acres. Each dot on the map represents 8,000 acres where irrigation facilities existed in 1977, the latest year for which information of this kind is available. The colored area shows the lands underlain by the Ogallala aquifer, where much of the postwar growth in irrigated acreage has been concentrated. Crops are raised here mainly to feed cattle.

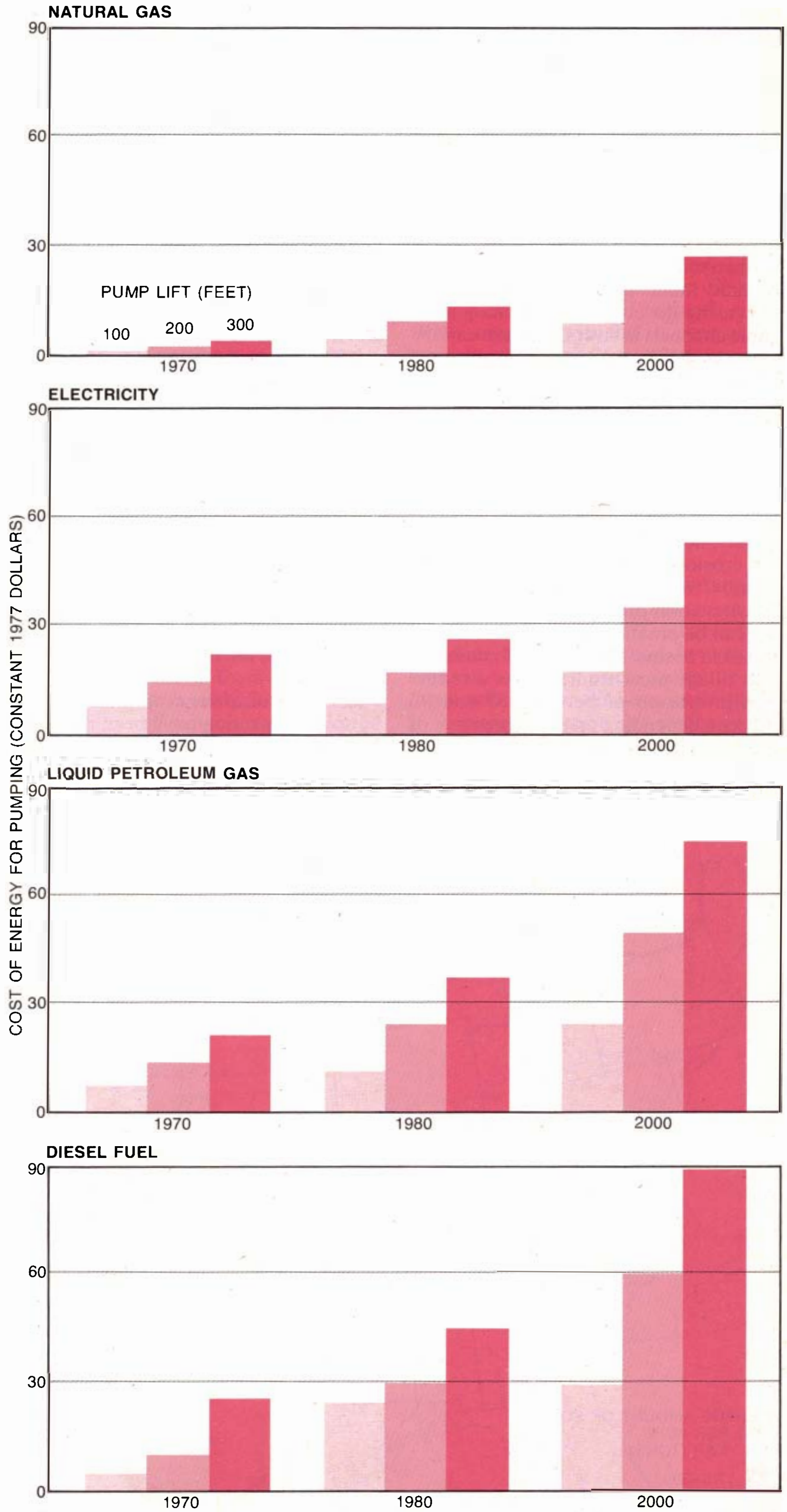
© 1983 SCIENTIFIC AMERICAN, INC

caused erosion-prone land formerly in pasture to be plowed for crops. Land devoted to these crops has increased further since 1974.

Although an increase in the acreage devoted to the production of corn, soybeans, wheat and cotton in response to a projected increase in the demand for food and fiber could lead to more erosion, the advent of "conservation tillage" may lessen the effect. Conservation tillage is defined as any method of preparing the soil that reduces the loss of soil in water runoff compared with conventional unridged, or clean, tillage. Common conservation-tillage practices include no-till, minimum-till, ridge-plant, plow-plant and chisel-plowing techniques. Since the 1960's conservation tillage has spread rapidly, and it is now practiced on an estimated 25 percent of the tilled cropland. The rapid growth of conservation tillage has come largely in response to economic incentives: it generally costs less than conventional tillage in both labor and fuel. Still, the result is also less erosion. Pierre R. Crosson of Resources for the Future predicts that by the year 2010 between 50 and 60 percent of the nation's cropland will be farmed by means of conservation tillage. He also estimates, however, that productive cropland may increase by as much as 75 million acres by 2010 and that "much of that additional land would be more erosive than land now in crop production." Thus substantial amounts of erosion may still prevail.

Soil erosion can affect productivity by reducing the capacity of the soil to hold water, by removing plant nutrients, by degrading the structure of the soil and by reducing the uniformity of the soil within a given field. The severity of these effects differs widely, depending on the region, the soil, the farming practices employed, the crop and the length of the growing season. Assessing the reduction in yield caused by erosion at a specific site is difficult because many other factors can affect yield. As erosion has progressed over the years, farmers have applied more fertilizer both to substitute for eroded topsoil and to enhance fertility. Farmers have also changed tillage practices, introduced different types of machinery and adopted new varieties of plants.

In spite of these complicating factors, experts who have studied the effect of erosion on productivity conclude that in some cases the loss of even an inch of topsoil may be significant. Where erosion has substantially reduced the topsoil depth, so that plant roots are reaching less fertile subsoil, the farmer's costs of production are higher because more fertilizer is needed to maintain yield. Furthermore, adding nutrients in the form of fertilizer can increase yield only if enough water is available in the soil;



ENERGY COSTS of lifting an acre-foot of water from various depths are estimated in this bar chart for the years 1970, 1980 and 2000. Projected increases in pumping depth, brought on by the drawdown of aquifers, combined with rising fuel prices, brought on in part by deregulation, are likely to result in a sharp increase in the costs of pumping, a development that is expected to reduce the competitive advantage of irrigated farming over dryland farming in many areas. The chart is based on a study by Kenneth D. Frederick of Resources for the Future.

© 1983 SCIENTIFIC AMERICAN, INC

hence differences in the capacity of soils to supply moisture are often the primary determinants of differences in crop yield.

In addition to reducing productivity soil erosion is also the main contributor to agricultural water pollution. Runoff from farmland carries sediment, nutrients, herbicides and pesticides, all of which are considered pollutants when they are present in excess. Sediment washed off farms is reducing the lifetime of many inland lakes and reservoirs and increasing the cost of maintaining navigable channels in rivers. Excessive nutrients can lead to the eutrophication of bodies of water and can raise the cost of purifying public water supplies. Excessive herbicides and pesticides can be detrimental to wildlife. Because the damage happens off the farm, however, farmers have little or no incentive to take such factors into account.

Technical solutions to the problem of soil erosion exist. Steep lands can be terraced. Crop residues can be left in place. Conservation tillage and contour plowing can be practiced. Runoff can be collected in basins. In the case of conservation tillage most studies show a reduction in erosion of between 50 and 90 percent, depending on the percentage of the field left covered by crop residues.

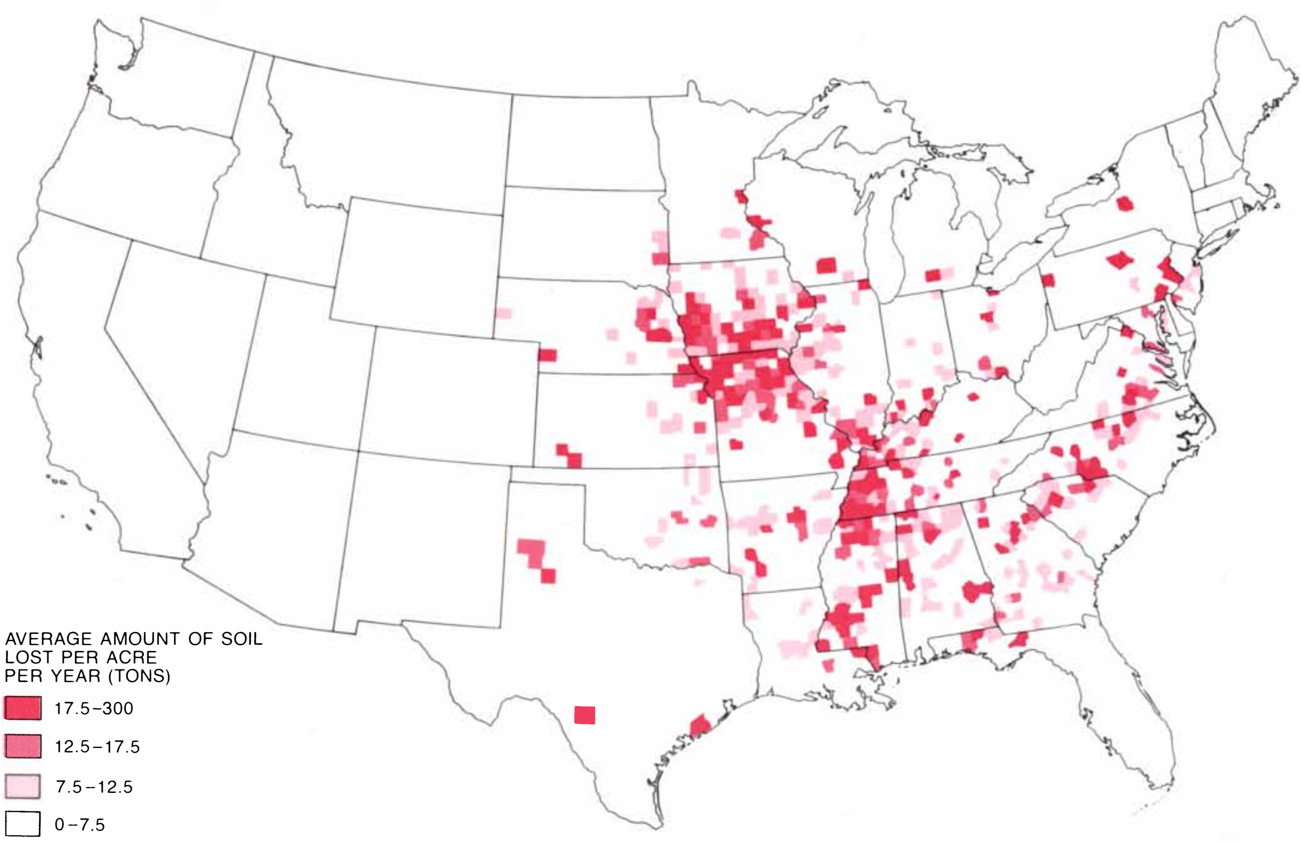
If such measures can control erosion, why are they not more widely practiced? Contributing factors may include the personal preferences and beliefs of farmers, the physical characteristics of the land, tenure arrangements, tax policies and farm-commodity programs. The main reason, however, is simply that many conservation practices do not pay for themselves. Because erosion degrades productivity and water quality, because there are few incentives for the farmer to adopt most of the conservation practices and because erosion is likely to increase as the nation responds to the demand for exports, many observers have reached the conclusion that erosion is a problem worthy of greater public concern and management.

Heavy farm machinery can lower productivity by compacting the soil. Compaction can reduce water percolation, resulting in drainage problems and increased runoff, and it can inhibit root penetration. Some plants, such as potatoes and sugar beets, are more sensitive to these effects than other plants. Compacted soils can also inhibit the transmission of gases and heat, another effect detrimental to the growth of most plants.

Instances of lower crop yields caused by soil compaction have been reported, but there is also some evidence to the contrary. For example, soybean yields in Minnesota have been shown to be higher on moderately compacted soils than on noncompacted soils. Actually little is known about the extent or significance of soil compaction. In any case, technologies and management practices are available to relieve problems caused by soil compaction. They include careful timing of farming operations, reducing the number of trips into the field, confining wheeled traffic to designated paths, deeper plowing and wider equipment. A recent study by the Congressional Office of Technology Assessment concluded that existing technology can greatly reduce the possibility that soil compaction will become an important constraint on productivity.

Excessive salinity is considered the most pervasive environmental problem associated with irrigated agriculture. When irrigation water is applied to crops, evaporation and transpiration remove the pure water, leaving salt as the residue. If rainfall or further irrigation does not flush the salt away, the productivity of the soil can suffer. Between 25 and 35 percent of the irrigated western croplands are thought to have excessive salinity.

Two areas with serious salinity problems are the Lower Colorado River Basin and the west side of the San Joaquin valley of California. In the San Joaquin area it is estimated that yields have already been reduced on more



SOIL EROSION is a serious problem on only a small fraction of the nation's cropland, although the areas affected include some of major importance for food production. The maps on these two pages show the average tonnage of soil lost per acre to water- and wind-caused

© 1983 SCIENTIFIC AMERICAN, INC

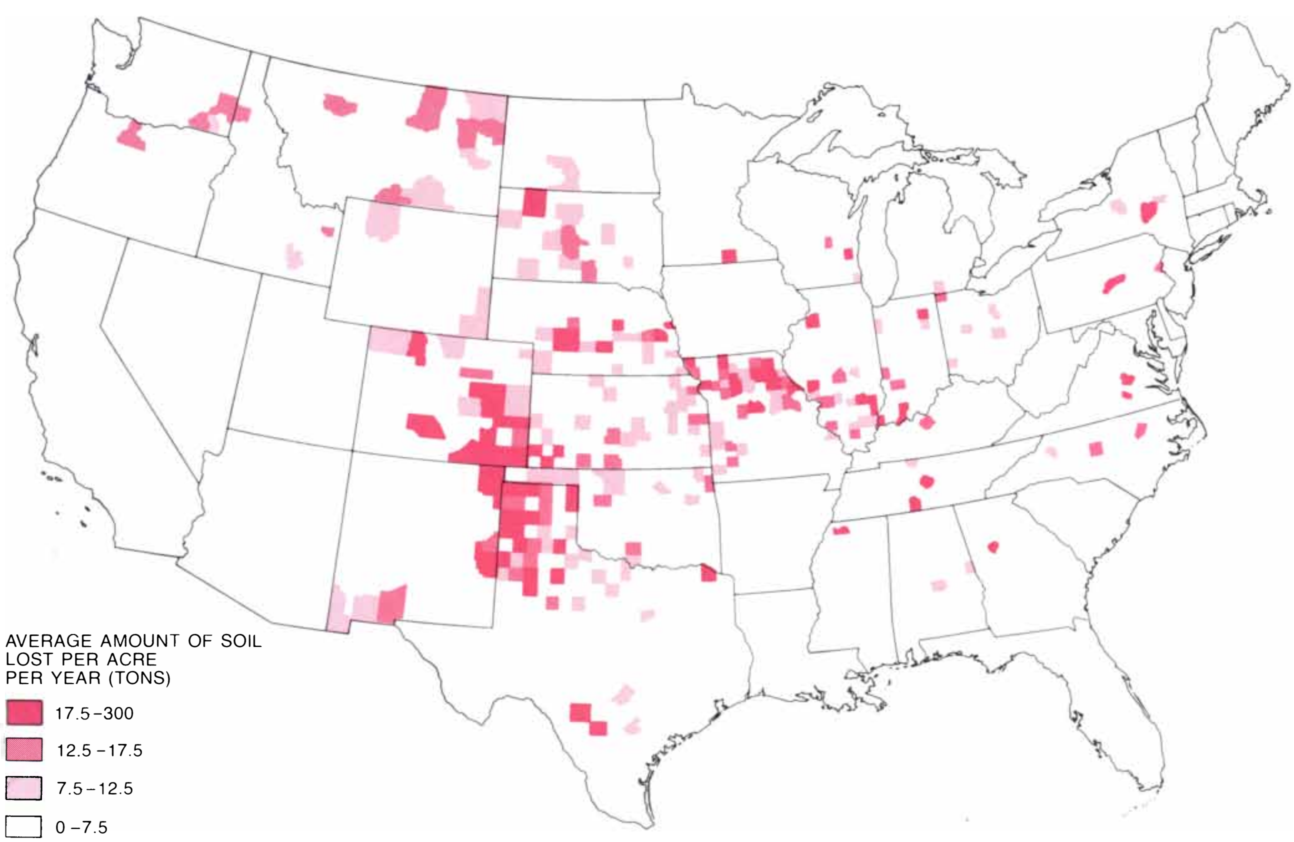
than 400,000 acres, with an estimated economic loss of $31.2 million per year. Farmers have asked for government investment to provide water of higher quality, but in the meantime there are other remedies farmers could adopt themselves. One technique is to apply irrigation water more efficiently through the use of drip-irrigation systems or lined irrigation ditches. Another approach is better planning for the recirculation of the irrigation water. The first use of the water would be for those crops most sensitive to salt. Subsequent uses (as the water became saltier) would be on less sensitive crops. Regardless of who pays for the investments needed to reduce the salinity of soils, it appears that at least some investment would be worthwhile.

The rising cost of energy has led many people to consider alternative energy sources. One possible alternative is to exploit crop residues for direct burning or for conversion into fuel. Although these schemes are technically feasible and enormous quantities of residue are available (an estimated 400 million dry tons annually), such an approach is not necessarily either economic or desirable. Crop residues are bulky and not easily transported, and they would have to be hauled at considerable cost to the plants where they would ultimately be converted. Several studies have found that various crop residues would be at an economic disadvantage with respect to coal when prices are calculated at the point of delivery to local utilities. Furthermore, residues left in place have a useful role in reducing soil erosion and recycling nutrients. For this reason farmers in many areas are reluctant to have crop residues removed. In short, the potential of crop residues as an alternative energy source is constrained by both economic and environmental considerations and in any case is nowhere near the 400 million dry tons available in principle. One estimate is that at most there would be 80 million dry tons available annually.

Granted that the quantity and the quality of agricultural land and other inputs are important to the future of American agriculture, it is nonetheless true that the Malthusian specter has been avoided so far mainly because of technological advances that have greatly improved productivity for a given level of inputs. Here too there are concerns that the growth of productivity cannot be sustained at the high level of the past few decades. Even some of the more optimistic investigators think the U.S. will do well simply to maintain the same rate of absolute yield increases in the future.

The problem is not really a lack of possible technological breakthroughs. There are plenty of these in prospect: the hybridization of crops such as wheat, the enhancement of photosynthetic efficiency, the twinning of livestock, the breeding of drought-resistant plants and so on. Instead many observers fear that the development of these technologies will be hampered by limited expenditures on research and development and on the extension programs that bring the results of such research to the farm. There has been little growth in Federal support for agricultural research since 1975, compared with the 3.8 percent annual growth rate from 1939 through 1965. Private-sector expenditures for agricultural research and development are difficult to estimate, but they are probably as large as the public-sector expenditures. Even if the private-sector research increases, however, the expenditures are more likely to be for innovations that are patentable, such as machines, drugs and pesticides.

The projected rate of technological change is of crucial importance; it has been technological progress that has enabled the U.S. to meet food demands at comparatively low prices without being faced with severe resource constraints. If increased productivity is not forthcoming or is achieved only after a long delay, and if predicted future demands do indeed materialize, then either prices



erosion in 1977. The map at the left shows the erosion of cropland used to raise soybeans; most of the erosion in this case is attributable to water. The map at the right shows the predominantly wind-caused erosion of cropland used to raise wheat. Keys indicate erosion rates.

© 1983 SCIENTIFIC AMERICAN, INC

will have to rise, exports will have to be curtailed or substantially more resources will have to be diverted from other needs and devoted to food production.

The technological revolution that agriculture experienced over the past three decades was accompanied by enormous changes in what is often called the structure of agriculture. Between 1950 and 1980 the number of people living on farms fell from 23 million to fewer than eight million and the average size of farms doubled, to just over 400 acres. Although a fairly large number of small farms still exist, the bulk of U.S. food production comes from a diminishing number of farms. Workers at the Department of Agriculture have predicted that if present trends continue, by the year 2000 the nation's 50,000 largest farms will account for 63 percent of all agricultural sales, up from 31 percent in 1974. Also projected is increased specialization by individual farms in a limited number of commodities, in contrast to the product diversification and virtual self-sufficiency that once characterized many family farms in the U.S.

The structure issue is important mainly for its political and social ramifications, but it may also have a bearing on future agricultural productivity. Increasing scale has often been accompanied by greater mechanization and more irrigation. Both have made the agricultural system more dependent on energy and other inputs produced off the farm. Crop specialization by the individual farm has led to localized monoculture, whereas regional specialization has made the food-production system more vulnerable to climatic variation and the distribution system more susceptible to the effects of energy shortages and transportation stoppages.

Some of the same technological advances that have made American agriculture so productive have also reduced the ability of some crops to withstand pests and disease. Modern plant-breeding practices have resulted in crops with comparatively little genetic variability. More than 70 percent of the acreage planted in corn, for example, is planted with the descendants of only six inbred varieties. The narrow genetic base of major crops such as corn can make them more vulnerable to disease, insects or erratic climate.

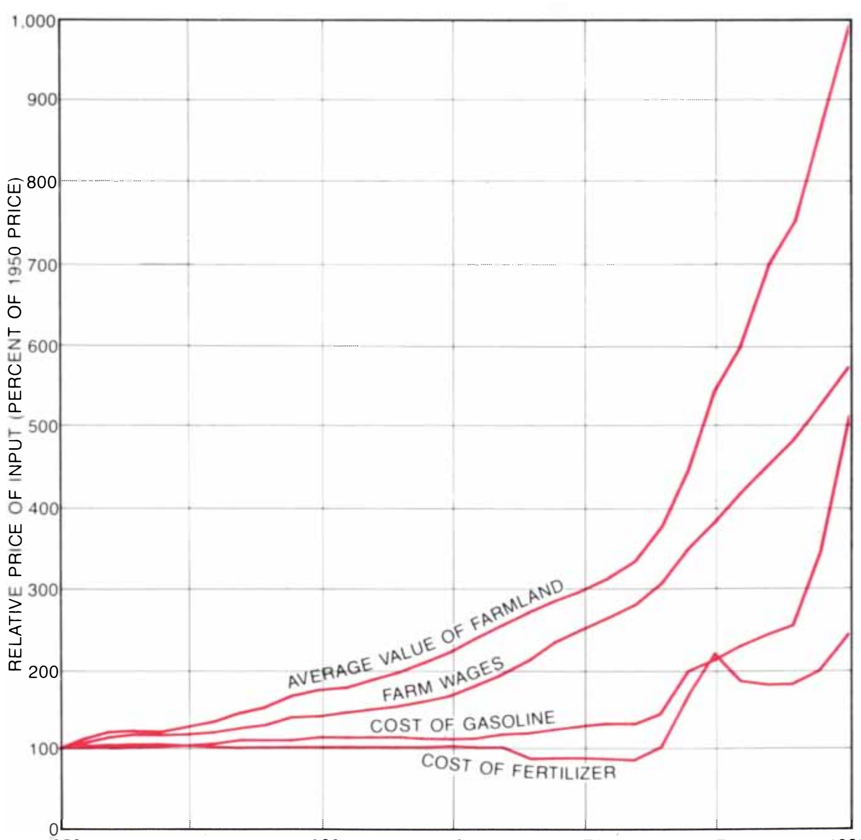
The vulnerability was vividly demonstrated in 1970 when a mutation of the blight-causing fungus *Helminthosporium maydis* caused an epidemic of corn-leaf blight in the U.S. An estimated 15 percent of the corn crop was lost. Although seed companies quickly introduced resistant seed and recovery was rapid, the single-season epidemic was perceived as threatening. Farmers became aware that the stability that comes from ecological diversity had been sacrificed in order to get the increased productivity and reduced production costs associated with plant uniformity.

There are ways of reducing the risks of having monocultured crops. The improvement of germ-plasm collections would make it possible to respond quickly to epidemics or pest infestations. Strengthening plant-breeding programs and upgrading the monitoring of diseases and pests can also improve and protect plant resiliency. Agricultural research will have to be better funded and in some respects reoriented if these and other improvements are to be implemented.

Of all the concerns about the vulnerability of American agriculture the possibility of an adverse change in the climate is probably the most difficult to analyze. The difficulty results less from limited data than from limited understanding of the reasons for climatic variation. The available evidence does suggest that agriculture has developed in an abnormally warm period. Furthermore, warm periods such as the present one appear to last for no longer than 10,000 years, and this one has already lasted for almost that long. Climatologists do not agree on whether the earth is now in a long-term warming trend or a long-term cooling trend. In any case, long-term climatic trends are probably not as important to American agriculture as annual fluctuations in weather.

Human modifications of climate further complicate the analysis. The cutting of forests, the creation or expansion of deserts, the production of acid rain and the addition of particulates, ozone and carbon dioxide to the atmosphere can change climate, but the ultimate impact on crop yields is uncertain. Some crops such as tomatoes and strawberries may actually benefit from more-acidic rain, whereas others such as lettuce and spinach could be injured.

In a report published in 1980 the National Defense University made an effort to predict changes in crop yields under various conditions of temperature and mean precipitation. Agricultural experts were asked to estimate yield changes for several combinations of conditions during the reproduction period of various crops in various countries. The estimates for corn in the U.S. showed that maximum yields would be achieved if the temperature were slightly cooler during the reproduction period and if precipitation were 20 percent



RELATIVE PRICE INCREASES of four major agricultural inputs—land, energy, labor and fertilizer—are plotted here for the period from 1950 through 1981. The largest increase by far has been in the average value of farmland, which has risen by more than 900 percent over the past three decades. The graph is based on data gathered by the Department of Agriculture.

© 1983 SCIENTIFIC AMERICAN, INC

above normal [*see illustration at right*]. Soybeans, on the other hand, would reach a maximum yield with no change in temperature but with a 40 percent increase in precipitation. Although predicting climatic change and its effects remains guesswork, it is safe to assume there will be some unfavorable weather.

There is no simple answer to the question of whether the bounty of American agriculture can be sustained in the long term. Few of the experts, if any, foresee circumstances in which future agricultural output will be insufficient to feed the domestic population, which is now growing at less than 1 percent per year. The real uncertainty lies in the future level of exports, which already claim the output of one out of every three acres of U.S. cropland. So far agricultural exports have been regarded as an unqualified benefit to the U.S., raising farm income, generating foreign exchange and absorbing the chronic grain surpluses that have plagued American farmers at various times, including the present. Until recently the only serious objection raised to farm exports was based on their impact abroad; the argument was that they kept prices in the consuming countries so low that local farmers found it impossible to compete.

As the cushion of excess capacity in U.S. farming has diminished, however, a few voices have begun to ask whether unlimited export expansion is in the national interest. Should we encourage foreign nations to become dependent on U.S. grain at low prices that may not be maintained in the long run? If American agriculture is pushed to its limits, they argue, it will not be by domestic demand alone. The future expansion of farm exports is a fundamental policy question that deserves the same attention that has recently been given to the threats to agricultural supply.

On the supply side several of the most important constraints—those on land, water and energy—will probably manifest themselves not through physical shortages but through rising prices. Farmers' responses to rising input prices may bring increasing pressure for interregional shifts in crop and animal production as well as changes in production practices. Increased irrigation costs and rising demand for grain, for example, may move the livestock industry back to the grazing patterns of the 19th century. If that happens, cattle herds will put greater demands on the land than ever before, and higher prices for both cattle and land will be the likely outcome.

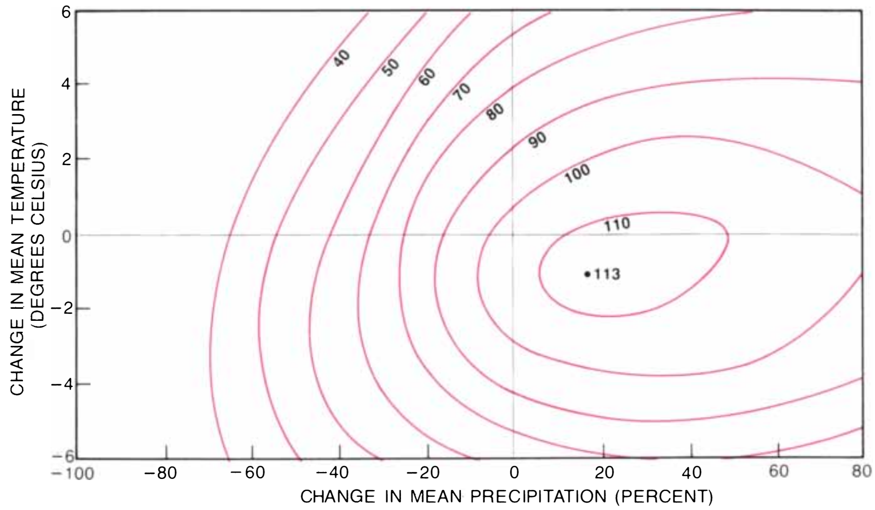
Other concerns—soil erosion, excessive soil salinity, climatic change and lags in technological growth—seem less tractable. If future increases in productivity are more in line with the moderate growth experienced between 1900 and 1950 than with the more recent high-growth era, and if increased export demands for American agricultural products do indeed materialize and are accommodated, then we may witness a return to an agriculture in which increased production depends mainly on the cultivation of more land rather than on more output per acre. The problems of reduced soil fertility and adverse climatic change would then be far more serious than they would be if increased productivity were likely.

The high degree of crop monoculture that characterizes American agriculture is a serious threat only if the agricultural establishment fails to take some simple preventive measures, such as maintaining seed collections and a healthy and diversified seed industry. Unfortunately government has been somewhat reluctant to take the necessary leadership role in this area.

In general, predictions based on past and present trends do not suggest anything approaching a crisis in American agriculture. The predictions describe only the most probable events, however, and not those that are less likely but still possible. Perhaps American agricultural policy should be focused more sharply on insuring against extreme deviations from the likeliest event. The most obvious example of such a deviation is a change in the climate to one less favorable to agriculture. The amount of insurance that might be appropriate would depend, of course, on the costs and benefits of the preventive strategy selected. The possibilities include wider spatial distribution of agricultural lands; more diversity in agricultural production by region; increased funding for agricultural research; the establishment of private and public grain reserves; programs to reduce specific threats to agriculture; modifications to government programs that encourage a structure of agriculture vulnerable to extreme events, and adjustments to the prices of agricultural products to reflect more of the private and social costs of food production.

Needed research relevant to these strategies would include work on determining the future costs of soil erosion, compaction and salinity; on gauging the effects on agricultural productivity of a severe climatic change; on assessing the costs of converting potential cropland into actual cropland, and on estimating the trade-offs in productivity resulting from wider and more diverse spatial distribution of agricultural enterprises. In most situations it will be easier to identify the costs incurred when implementing a preventive strategy than it will be to estimate the benefits from protecting against an extreme event.

The debate over American agricultural policy should not await research findings, if for no other reason than that policy debates can help to identify areas of ignorance. With the present farm economy beset by production surpluses and cost and credit squeezes on the farmer, it may seem strange to suggest debating how the U.S. might confront possible long-term shortages. The urgency of short-term problems, however, must not obscure the importance of considering how to organize an agricultural sector that can survive and flourish for as long as we expect our society to endure.



EFFECT OF CLIMATIC CHANGE on the estimated yield of the U.S. corn crop was forecast as part of a recent study done by the National Defense University. The results, displayed in the form of a contour diagram, showed that the maximum corn yield would be achieved if the temperature were slightly colder than normal during the crop's reproduction period and if the precipitation were 20 percent above normal. The contours connect combinations of temperature and precipitation that would result in the same relative yield. For example, the curve marked *80* shows the climatic conditions under which yields are projected to be 80 percent of the average yield. Similar diagrams were generated for a number of other crops in the U.S. and abroad.

© 1983 SCIENTIFIC AMERICAN, INC

The Lattice Theory of Quark Confinement

The force between quarks in a particle such as the proton has been simulated by imposing a discrete lattice on the structure of space and time. The results suggest why a free quark cannot be isolated

by Claudio Rebbi

The development of quantum mechanics put to rest the uncritical acceptance of the idea that elementary particles are the "building blocks" of matter. Often such particles do not act like hard, impenetrable blocks at all, and in many circumstances they must be described as waves. Until recently, however, it still seemed that elementary particles were like building blocks at least to the extent that each particle could in principle be isolated and observed as an individual entity. The electron, the proton and the neutron, for example, can be separated from one another and observed as individual packets of waves. Even this limited interpretation of the building-block metaphor fails in the case of the quark, the supposed constituent of the proton, the neutron and many related particles. Apparently a quark cannot be isolated; although there is abundant evidence for the existence of quarks and antiquarks bound together in pairs and triplets, an individual, or free, quark has never been observed.

As the experimental evidence has accumulated, it has begun to seem that if quarks are real particles at all, they must be permanently bound within nuclear particles. Any theory of quark interactions ought to account for this phenomenon, which is called quark confinement. It is easy to construct pictorial models of particles such as the proton in which the constituent quarks are confined. For example, the quarks can be thought of as being fastened to the ends of an unbreakable string; they are then free to move about within the volume defined by the length of the string but cannot wander away from one another. It is a formidable task, however, to formulate a theory that can account for the permanent binding of quarks and the structure of nuclear particles without violating the constraints imposed by the theory of relativity, quantum mechanics and the principles of ordinary causality.

After several years of both experimental and theoretical investigations most particle physicists are confident they at last have a theory capable of explaining the interactions of quarks. One reason for confidence is that the theory is a mathematical analogue of the most successful physical theory ever developed: the quantum theory of interactions in an electromagnetic field. The latter theory is called quantum electrodynamics, or QED, and the conceptual similarity of the theory of quark interactions to QED is reflected in the name of the new theory: quantum chromodynamics, or QCD.

The difficulty that has delayed full acceptance of QCD is that its mathematical complexity makes any rigorous, analytical prediction from it exceedingly difficult. Indeed, up to now the most eagerly sought prediction of QCD, namely the demonstration of quark confinement, has not been forthcoming. Recently, however, my colleagues and I at the Brookhaven National Laboratory have applied mathematical methods that rely heavily on the capabilities of the digital computer to the problem of confinement, and a numerical breakthrough has been achieved. Because the method explores the implications of QCD by making a series of increasingly accurate approximations, the results of the calculations do not carry the same force as a logical deduction from accepted first principles. Nevertheless, the numerical results have provided strong evidence for the confinement of quarks.

The framework for the calculations is a pioneering suggestion made in 1974 by Kenneth G. Wilson of Cornell University. Wilson proposed that QCD be formulated on a cubic lattice, an array that divides space and time into discrete points. The lattice is only an approximation to real space-time, but it allows calculations to be made that would otherwise be impossible. As the mesh of the lattice is made progressively finer the values of physical quantities defined on the lattice converge to the values QCD would predict for them in ordinary, continuous space and time. Our numerical approximations show that for an extremely fine lattice, confinement is a consequence of QCD. For reasons that will become clear both QCD and QED are called gauge theories; the computational method I shall describe is therefore called a lattice gauge theory.

The original impetus for the quark model was the need to bring order to the large number of particles that exhibit strong interactions, or in other words those subject to the strong force. The proton and the neutron are members of this class, and indeed it is the strong force that binds them in an atomic nucleus. The existence of many other strongly interacting particles has been inferred from the decay products of collisions in accelerators. Most such particles live for an extremely short time, as short as 10^{-24} second, before they decay into other particles. All particles that are subject to the strong force are called hadrons, from the Greek adjective hadros, meaning robust or heavily built.

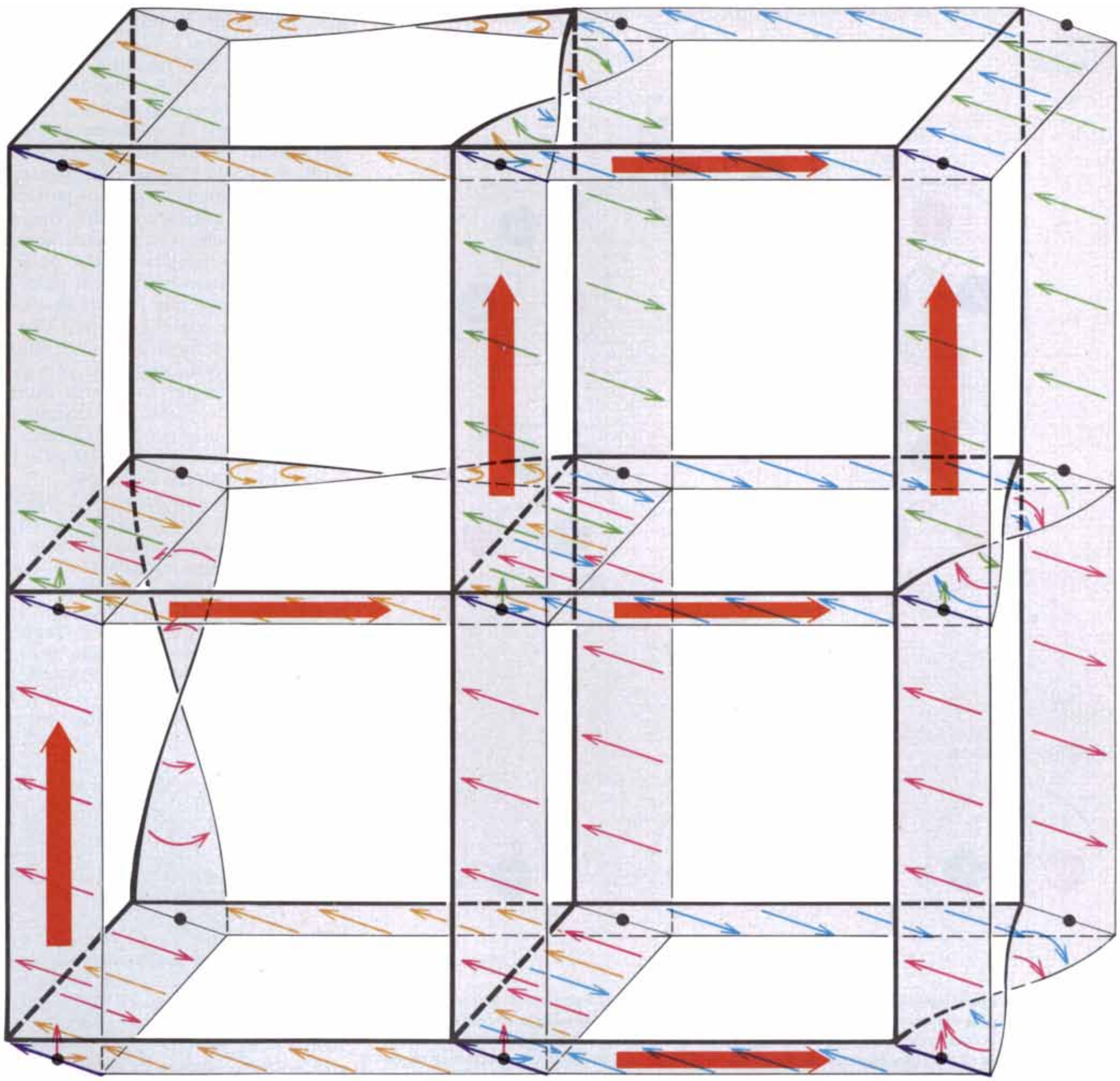
In 1962 Murray Gell-Mann of the California Institute of Technology and Yuval Ne'eman of Tel-Aviv University proposed a scheme for classifying the hadrons in symmetrical patterns. The scheme was based on the mathematical theory of groups and was called the eightfold way. A short time afterward Gell-Mann and, independently, George Zweig, also of Cal Tech, proposed a physical interpretation of the eightfold way. The mathematical classification could be explained by assuming that all hadrons are built up of more fundamental constituents, which Gell-Mann called quarks.

At the time every known hadron could be understood as some combination of three basic quarks (and their corresponding antiquarks): the up or

© 1983 SCIENTIFIC AMERICAN, INC

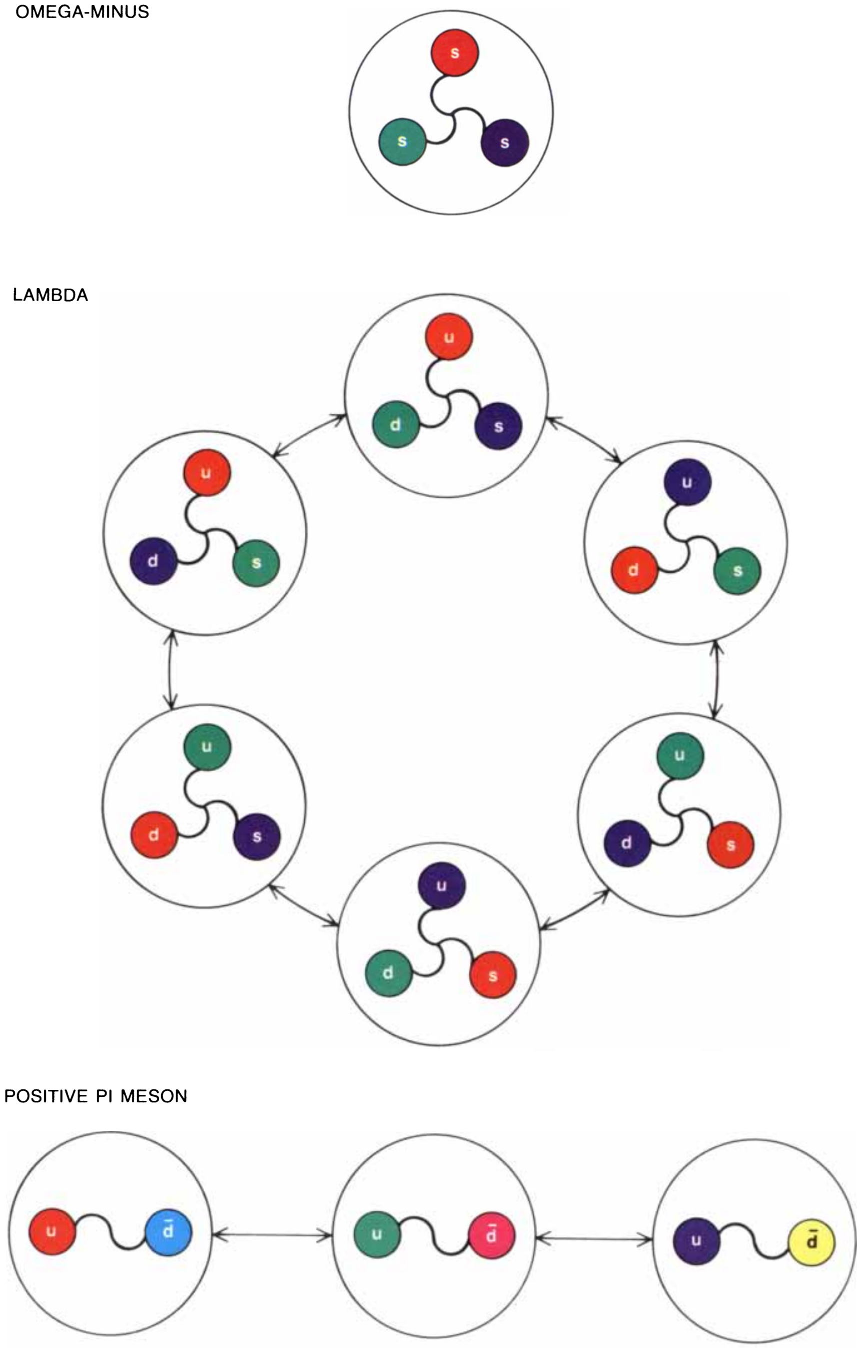
u quark, the down or *d* quark and the strange or *s* quark. The proton, for example, is a combination of two *u* quarks and a *d* quark, whereas the neutron is a combination of a *u* quark and two *d* quarks. The positively charged pi meson is a combination of a *u* quark and a *d* antiquark. Since the quark hypothesis was put forward more hadrons have been discovered and it has become necessary to add at least two more quarks, the charm or *c* quark and the bottom or *b* quark, to the catalogue of elementary particles. Nevertheless, the quark model remains a highly successful classification scheme: more than 100 hadrons are known, and they can all be described in terms of the quark model.

In spite of the success of the model in classifying hadrons, certain features ascribed to the quarks initially made the



LATTICE OF POINTS is a schematic way of representing space and time; with such a lattice physicists seek to understand the field of force that gives rise to the permanent confinement of quarks. Quarks and antiquarks, which are thought to be the constituents of the neutron and the proton and all other particles subject to the strong nuclear force, have never been observed in isolation. Instead they are bound together in triplets or pairs. The field responsible for the confinement is called the chromoelectric field, and it is classified mathematically as a gauge field. A gauge is analogous to a ruler or a pointer; it is used to compare physical quantities defined at different lattice sites. The length of the ruler and the direction of the pointer can change as they are moved about in space-time. Particles such as quarks lie only at the points, or vertexes, of the lattice. In the illustration the variables that specify the state of a particle are represented by purple arrows, which must point in one of two directions at each vertex in the front lattice plane. To compare the orientations of two purple arrows one of them must be moved next to the other. A set of rules, represented in the illustration by the bands that join neighboring vertexes, must be defined in order to specify the changes made in the orientation of the arrows during transport. The set of rules is a gauge field; it enables an arrow to be transported upward or to the right of a vertex, around a square array of four vertexes called a plaquette and back to its starting point. (The transported arrows are colored green, yellow, blue or pink.) Any set of bands that returns the transported arrow to its original orientation represents the lowest energy state of the gauge field. A set of bands that resembles a Möbius strip, however, cannot be untwisted: a transported arrow returns to its starting vertex with the opposite orientation. The energy of the gauge field defined on the lattice is stored in such plaquettes. The red arrows represent the lines of force to which the gauge field gives rise.

© 1983 SCIENTIFIC AMERICAN, INC



COLOR is a quantum-mechanical property of quarks that was introduced to reconcile the quark model with the exclusion principle of Wolfgang Pauli. The principle states that no two quarks within a small region of space can occupy the same quantum-mechanical state. Before the introduction of the color hypothesis the quark model seemed to predict the existence of particles in which the principle is violated. For example, the constituents of the omega-minus particle are three strange or *s* quarks that were thought to be in the same quantum-mechanical state. The paradox is resolved by assuming that each of the three *s* quarks takes on one of three colors, such as red, purple and green. Since the quarks differ in color, they are in different quantum-mechanical states and the exclusion principle is saved. Color is never observed in isolation, and so the colored quarks must be combined in such a way that the omega-minus particle is colorless (or white). The color hypothesis states further that all possible combinations of the quark colors are equally likely, as long as the composite particle remains colorless. For example, the lambda particle, which is made up of three quarks, can occupy any one of six colorless states with equal probability, namely the six permutations of the three colors red, purple and green. The positive pi meson, which is made up of a quark and an antiquark, can occupy any of three colorless states with equal probability: red and cyan (antired), green and magenta (antigreen) or purple and yellow (antipurple). The confinement of quarks and of color is represented here by showing the quarks linked to one another by an unbreakable string. The quarks can move about almost freely inside a particle as long as they stay within the bounds of the string.

physical reality of quarks difficult to accept. The most fundamental problem is the failure to detect a free quark. The proton and the neutron are strongly bound in the atomic nucleus, yet given enough energy in a nuclear collision they can be set free. Any theory that describes the interactions of quarks, however, not only must account for their binding into hadrons but also must lead to permanent confinement.

Almost equally unsettling was the fact that in certain hadrons the quark constituents seemed to violate a fundamental principle of quantum mechanics, namely the exclusion principle of Wolfgang Pauli. The exclusion principle applies to a broad category of particles including the quarks and states that no two such particles within a small region of space can simultaneously occupy the same quantum-mechanical state. In practice the principle implies that two quarks of the same kind, say two *u* quarks, cannot form a hadron unless they have opposite spins. The spin of a quark is like the spin of the earth, except that the quark's spin is quantized: it can assume only one of two values. Hence in any group of three quarks there must be at least two with the same spin.

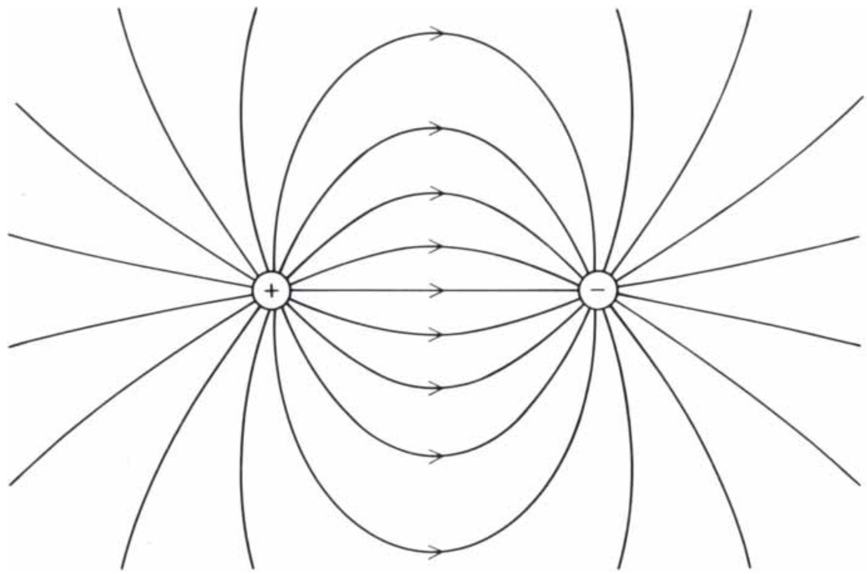
There are several hadrons in which three identical quarks must approach one another closely enough to bind together. The omega-minus particle is one such hadron. It was predicted by the quark model, and its subsequent discovery in 1964 by Nicholas P. Samios, Ralph P. Shutt and their collaborators at Brookhaven gave strong support to the model. On the other hand, the omega-minus was also quite puzzling because the quark model predicts it must be made up of three *s* quarks, whose close association seemed to violate the exclusion principle. For these reasons and others physicists initially preferred to regard the quark as a mathematical convenience; the question of its physical existence was temporarily set aside.

The quark model and the exclusion principle were reconciled as a result of ideas developed by Oscar W. Greenberg of the University of Maryland at College Park and, independently, by Moo-Young Han of Duke University and Yoichiro Nambu of the University of Chicago. What is needed is to assume that each kind of quark can exist in any of three states. For example, if an *s* quark in state *A* is combined with an *s* quark in state *B* and an *s* quark in state *C* to form the omega-minus particle, the exclusion principle is saved. In order to label the states of a quark physicists have whimsically taken to calling them by the names of colors: a quark can come in the three colors red, purple and green, and an antiquark can come in the three complementary colors cyan

© 1983 SCIENTIFIC AMERICAN, INC

(antired), yellow (antipurple) and magenta (antigreen). The prefix "chromo" in quantum chromodynamics refers to the color terminology.

The introduction of color had to be supplemented by another hypothesis if the successful classification of hadrons was to be maintained. Although the new color degree of freedom made possible a quark model of particles such as the omega-minus, it also led to a multiplication of other hadrons. The lambda particle, for example, is made up of a *u*, a *d* and an *s* quark; if each quark can exist in any of three colors, it would seem there should be nine lambda particles, one for each color combination, rather than the single particle that is observed. To avoid such redundancy one adds the hypothesis that the quarks in a hadron can assume only those combinations of colors that leave the hadron colorless, or white, if the rules of color addition (with ordinary light) are assumed. The three quarks in a proton or a lambda particle must include one red, one purple and one green, whereas the quark and the antiquark in a pi meson can be red and cyan, purple and yellow or green and magenta. Because the "total" color is always the same, in the quantum-mechanical sense that each colorless state can occur with equal probability, there is effectively only one lambda particle and only one positive pi meson.

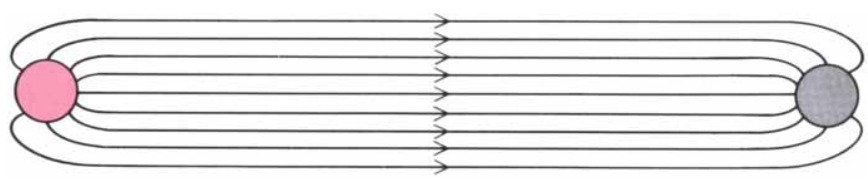


LINES OF FORCE of the electromagnetic field spread out in space. The field shown is the one generated by two particles of opposite electric charge; although the lines of force are densest in the region between the particles, they extend in other directions as well. The intensity of the field at any point (that is, the strength of the force "felt" at a given point by a unit electric charge) is proportional to the number of lines crossing a surface of unit area, orthogonal to the lines of force, that passes through the point. The electromagnetic force that is generated by a single source of electric charge diminishes as the square of the distance from the source.

In the late 1960's strong evidence that the quarks in hadrons are real particles instead of mere mathematical entities came from a variety of experimental results. Of notable importance was a series of experiments done at the Stanford Linear Accelerator Center (SLAC) by Jerome I. Friedman and Henry W. Kendall of the Massachusetts Institute of Technology and Richard E. Taylor of SLAC. High-energy electrons were directed against a fixed target of protons in order to probe the protons for internal structure. By examining the decay products of the collisions it was possible to show that inside the proton there are constituents with all the properties attributed to quarks.

Moreover, although no free quarks were detected, the experiments showed that within the proton the quarks are in a nearly free state of motion. This result was quite puzzling: how could forces strong enough to keep quarks permanently bound together also allow them to move about almost freely when they are at close quarters inside a proton? The three quarklike objects in the proton appeared bound to one another like the three stones tied together in a bola, the South American hunting device. The stones in the bola move freely as long as they remain within the limits of the connecting string; the string, however, keeps them from flying apart.

The new experimental evidence for quarks combined with the introduction of color gave strong impetus to the formulation of a theory of quark dynamics. Color could serve as a source for a new field called the chromoelectric field, which would give rise to a new kind of interaction among colored particles. In 1973 H. David Politzer of Cal Tech and, independently, David Gross of Princeton University and Frank Wilczek of the University of California at Santa Barbara realized that a dynamic interaction based on the chromoelectric field would lead to a progressively weaker force between quarks as they approached one another. The prediction explained the almost free motion of quarks inside protons that had been observed in the SLAC experiments. It was then conjectured that the same interaction could be responsible for the confinement of quarks, although at the time there were no theoretical results to support such an appealing idea. Nevertheless, once the chromoelectric field is introduced the rather ad hoc assumption that all hadrons are colorless and the observed confinement of quarks can be understood as two aspects of the same phenomenon. If one quark, say a red one, is pulled away from a hadron, both the quark and the fragment left behind are colored. If color, like electric charge, is the source of a field, there could be an attractive force between the two colored fragments. Confinement of the quarks might then result if the attraction between the two



COMPRESSION of the lines of force between two particles into a thin tube of uniform cross section would make the force between the particles constant, regardless of the distance between them. A surface of unit area orthogonal to the tube would always meet the same number of lines of force, no matter where the surface was placed along the tube. Because the force that binds the particles remains constant, increasing the separation of the particles by a given increment would always require the same amount of energy, no matter how far apart they were at the outset. An infinite amount of energy would be needed to free one of the particles from the other. Such compression of the lines of force would therefore lead to the permanent confinement of the two particles. If the radius of the tube is negligible, the bundle of lines of force resembles a string. A model explaining the confinement of quarks by this principle was formulated mathematically in 1968 by Gabriele Veneziano of the European Organization for Nuclear Research (CERN) and interpreted as a string by Yoichiro Nambu of the University of Chicago. The idea that the string could be physically realized as a bundle of lines of force was proposed in 1973 by Holger B. Nielsen and Paul Olesen of the Niels Bohr Institute in Copenhagen.

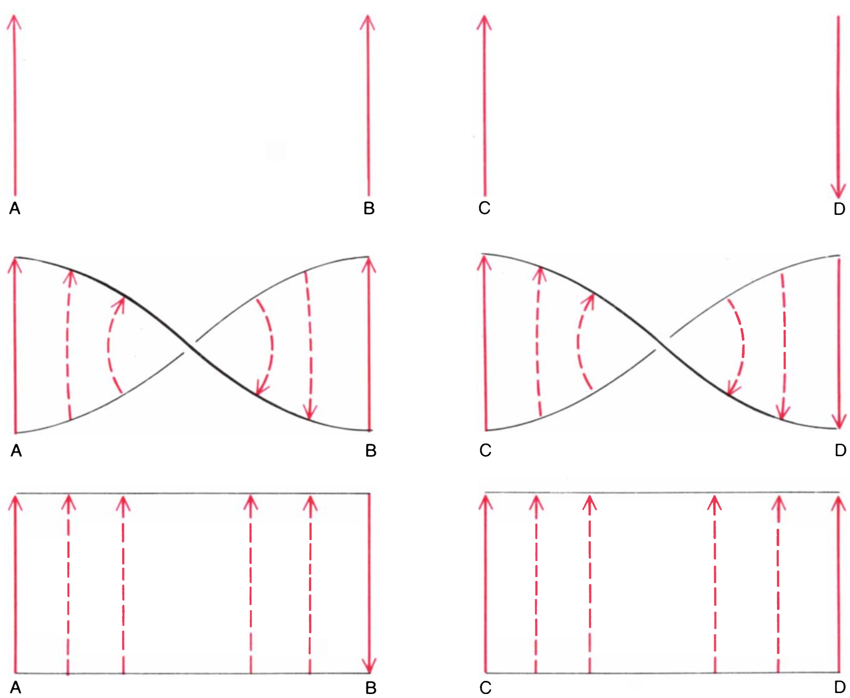
© 1983 SCIENTIFIC AMERICAN, INC

fragments is so strong that it is impossible to separate them beyond some limit.

In the early 1970's a dynamic model independent of the quark model was developed to account for certain properties of hadrons that had not been explained by quarks. According to the dynamic model, the hadron is not a point or a spherical particle but can better be understood as a string. The string can rotate or vibrate in ways prescribed by the laws of relativistic dynamics, and its end points are required to move at the speed of light. Calculations showed that the force acting along the string must be enormous: about 14 tons. The quantized vibrations of the taut string give rise to various states that could be identified with certain hadrons.

It is evident that the string model of the hadron and the bola analogy can be combined. If three quarks or a quark and an antiquark are placed at the end points of the relativistic string, the tension of the string could explain the permanent binding of the particles. The string model, however, like the original concept of the quark, is itself a mere mathematical abstraction: the string is a one-dimensional object. Could it nonetheless be an approximation of some other structure that is more acceptable from a physical point of view?

In 1973 Holger B. Nielsen and Paul Olesen of the Niels Bohr Institute in Copenhagen pointed out that the string could be interpreted as a bundle of lines of force of a suitable field. In an electromagnetic field, lines of force have the familiar patterns made by iron filings on a sheet of paper held over a magnet. The intensity of the field is proportional to the density of the lines of force. Thus when the lines of force spread out, as they do at points increasingly distant from the poles of an ordinary magnet, the intensity of the field diminishes. If the lines of force are squeezed into a tube of uniform cross section, however, the intensity of the field remains constant all along the tube. The force needed to separate a quark and an antiquark at opposite ends of such a tube would also remain constant no matter how far apart the two particles were placed. In order to liberate one of the quarks an infinite amount of energy would have to be supplied.



DIRECTIONS OF TWO ARROWS defined only at the vertexes of a lattice cannot be compared unless there is some way to transport the arrows from one vertex to the other. In the diagram at the top the arrows at vertexes *A* and *B* seem to point in the same direction, whereas the arrows at *C* and *D* seem to point in opposite directions. The comparison, however, presupposes the existence of the paper on which the diagram is drawn, which acts as an intermediary along which one arrow can be moved next to the other. One might just as well assume that the points in each pair are connected by a half-twisted ribbon of paper, as in the middle diagram; the comparison then gives the contrary result. Without a gauge field to supply a rule of transport from point to point the orientations of the arrows at adjacent points cannot be compared. The diagram at the bottom shows that the result of a comparison need not change when the direction of an arrow is reversed at any point; the comparison remains valid if the ribbon is correspondingly twisted or untwisted. The property of the gauge field that makes it possible to compensate for a change in the direction of any individual arrow is called local gauge invariance.

The quantum-mechanical reality of quarks and strings requires that the lines of force associated with the color interaction of quarks act quite differently from the lines of force associated with the electromagnetic interaction of electrically charged particles. Because both forces propagate in the vacuum, one might assume that any differences between them would be caused by the intrinsic nature of the forces themselves and not by the interaction of the forces with the vacuum. In classical, or Newtonian, mechanics the assumption would be sound; indeed, there can be no interaction between a field and the classical vacuum because the classical vacuum is by definition a state with no matter and no energy in it. In quantum mechanics, however, even the vacuum has a structure, which can alter the propagation of fields and forces.

The structure of the vacuum is a consequence of the uncertainty principle of Werner Heisenberg. One version of the uncertainty principle states that for any physical event there is an uncertainty about the energy released during the event that is related to an uncertainty about the exact time of its occurrence. More precisely, the product of the uncertainty about the energy and the uncertainty about the time is not less than some numerical constant. For an event confined to an extremely short interval there is a correspondingly large uncertainty about its energy. During any short interval, therefore, there is a substantial probability that the quantum-mechanical vacuum has some nonzero energy.

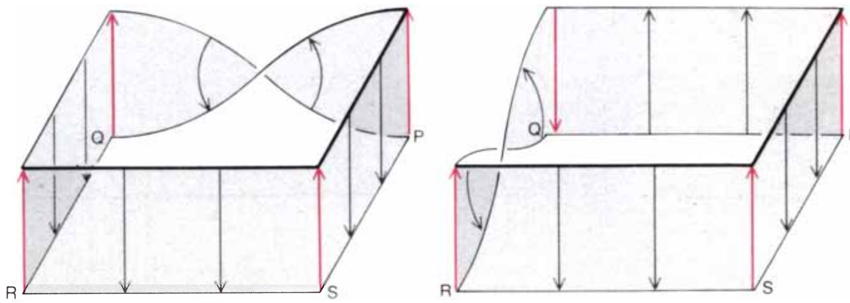
The energy of the vacuum can manifest itself in the spontaneous creation or annihilation of a particle and its antiparticle or in the appearance and disappearance of an electric or a chromoelectric field throughout various regions of space. Such variations of a quantum field are called fluctuations. In the electromagnetic field between two electrically charged particles, for example, the presence of quantum fluctuations implies that the interactions between the two charges are not strictly determined by the classical field predicted by Maxwell's equations. Instead the measured electromagnetic field is the average of all the fields that can be generated by the quantum fluctuations, weighted according to the probability that a given fluctuation will occur.

For most practical applications of electrodynamics the effects of the quantum fluctuations are very small. A measurement of the field between two macroscopic electrically charged objects would be consistent with the value predicted by the classical theory. In high-energy collisions of charged particles, however, the quantum-mechanical fluctuations become much more important

© 1983 SCIENTIFIC AMERICAN, INC

and must be taken into account to calculate the electromagnetic effects. In the standard method (which for many problems is quite successful) one first calculates the properties of the field in the classical vacuum. One then builds on the result of the classical calculation by correcting it for quantum-mechanical fluctuations of a progressively higher complexity, according to what is called a perturbative expansion. In the quantum theory of electromagnetism the larger, or the more complex, a fluctuation, the less the probability that it will take place. Hence almost all the corrections to the classical electromagnetic field that must be made in the quantum-mechanical calculations are the outcome of small fluctuations.

One might suppose the properties of the chromoelectric field between two quarks could be deduced in a closely analogous way. It would seem, at least in principle, that a perturbative expansion could give the strength of the field at any point to any degree of accuracy needed. It turns out, however, that the method of perturbative expansion works only if the field calculated for the classical vacuum is the dominant effect. In other words, the method works only if the corrections that must be made to take account of the fluctuations are small and become smaller still as fluctuations of increasing size are considered. For quantum-mechanical phenomena that depend primarily on the effects of large fluctuations the perturbative expansion does not converge, that is, the series of calculations does not approach a constant, finite value. Such phenomena are said to be nonperturbative. The confinement of the lines of force of the chromoelectric field between two quarks, and hence the permanent binding of the two quarks, is a nonperturbative phenomenon.



FRUSTRATED PLAQUETTE is given a unit twist by the four gauge fields that make up its sides. The gauge fields, represented by ribbons, can be twisted or untwisted locally to compensate for the reversal of an arrow at any point, but the overall twist in the four ribbons of the plaquette cannot be removed. The twist between *P* and *Q* (*left*) can be removed by reversing the direction of the arrow at *Q*. The effect, however, is merely to transfer the twist to the part of the ribbon between *Q* and *R* (*right*). No matter which arrows are reversed, or in other words no matter how many local gauge transformations are performed, an odd number of the four ribbons must stay twisted. Hence the plaquette remains frustrated in spite of local gauge invariance.

How, then, can confinement be demonstrated? Workers in theoretical physics recognized that a new approach had to be devised, in which large fluctuations of the quantum-mechanical field are considered at the outset of the calculations. The lattice method suggested by Wilson is such an approach.

The lattice is generally a cubic one and can be thought of as the edges and vertexes of a collection of densely stacked cubes. The lattice extends in time as well as in space, that is, each point on the lattice designates both a position in space and a moment in time. To visualize the lattice one can think of an array of cubes in which two of the axes are labeled with spatial coordinates and the third axis is labeled with temporal coordinates; the full lattice has three spatial dimensions as well as the time axis, and so it is a four-dimensional structure. Between any two neighboring vertexes of the lattice there is a link, which can be pictured as a line connecting the two vertexes. A small square bounded by four links is called a plaquette. In Wilson's formulation the vertexes, links and plaquettes of the lattice are all that is left of ordinary physical space and time.

Links and plaquettes on the lattice must be regarded as entities at a different level of abstraction from the vertexes, or lattice points. Although links and plaquettes are defined by lattice points, there are no additional lattice points along a link or within a plaquette. In other words, space and time on the lattice are quite unlike ordinary space and time, which always include an infinite number of points between any two given points.

Wilson's introduction of a space-time lattice is not meant to imply that physical processes really take place on a lattice. Space-time, according to all current evidence, is continuous. Instead the lattice represents what theoretical physicists call a regularization, a temporary artifact for making calculations that would otherwise be impossible. Applied to the problem of confinement, the strategy is as follows. All particles are defined only at the vertexes of the lattice, and the strength of the field is defined only along the links of the lattice. (Actually what is defined at each vertex is the probability that a particle will be found there. The probability of finding a particle between two adjacent vertexes is not defined.) When no particles are present, the symmetry of the fluctuations on the lattice dramatically simplifies the calculation of the average electric or chromoelectric field generated by strong fluctuations. The fields, which are vector quantities and therefore have both magnitude and orientation, are just as likely to point in one direction along a link as they are to point in the opposite one. Hence in the vacuum state, with no particles, the mean value of the electric or the chromoelectric field throughout the lattice is zero.

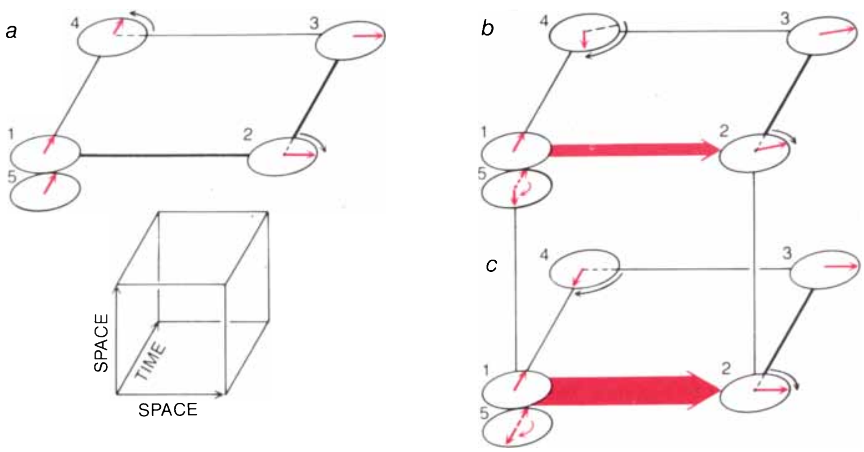
A similar although slightly more elaborate calculation can be done for large fluctuations of the field when a single particle and its corresponding antiparticle are defined on the lattice. On the average the fluctuations of the field again cancel except along the links of the lattice that make up the shortest path between the particle and the antiparticle. The results do not depend on the kind of field defined on the lattice; the particles can be a quark and its antiquark or an electron and a positron. Thus confinement is a natural outcome of defining the field on the lattice.

The next step in the strategy is to remove the lattice and regain ordinary, continuous space-time. The lattice spacing is made progressively smaller so that the vertexes of the lattice become closer and denser in space-time. If the reduction in the lattice spacing proceeds to the mathematical limit, continuous space-time is recovered. At the limit the procedure gives the average field after all the quantum-mechanical fluctuations have been taken into account.

What is gained by introducing the lattice? The strong fluctuations that must be responsible for squeezing together the lines of the color force are considered from the outset along each lattice link. On the other hand, there is a price that must be paid: as long as the lattice is relatively coarse the quantum fluctuations give rise to the confinement of the electromagnetic field as well as the chromoelectric field.

It is well known that electric charges, unlike color charges, do exist in isolation. Since the lattice method predicts the confinement of electric charge, one must be skeptical about the lattice approach unless it can be shown that as the mesh of the lattice is made finer the elec-

© 1983 SCIENTIFIC AMERICAN, INC



GAUGE-FIELD CONCEPT can be generalized by allowing the quantities defined at each lattice point to vary over a continuous range. For example, the direction of an arrow at each point could be allowed to form an arbitrary angle with the vertical. A gauge field makes it possible to compare the angles at different lattice points. If an arrow does not return to its original orientation after a complete circuit of a plaquette, the angular difference between the initial and the final orientations (a quantity called the phase angle) measures the degree of frustration of the plaquette. The electric field is such a gauge field, and the strength of the field in a given direction at any point is measured by the degree of frustration of an associated plaquette. The illustration shows how the frustration of plaquettes is associated with a single component of the electric field, namely the spatial component in the left-right direction. Strictly speaking, the lattice extends in four dimensions, the three spatial ones and time, but one can visualize the lattice as having only two spatial coordinates and a temporal one. As a dial is transported counterclockwise around a plaquette that includes the left-right spatial component and the time dimension, the gauge field causes an arrow on the dial to rotate. If the arrow returns to its starting position after the dial makes a full circuit of the plaquette, the strength of the field is zero in the spatial direction and there is no field line (*a*). As the phase angle increases, the strength of the field along the spatial direction increases as well (*b–c*), which is represented by the increasing thickness of the field lines (*colored arrows*) along the spatial coordinates. Mathematically the rotation of the arrow on the dial can be identified with the rotation of an arrow in the plane of complex numbers, that is, numbers that have both a real part and an imaginary part.

tromagnetic field begins to act in accordance with its well-established properties. In other words, what one would like to show is that at some stage in the shrinking of the lattice the electromagnetic lines of force break out of their confinement to a line of lattice links, whereas the chromoelectric lines of force remain squeezed together all the way to the limit of continuous space-time. It is not a straightforward matter to demonstrate that things happen in precisely this way, but in the past few years the demonstration has been achieved by making elaborate numerical calculations with the aid of high-speed computers.

Theoretical physicists express the fundamental difference between the electromagnetic field and the chromoelectric field by saying that QED is an Abelian gauge theory whereas QCD is a non-Abelian gauge theory. The terms refer to the Norwegian mathematician Niels Henrik Abel. The distinction between Abelian and non-Abelian is drawn from the mathematical theory of groups, which describes the symmetries inherent in a sequence of operations, such as a sequence of rotations. If the operations that are members of a group can be carried out in any sequence with the same final result, the group is Abelian. For example, the group of rotations about a single axis is Abelian, because such operations have the same effect regardless of their sequence. On the other hand, if the sequence in which two or more operations are carried out does affect the final outcome, the group of operations is a non-Abelian one. The rotations of a cube about its three axes form a non-Abelian group: when the cube is turned about a vertical axis and a horizontal one, the result depends on which operation is done first.

In order to understand how ideas from group theory apply to QCD and QED, one must understand the concept of a gauge field. The concept can best be illustrated for isolated points of space and time arranged in a lattice. Particles can rest on any of the lattice vertexes or hop from one vertex to another; as the particles move through the space-time lattice they can change their state, where a state is defined by quantities that can vary over a certain range of values. I shall make the simplifying assumption that the state of a particle is described by just one variable, which can take on exactly two values. For example, at each vertex of the lattice there might be a variable whose value is either $+1$ or -1, indicating the sign of the electric charge. The two possible values can be represented by an arrow that points either up or down.

In describing how interactions propagate in space and time it is essential to be able to compare the values of variables at neighboring points. To compare the length of two objects in separate regions of space one needs a measuring stick, or gauge, that can be moved next to one object, marked and then moved to the second object, where the comparison is made. Similarly, on the lattice one must be able to compare the orientation of the arrows at neighboring vertexes. At first such a comparison seems trivial. Suppose two vertexes of the lattice are represented on a sheet of paper and at each vertex there is an arrow directed upward. Is it not obvious that the two arrows point in the same direction? The question, however, presupposes the existence of the sheet of paper on which the lattice is drawn. The paper acts as an intermediary that allows the two orientations to be compared. In a sense, one transports the arrow from one vertex to the other with the eye, and one concludes that the two arrows had the same orientation before the transport because after the transport they match.

Suppose the sheet of paper between the two arrows had been given a half twist [*see illustration on page 58*]. An upward-pointing arrow transported along the twisted paper would point downward when it reached the second vertex. Because the direction of an arrow is defined only at the individual, isolated vertexes of the lattice, there is no way to decide which of the two methods of transport is correct. Indeed, without the sheet of paper or some similar assumption about the effects of transport no comparison of directions at different vertexes is possible.

In field theory a gauge is any standard of measurement, analogous to the distance between two marks on a metal bar or the direction of an arrow with respect to a dial, that can change under the influence of a field as the gauge is moved about in space and time. A field that can effect such changes is called a gauge field, and it specifies explicitly the assumptions that must be made about the transport of the gauge. In my simple example the gauge field is a set of rules for transporting the arrows along the links of the lattice from one vertex to the next. One can think of the gauge field as a ribbon that connects neighboring lattice sites and thereby allows the arrows at different points to be transported and compared.

It is important to note that the arrow representing the state at any lattice vertex can be reversed, provided the ribbon representing the gauge field is correspondingly twisted or untwisted. If the arrows and the ribbons are adjusted in

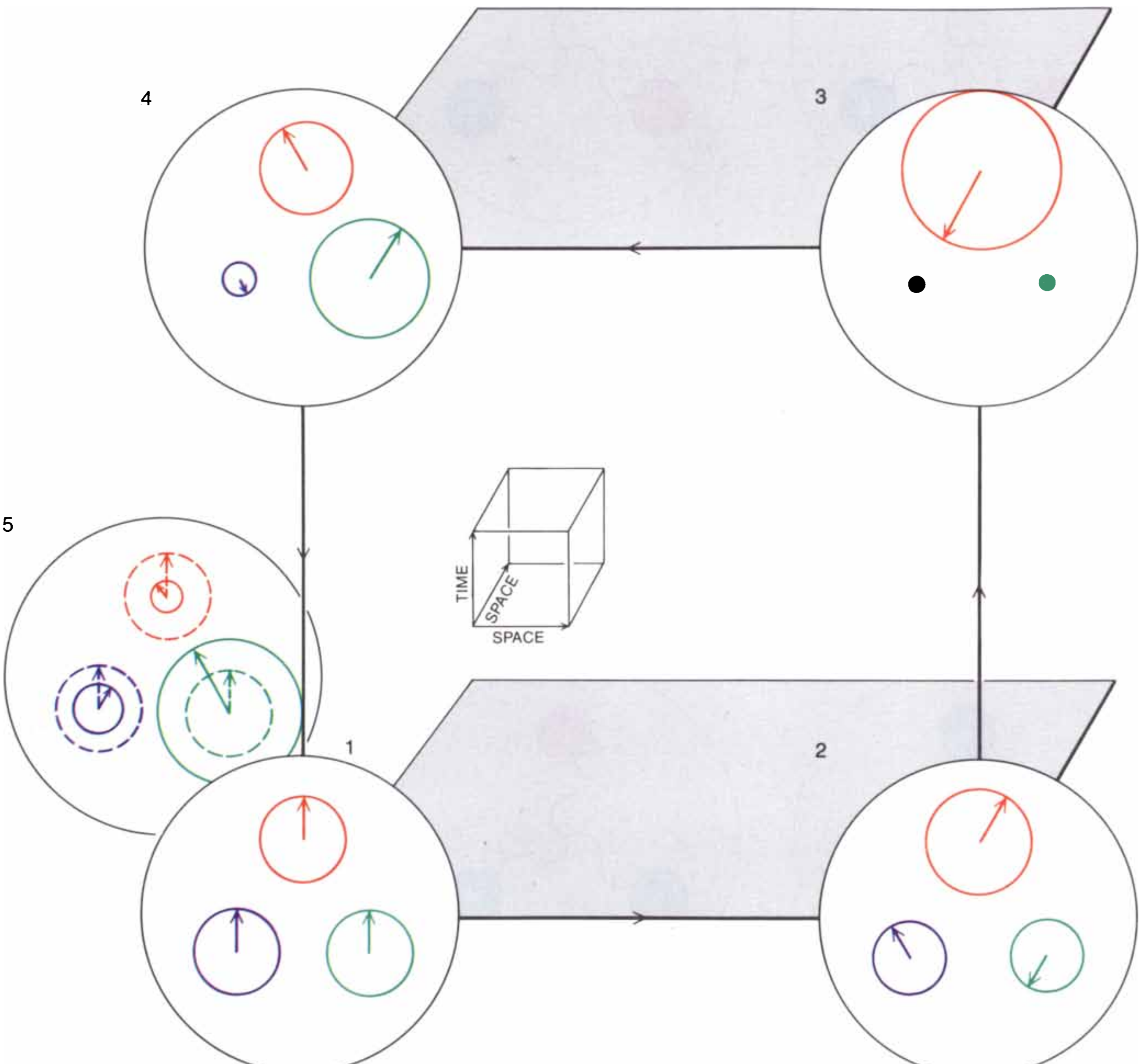
© 1983 SCIENTIFIC AMERICAN, INC

coordination, the result of any comparison between state variables does not change, nor does the physical information represented by the system. The possibility of reversing the arrows without modifying the physical information is called local gauge invariance.

Because of local gauge invariance one might think a gauge field is nothing but an unnecessary complication. What is the point of introducing the twisted ribbon if it can then be untwisted by a local gauge transformation? The objection is valid as long as one considers only pairs of lattice points. The importance of the local gauge invariance becomes apparent, however, when one considers a plaquette on the lattice: a set of four points in a square, connected along the lattice links by four ribbons that constitute the gauge field. Suppose one of the ribbons is twisted and the other three are not [*see illustration on page 59*]. The single twisted ribbon can be untwisted (and one of the arrows can be reversed), but not without introducing a half twist in one of the other three ribbons. No matter which arrows are reversed and no matter how many local gauge transformations are carried out, an odd number of the four ribbons must be twisted. A plaquette whose ribbons cannot all be untwisted is called a frustrated plaquette.

The frustration of the plaquette manifests itself in another way. An upward-pointing arrow that is transported along the ribbons around the frustrated plaquette will return to its starting vertex pointing down. Thus on the lattice the frustrated plaquettes are those that generate a mismatch in orientation when an arrow is transported all the way around.

It is not difficult to apply the idea of a



CHROMOELECTRIC FIELD is a gauge field similar in principle to the electromagnetic field but more complicated mathematically. At every point on a lattice there are three arrows instead of one; they correspond to the color charges of a quark. Moreover, the color gauge field affects not only the direction of each arrow but also its length. The lengths of the arrows are not independent of one another: the square root of the sum of the squares of the lengths must equal 1. The strength of the chromoelectric field along each link of the lattice is dependent on the phase angles and the change in the configuration of the three arrows after a complete circuit of a plaquette.

© 1983 SCIENTIFIC AMERICAN, INC

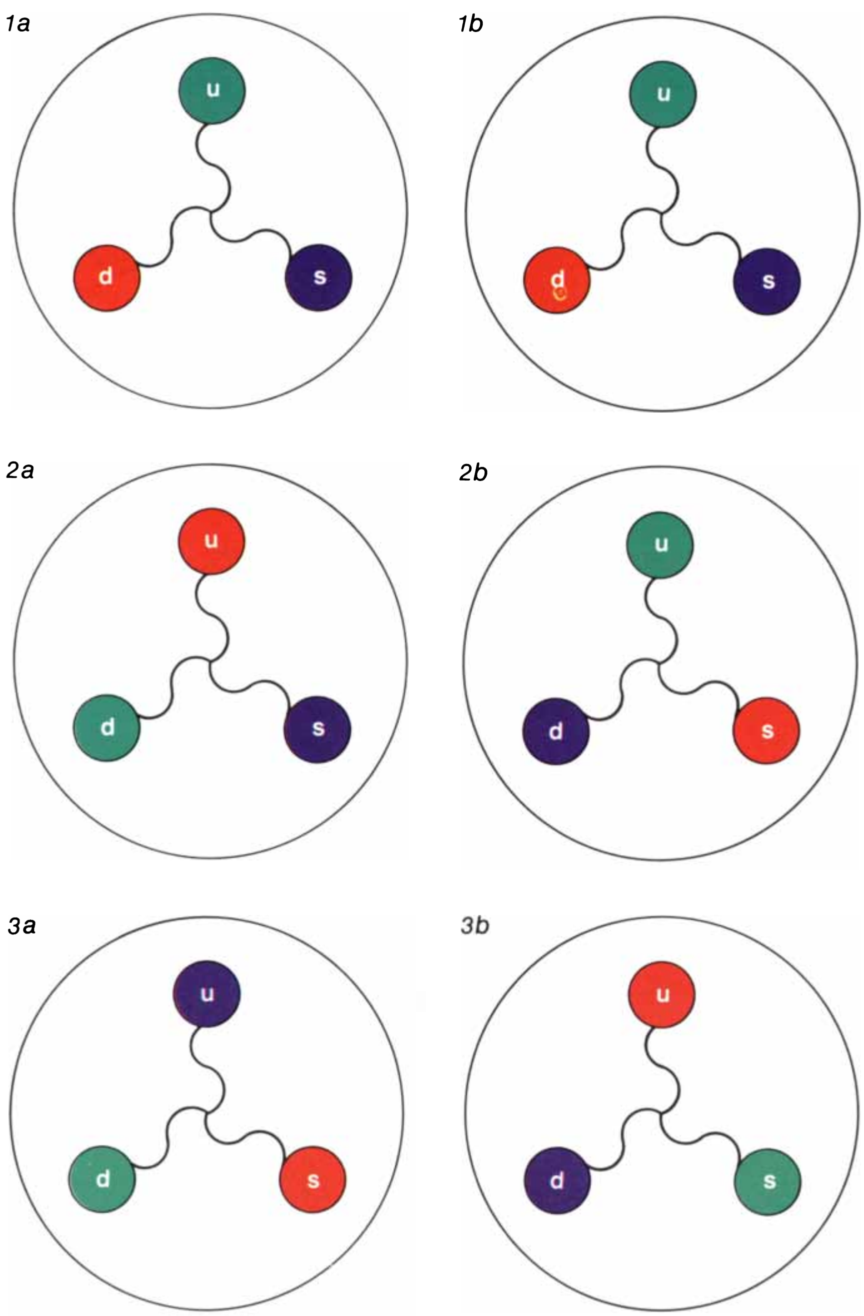
frustrated plaquette to ordinary space-time. In physical space-time the ribbons along which the arrows move cannot be visualized directly; just as the sheets of paper that define the gauge field on the lattice are not themselves part of the lattice, so a gauge field defined in ordinary space-time is not itself part of space-time. Mathematically the higher-dimensional abstract space that specifies the rotations of the arrow is called a connection in a fiber bundle [see "Fiber Bundles and Quantum Theory," by Herbert J. Bernstein and Anthony V. Phillips; SCIENTIFIC AMERICAN, July, 1981]. Nevertheless, it is possible without complete mathematical understanding to imagine that an arrow moved about along some closed path in space-time could return to its starting point with its direction changed.

The idea of a frustrated plaquette has an important physical interpretation. In any gauge field the energy of the field resides precisely in the plaquettes that are frustrated. The plaquettes that are not frustrated, in which all the ribbons can be untwisted and all the arrows can be oriented in the same direction, are associated with the vacuum state of a physical system, the configuration with no energy. Frustrated plaquettes are the sites of the fluctuations in a quantum-mechanical field.

The idea of a gauge field can readily be generalized to more complex situations. The arrows, for example, can be allowed to form an arbitrary angle with respect to a fixed direction rather than being constrained to point up or down. The gauge field now specifies the angle by which the arrow rotates when it is transported along a link. Consequently the frustration of a plaquette can take on a continuous range of values, measured by the angular difference in the direction of the arrow after it is transported around the entire plaquette. The degree of frustration of a plaquette, expressed in suitable mathematical units, is called its action. The action of the entire system is the sum of the actions of all the individual plaquettes.

It turns out the electric field is a gauge field that specifies the continuous rotations of an arrow about a single axis, although this fact is not apparent in any but the most unified and sophisticated formulations of the concept of the electric field. It is more common to think of the electric field as a set of vectors, with one vector at each point in space giving the magnitude and direction of the field at that point. Understood according to the more comprehensive framework of the gauge theory, however, the magnitude of each vector is directly proportional to the frustration of an associated plaquette.

The rotating arrow of the gauge field associated with electromagnetism actually represents a complex number (one with both a real part and an imaginary part), which changes in value as the arrow is carried around the plaquette. (The transported arrow and the complex number must not be confused with the vector that represents the electric field itself at each link on the lattice.) The plaquette is a loop in space-time, and so it is necessary to imagine the transported arrow as moving forward and backward in time as well as in space. The arrow is carried along a certain axis in space, say the positive x axis, then forward in time, then back to its starting point on the x axis and finally backward to its starting point in time. The amount



GROUP OF COLOR TRANSFORMATIONS in the chromoelectric field is classified mathematically as a non-Abelian one. A non-Abelian group is an abstract set of operations (together with the objects on which the operations are carried out) in which the final outcome of several operations is dependent on the sequence in which they are done. For the chromoelectric gauge theory the operations are interchanges of colors among quarks. In the illustration two interchanges of color are applied to a given quark configuration (*1a, 1b*). At the left an exchange of red and green (*2a*) is followed by an exchange of purple and red (*3a*). At the right the order of the two exchanges is reversed (*2b, 3b*). The final configurations of the quarks are not the same.

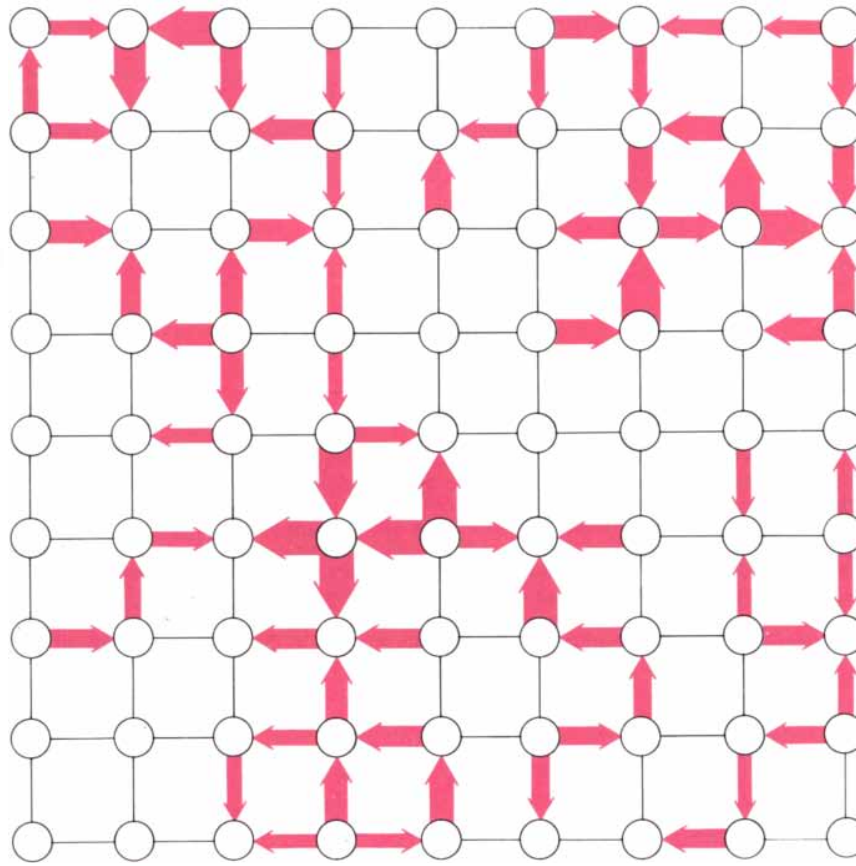
© 1983 SCIENTIFIC AMERICAN, INC

by which the direction of the transported arrow changes as a result of the transport is called the phase angle. According to gauge theory, the magnitude of the x component of the vector that ordinarily gives the strength of the electric field at a point is a measure of the phase angle generated around a space-time loop that begins in the x direction. The quantum-mechanical fluctuations at a point in the electric field can therefore be thought of as fluctuations in the amount of rotation an arrow would undergo if the arrow were transported around a plaquette extended along one spatial dimension and along the temporal dimension [*see illustration on page 60*].

It is difficult to suppress a feeling of unreality when one is asked to regard the electric field—an entity that can become quite tangible when one gets a shock—as the abstract space of phase rotations. Nevertheless, as one goes more deeply into the study of the physical world, tangible facts and mathematical concepts become intertwined. The abstract idea of phase space at last makes contact with the distinction between QED and QCD, that is, with the distinction between an Abelian gauge theory and a non-Abelian one. The fundamental group operations in QED are the rotations of phase angles. The result of two successive phase rotations, say one of 30 degrees and one of 50 degrees, does not depend on the sequence in which they are done. The net result in either case is a rotation by 80 degrees. Such an outcome is characteristic of an Abelian group and QED is therefore an Abelian gauge theory.

If the state variables defined at the vertexes of a lattice are the colors of quarks, the transported gauge is made up of three arrows instead of just one, and the arrows can vary in length as well as in angle [*see illustration on page 61*]. Each arrow represents one of three color charges; the color charges together determine the color of the quark. As the quark is transported along a link of the lattice it can change its color from red to green; hence a red quark can exchange colors with a green quark, and then the newly green quark can exchange colors with a purple quark. The final result of the exchanges, however, depends on the sequence of events. Since an outcome that depends on sequence is characteristic of a non-Abelian group, QCD is a non-Abelian gauge theory.

The non-Abelian nature of QCD introduces an extra degree of freedom into the fluctuations of the chromoelectric field. Moreover, the existence of three kinds of color also implies that the plaquettes of the chromoelectric field can be frustrated in many more ways than the plaquettes of the electromagnetic field can. It is likely that the extra degree of freedom and the additional disorder of the non-Abelian theory are responsible for the confinement of quarks within hadrons.



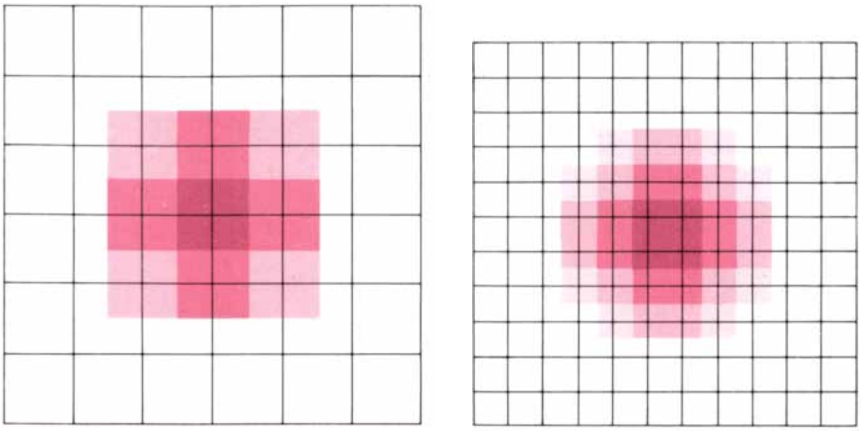
FLUCTUATIONS of the lines of force on a lattice are characteristic of the quantum-mechanical vacuum. The fluctuations are a consequence of the uncertainty principle of Werner Heisenberg. The principle states that the product of the uncertainty about the energy of a system and the uncertainty about the time of an event is not less than a numerical constant. Hence during any short interval there is a correspondingly large uncertainty about the energy of the system, which manifests itself in spontaneous fluctuations of the quantum field. The fluctuations cancel one another in the absence of charged particles, and so a measurement of, say, the electric field in the vacuum would give the value zero because it would determine only the average of all the fluctuations. Nevertheless, the fluctuations of the lines of force in the vacuum must be carefully considered in calculating the field of force between two charged particles.

I have already stated that confinement on a coarse lattice can be demonstrated by a relatively simple calculation for large fluctuations of the quantum-mechanical field. For smaller fluctuations on a finer lattice, however, the calculation of the field becomes much more difficult. The value of any physical quantity measured in an experiment is the quantum-expectation value, which is a weighted average of all possible values the quantity can have. The measured electric or chromoelectric field is the average of all the possible configurations of the field to which fluctuations can give rise. Not all configurations contribute equally to the average, and so each configuration must be weighted, or multiplied by some factor based on the probability of the configuration. In principle, therefore, the quantum-expectation value of a physical quantity defined on the lattice can be calculated in two steps. First the value is calculated for each configuration of the field fluctuations and multiplied by the weight factor for that configuration. Then the products are added and the result is divided by the sum of the weight factors.

Even for a lattice defined over a small volume of space-time, however, the number of possible configurations is so large that a complete summation is out of the question. For the simplest-possible gauge theory, in which the gauge field between two lattice vertexes can be either twisted or untwisted, the number of configurations on a lattice extended for 10 sites in each direction is $2^{40,000}$, or more than $10^{12,041}$.

To calculate the quantum-expectation value of the field on the lattice one must therefore resort to techniques of statistical sampling. The techniques are analogous to the ones employed in opin-

© 1983 SCIENTIFIC AMERICAN, INC



SHRINKING A LATTICE on which particles and fields have been defined must be carried out in order to approximate the effects of the field in ordinary, continuous space and time. As the lattice spacing is reduced, however, the size of a composite particle such as a proton must be preserved, that is, the probability that a quark will be found at a given lattice vertex within the proton must be spread over an increasing number of vertexes. The illustration at the right shows the result of shrinking the lattice at the left to half its previous size. The color density represents the relative probability that a quark will be found at a given vertex. The shrinking can be accomplished by systematically changing the value of the quantity called the coupling constant, which determines the strength of the fluctuations in the quantum-mechanical field.

ion polling. One cannot ask the opinion of every person in the U.S. in order to determine how an issue is perceived by various groups in the population, and so a sample is selected. The result should reflect the actual opinion of the groups if the probability of selecting a respondent from a given group matches the portion, or weight, of the group in the population as a whole.

In a similar way the configurations of the field fluctuations are sampled by a computer program. The computer generates a large number of configurations (but many fewer than $10^{12,041}$), and the probability that a particular configuration is generated is set equal to the quantum-mechanical weight factor for that configuration. The sample of configurations tends to give the same average value for the quantum field as the total population of configurations does.

The importance of a configuration in calculating the quantum-expectation value is determined by its action, which is generally designated S; the weight factor is therefore given by a mathematical function of the action S. An elegant formula for the weight factor in quantum-mechanical systems was introduced by Richard P. Feynman of Cal Tech. First the total action of the configuration is divided by a constant, g^2, then the exponential of this quantity is determined; in other words, the number e (equal to approximately 2.7) is raised to the power S/g^2. The weight factor is inversely proportional to the result. Thus the weight factor is proportional to $1/\exp(S/g^2)$.

The formula for the weight factor implies that the higher the action of a configuration is, the less the configuration is weighted in calculating the average quantum field. Because the action S is not itself exponentiated but instead is divided by g^2, changing the value of g can have a significant effect on the weight factor of a given configuration. If g is large, S/g^2 is small and $1/\exp(S/g^2)$ is larger than it would be if g were small. The quantity g is called the coupling constant; hence when the coupling constant is large, the weight factor of a quantum-mechanical fluctuation with large action is higher than it is when the coupling constant is small.

The idea of a coupling constant may be familiar as a measure of the intrinsic strength of a force. In electromagnetic theory the coupling constant is an important physical quantity with a value of about 1/137. From the rather abstract perspective of the lattice gauge theory, however, g is to be understood as a quantity that takes on a fixed value for a given lattice spacing but can vary with the spacing. Once confinement is demonstrated on a coarse lattice the mesh of the lattice is made finer by carefully decreasing the value of the coupling constant. The process by which the lattice is made progressively finer until continuous space-time is recovered is called renormalization.

Strictly speaking, it is not yet possible to renormalize the lattice for fluctuations of the chromoelectric field, or in other words to shrink the lattice spacing to zero. Nevertheless, it is possible to make the lattice spacing smaller while still keeping it greater than zero and to search for indications that the lines of force do or do not remain confined on finer lattices. As the lattice spacing is reduced all physical objects should remain the same size. On a coarse lattice the probability distribution for a proton might be defined as nonzero across only three lattice spacings. When the lattice spacing is reduced to half its original value, the probability distribution must stretch over six lattice spacings. Since reducing the value of g lowers the probability of lattice configurations that have a large action, the decrease in g has the effect of "zeroing in" for a closer look at the fluctuations of the field, at the lines of force and at the particles defined on the lattice. Thus reducing the value of g makes the proton look larger on the lattice.

The numerical investigation of the properties of a gauge field on the lattice proceeds by limiting the lattice to a large but finite volume. The number of lattice vertexes, links, plaquettes and state variables is therefore finite, although it may be larger than 100,000. Initial values of all the variables are stored in the memory of a large computer. By randomly varying the elements in the starting configuration according to a suitable algorithm the computer generates a sample of as many as 100,000 configurations. Finally it calculates the average quantum-mechanical effects to which the configurations in its sample give rise.

Because of the element of randomness in the calculation, the method is called a Monte Carlo simulation. Before my own work and that of my colleagues on quark confinement the Monte Carlo method had been applied with considerable success to the analysis of the properties of thermodynamic systems, and Wilson had emphasized its suitability for the analysis of lattice gauge theories in quantum mechanics. In 1979, working at Brookhaven, Michael J. Creutz, Laurence A. Jacobs and I first applied Monte Carlo simulation to the study of Abelian gauge theories. We wanted to test whether or not the confinement of particles on the lattice observed at large values of the coupling constant in QED disappears as it should when the continuum limit is approached. The results were spectacular; they showed clearly that at some stage in the reduction of the coupling constant the lines of force on the lattice suddenly undergo a transition. The electric field, which for large values of g is confined to the lattice links between two electric charges, suddenly spreads out all around the charges.

It is useful to compare the sudden deconfinement of the electric field with the sudden change in properties observed when a solid changes to the liquid phase. What Creutz, Jacobs and I observed in our numerical calculations was a phase transition. Everything happened in the computer model as if we had a four-dimensional crystal that we could heat or cool by changing the value of g. At

© 1983 SCIENTIFIC AMERICAN, INC

a certain point the crystal underwent a phase transition; on one side of the transition electric charges are confined, whereas on the other side they are not. Our results were later confirmed independently by an extensive numerical simulation done at the European Organization for Nuclear Research (CERN) by Benny Lautrup of the Niels Bohr Institute and Michael Nauenberg of the University of California at Santa Barbara and by other investigators.

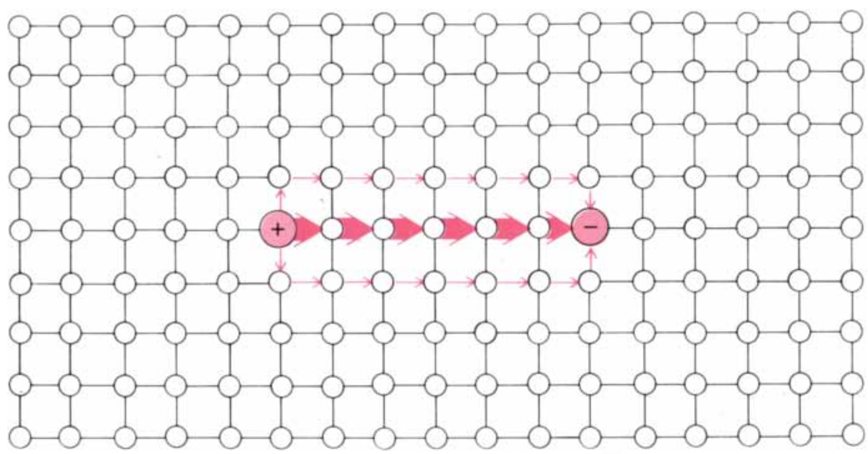
Soon after we got our results for the Abelian lattice gauge theory Creutz extended the Monte Carlo simulation to the non-Abelian model. His results were just as spectacular and physically more interesting; with QED the correct answer was known beforehand and the simulation was a test of the Monte Carlo method rather than of QED, but the outcome for a non-Abelian system was entirely unknown. Creutz's simulation showed that, contrary to the Abelian case, the non-Abelian lattice gauge theory undergoes no phase transition as the value of the coupling constant is gradually lowered. Thus what had long been sought is now finally obtained: a demonstration, albeit by numerical methods, that quantum chromodynamics leads to the confinement of quarks.

Monte Carlo simulation makes it possible to explore the predictions of QCD for many physical processes. For example, after Creutz showed that there is no deconfining phase transition in QCD he was also able to estimate the force holding the quarks together. The result is in excellent agreement with the prediction of the string model, which makes no pretense of describing the dynamics of quarks and hadrons in their full generality. Creutz's value for the string tension was confirmed independently by Wilson, by Gyan Bhanot of CERN, by me and by several others.

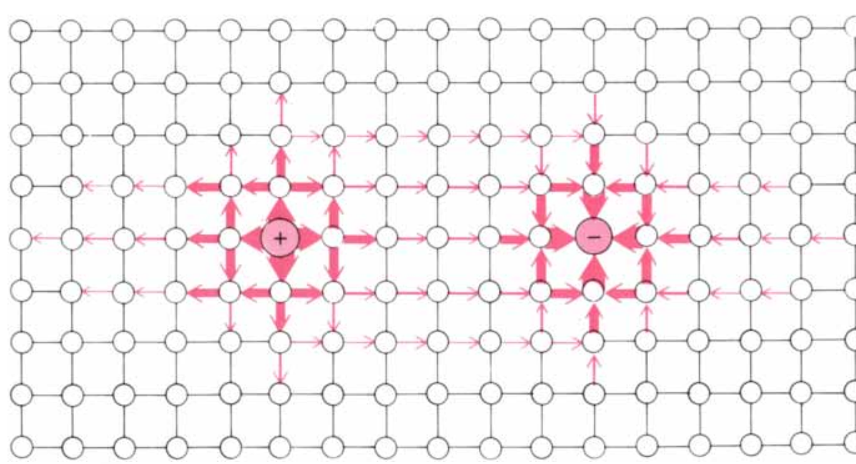
Another prediction of Monte Carlo simulation is the surprising result that at extremely high temperature quarks become deconfined and can move about freely. The temperature must be about two trillion degrees Celsius, much too high to be created in the laboratory but perhaps not too high for free quarks to be found in the interior of hot stars. The prediction also suggests that free quarks existed shortly after the big bang.

The current frontier in the understanding of quark interactions is the numerical simulation of changes in the chromoelectric field as the quarks move about. All the investigations described so far evaluate the chromoelectric field by assuming that the quarks are stationary. It is possible to employ the Monte Carlo method to simulate the quantum mechanics of moving quarks in a gauge field, but the amount of computation required puts the simulation almost out of reach, even for the most powerful computers. Herbert W. Hamber of the Institute for Advanced Study, Enzo Marinari and Georgio Parisi of the University of Rome and I, and independently Donald H. Weingarten of Indiana University, have recently proposed an approximation scheme whereby the computations become feasible. The approximation has been applied to a calculation of the masses of several hadrons, and the theoretical results are in good agreement with experimental values.

Lattice gauge theory has at last brought QCD to the stage where one can calculate its predictions and compare them with experiment. It is an elegant theory, based on relatively simple mathematical concepts yet rich in consequences. For the moment all its verifiable predictions are passing the test of experiment. What remains to be done is to learn how its consequences can be proved by logical deduction. Although a numerical approximation cannot satisfy the need for logical demonstration, it does augur well for such a demonstration in the future. In other branches of theoretical physics there have been many instances in which a definite result, obtained numerically at first, was soon proved by analytic methods. Several of my colleagues and I are confident that QCD will soon pass this test as well.



CONFINEMENT OF A FIELD, whether the field is the electric or the chromoelectric one, can be demonstrated by fairly simple calculations when the field is defined on a coarse lattice and the fluctuations of the lines of force are strong. The strong fluctuations cancel one another everywhere on the lattice except in the region between two charged particles. The lines of force are compressed into a thin tube, and so the two particles cannot be moved indefinitely far apart.



CALCULATION OF THE ELECTRIC FIELD generated by two oppositely charged particles defined on a lattice was done by the author with a Control Data Corporation CDC-7600 computer at the Brookhaven National Laboratory. The illustration shows the outcome of thousands of calculations, done for progressively finer lattice spacings. As the spacing is reduced the lines of force that are confined on the coarse lattice (*see upper illustration on this page*) suddenly undergo a phase transition and break out in all directions. The deconfinement of the electric field on the lattice confirms the plausibility of the lattice approach, because it is well known that the electric field spreads out in ordinary, continuous space. Similar calculations done by Michael J. Creutz of Brookhaven have demonstrated that as the lattice spacing is reduced for the chromoelectric field the lines of force remain confined, just as they are on the coarse lattice. Thus the confinement of quarks can be derived from the theory of the chromoelectric field.

© 1983 SCIENTIFIC AMERICAN, INC

Synthetic Vaccines

A short chain of amino acids assembled in the laboratory to mimic a site on the surface of a viral protein can give rise to antibodies of predetermined specificity that confer immunity against the virus

by Richard A. Lerner

Vaccination is one of man's most significant inventions. It has eliminated smallpox and brought such diseases as diphtheria, poliomyelitis and measles under control. Yet the procedure remains imperfect: the safety of a vaccine cannot be absolutely ensured. Immunization against viral infection is achieved by injecting a virus that elicits antibodies capable of neutralizing the agent of a disease; the injected virus is not itself likely to cause infection because it has been either attenuated or killed. An attenuated virus is one whose virulence in human beings has been reduced by mutation in the course of passage through some different animal host, so that the virus becomes adapted to survival in that animal rather than in human beings. (The cowpox virus, which confers immunity to smallpox, is in effect a naturally attenuated virus.) An attenuated virus is, however, a living organism and therefore a changeable one. It can mutate further, possibly increasing in virulence. Even in the absence of mutation it may possibly have some unknown long-term effect comparable to those of slow-acting "latent" viruses. These possibilities are eliminated if the virus is killed and thereby inactivated, but there have been instances where an incompletely inactivated virus has caused disease.

Other problems are common to both attenuated-virus and killed-virus vaccines. Any facility where viruses are grown to make a vaccine is a reservoir from which an agent of disease can accidentally be disseminated. Moreover, because viruses replicate only in a living system, they are commonly grown in cultured cells or fertilized eggs or are isolated from the blood of infected animals (from infected human beings in the case of the virus causing hepatitis *B*); the cultured cells, eggs or blood may contain undetected substances, notably other viruses, that can contaminate the vaccine. Some vaccines, even some killed-virus ones, must be kept under refrigeration from the time they are prepared until they are inoculated, and maintaining a dependable "cold chain" can be a formidable problem in underdeveloped parts of the world.

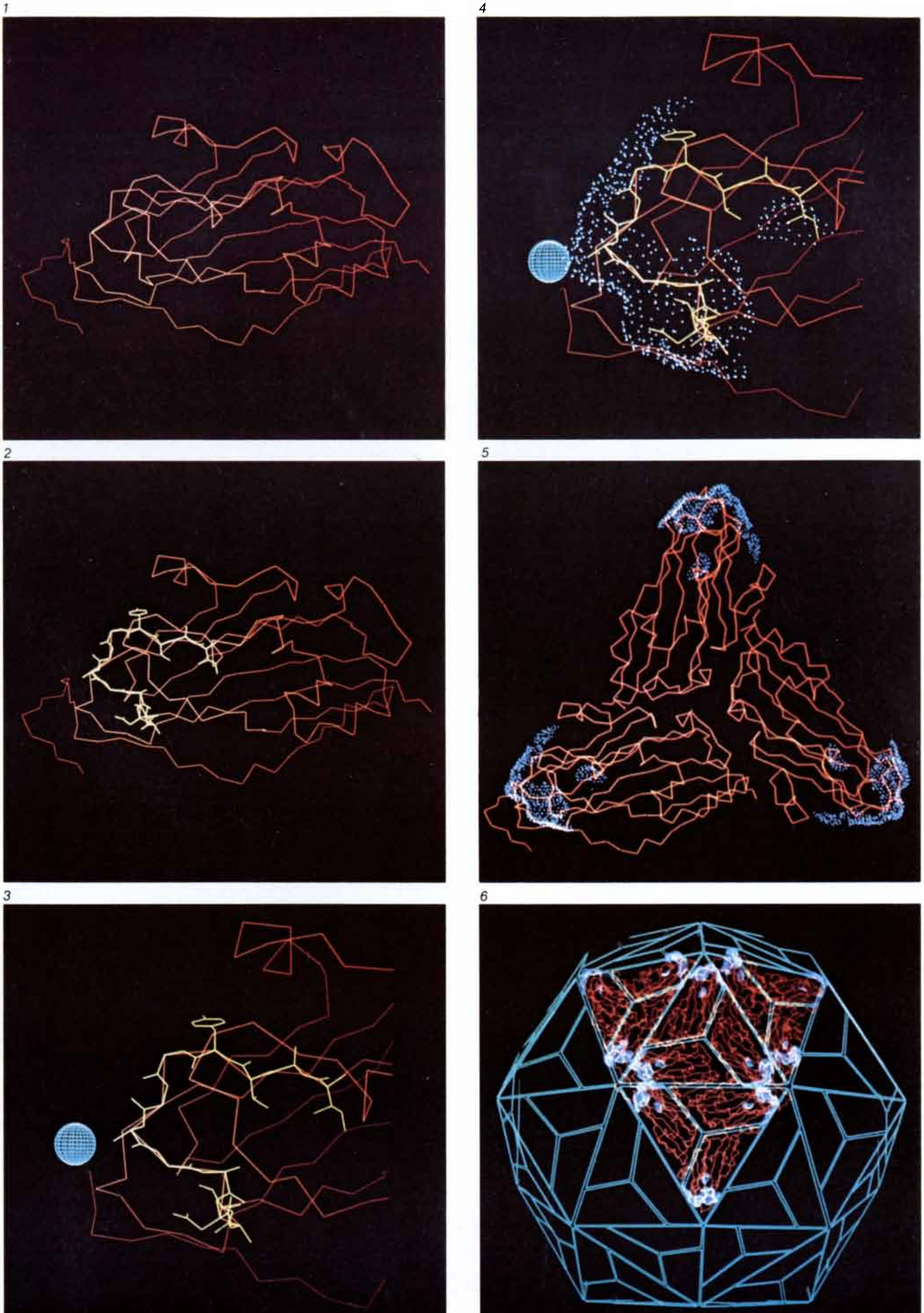
For all these reasons there has been increasing interest in the preparation of synthetic vaccines, which is to say vaccines containing not intact viruses but merely peptides (short protein chains) that have been constructed in the laboratory to mimic a very small region of the virus's outer coat and that can nonetheless give rise to antibodies capable of neutralizing the virus. Experimental vaccines of this kind have now been developed in a number of laboratories, including ours at the Research Institute of Scripps Clinic. In the process much has been learned about the immunological structure of proteins; indeed, the ability to synthesize peptides that elicit antibodies of precise and predetermined specificity may give molecular biologists a battery of tools whose significance may rival the clinical significance of synthetic vaccines.

Before reporting on our experiments and results and on what we have learned about protein structure I should describe the immune system's mode of defense against viral infection. A virus is a small package of genetic information (DNA or RNA) enclosed in a capsid consisting of many copies of a protein or of several different proteins; the capsid may in turn be enveloped in a membrane studded with proteins. Infection begins when a specific site on the surface of the virus binds to a specific receptor on the surface of the cell. The virus penetrates the cell membrane and appropriates the cell's biosynthetic machinery to make many copies of itself, which burst out of the cell and go on to infect more cells. The infective process may be blocked by the immune system if the host has previously been exposed to the same virus or has been immunized against it with a vaccine. In either case certain of the immune-system cells called lymphocytes will have been primed to recognize the virus. When receptors on one of these lymphocytes encounter the virus, an immune response is triggered, one result of which is the proliferation of plasma cells that secrete antibodies against sites on the surface of the virus. The antibodies bind to those sites, coat the surface of the virus and block its attachment to receptors on cells. The virus is neutralized.

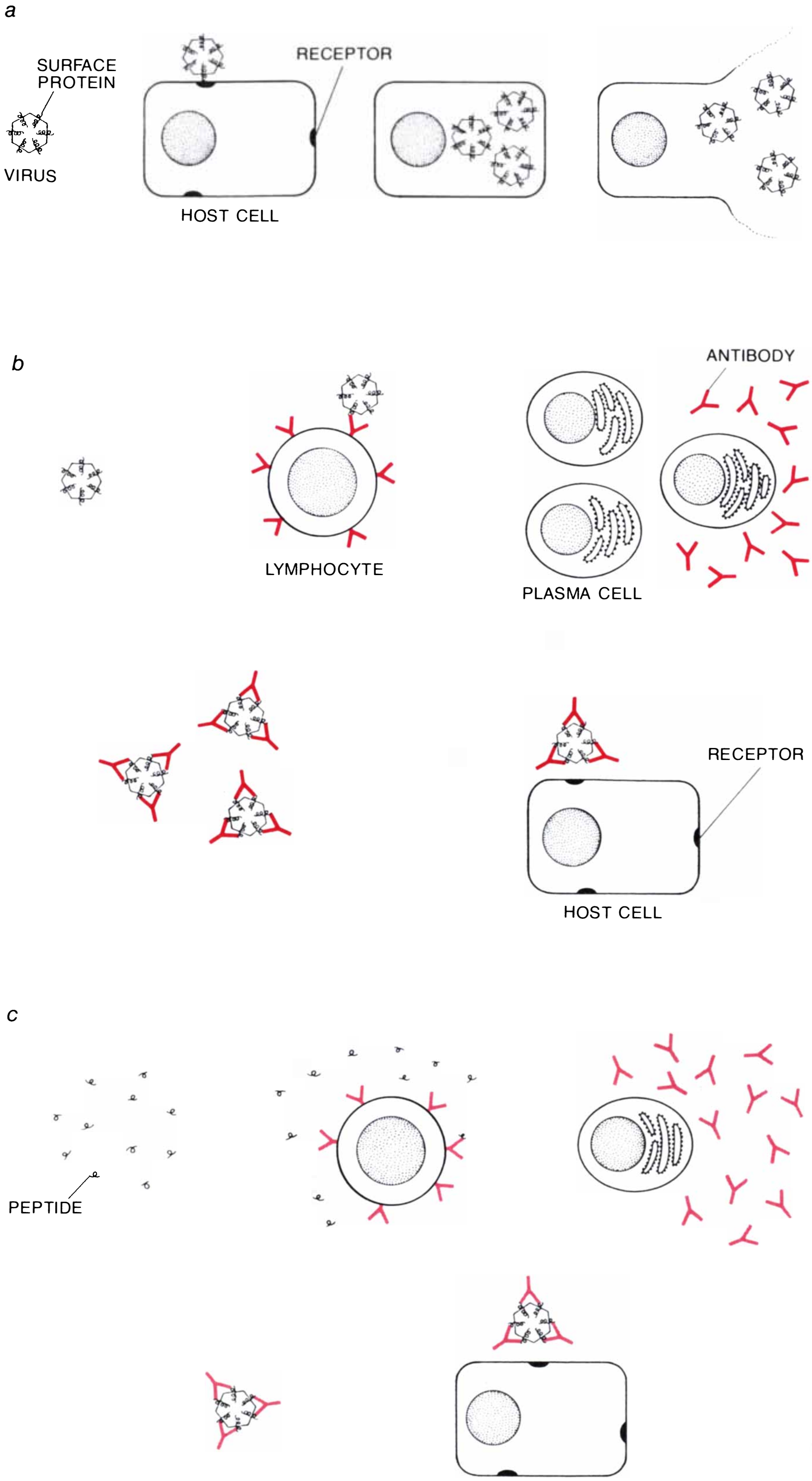
The site that is recognized by a specific lymphocyte and to which the antigen-binding site of a specific antibody subsequently binds is called an antigenic determinant. It is usually a limited patch of a viral surface protein. That being the case, might it not be possible to elicit antibodies capable of neutralizing a virus by injecting as a vaccine peptides that constitute antigenic determinants? The notion that a small part of an infectious agent might serve as a vaccine goes back more than 45 years, to studies done

SYNTHETIC VACCINES are designed with the help of computer-graphics programs. These displays generated by Arthur J. Olson of the Research Institute of Scripps Clinic show a method whereby parts of a viral protein that are on the surface of a virus, and therefore accessible to antibodies, can be identified. The backbone of the surface domain of the protein on the outer shell of the tomato bushy-stunt virus is displayed (*1*) on the basis of coordinates determined by Stephen C. Harrison of Harvard University and his colleagues. A single peptide of the protein is picked out in yellow, with the side chains of its component amino acids indicated in atomic detail (*2*). The peptide is enlarged and a sphere representing a water molecule is displayed (*3*). The sphere is rolled around the peptide to generate a map of the surface accessible to water (*4*); it does so, following an algorithm developed by Michael L. Connolly, by placing a dot at each point of its closest contact with the peptide, taking account of the sphere's own van der Waals radius (zone of influence, in effect) and that of each atom of the peptide and the rest of the protein. A similar dot-surface map is generated to show what parts of the peptide are still accessible to water when three copies of the protein are associated in an array on the surface of the virus (*5*) and when four such arrays (out of 60) are in position on the outer surface of the virus (*6*).

© 1983 SCIENTIFIC AMERICAN, INC



© 1983 SCIENTIFIC AMERICAN, INC



VIRAL INFECTION can be prevented by vaccination with the intact virus or a small peptide synthesized to mimic a site on a surface protein of the virus. A virus infects a cell (*a*) by attaching itself to a receptor on the cell, entering the cell and taking over the cell's machinery to make many copies of itself, which burst the cell. An attenuated or killed virus can be injected as a vaccine (*b*). When such a virus encounters a receptor on a lymphocyte that is genetically programmed to recognize and make antibody against a surface protein of the virus, the lymphocyte is stimulated to form a clone of antibody-secreting plasma cells. In any subsequent exposure to the virus the antibodies bind to the protein and neutralize the virus by blocking its attachment to host cells. A synthetic peptide mimicking part of the surface protein can serve as a vaccine (*c*). It too elicits antibodies able to bind to the surface protein and neutralize the virus.

by Walther F. Goebel of the Rockefeller Institute for Medical Research. He showed that immunizing mice with a sugar chain from the surface of the bacterium that causes pneumonia would protect the mice from subsequent infection. The feasibility of making immunogenic peptides to order was greatly increased in the 1960's when R. B. Merrifield of Rockefeller University developed a method for automatically linking amino acids, the subunits of proteins, to make a peptide.

It is not always easy to determine the amino acid sequence of a particular peptide in a large protein molecule, but a way around that difficulty was found. The amino acid sequence of a protein is specified by the sequence of the subunits called nucleotides in the nucleic acid (DNA or RNA) encoding the protein. A few years ago remarkable new techniques were developed for determining the nucleotide sequence of DNA. Now it became possible to very rapidly determine the complete sequence of the DNA of a virus (or of a DNA copy of a viral RNA). In principle, then, one could identify a nucleic acid sequence likely to encode a surface protein and translate that region according to the genetic code to derive the amino acid sequence of the protein. Then, exploiting Merrifield's procedure, one could synthesize short peptides in various regions of the protein, inject them in laboratory animals and see if they would give rise to antibodies that bind to the natural protein and perhaps neutralize the virus.

Before one could begin rationally designing vaccines, however, some difficult questions had to be answered having to do with the nature of an antigenic determinant and whether or not a short synthetic peptide could be expected to mimic one. Antibodies bind to protein sites by reason of shape and electric charge. The backbone of a protein is a chain of carbon and nitrogen atoms bristling with projections: the side chains of successive amino acids, each of which has a characteristic shape and array of charges. The sequence of the amino acids (which constitutes the primary structure of the protein) and their consequent short-range interactions determine how the chain twists locally into what is called the secondary structure: perhaps a helix or a sharp turn. Interactions among secondary structures in turn determine how the chain folds back on itself to form the complex tertiary structure characteristic of proteins.

It follows that there can be two kinds of antigenic determinant. If the antigen-binding site of an antibody molecule recognizes and binds to an array of contiguous amino acid side chains, one speaks of a continuous antigenic determinant. A discontinuous determinant, on the other hand, is one made up of groups of amino acids that are separat-

© 1983 SCIENTIFIC AMERICAN, INC

ed from one another along the chain but are brought into proximity by tertiary folding.

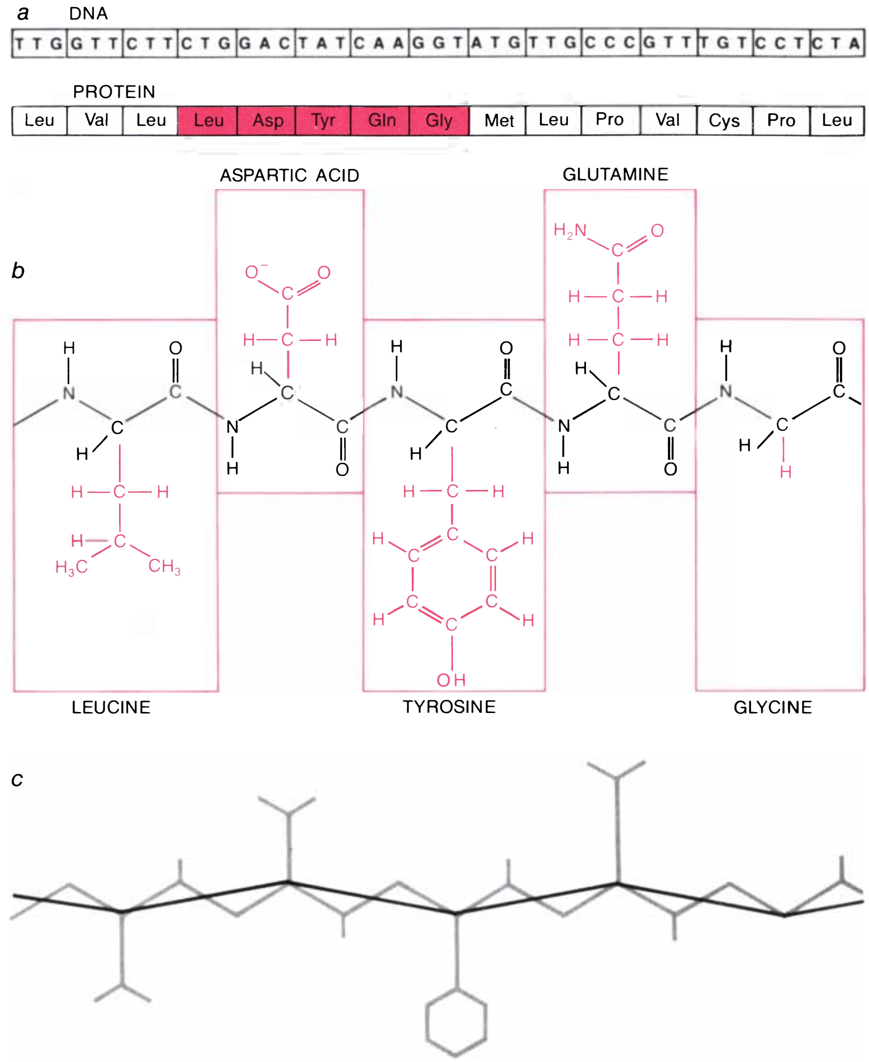
As of two or three years ago a considerable body of immunological doctrine suggested that synthesized peptides were not likely to mimic the natural antigenic determinants of intact viral proteins. In the first place it appears that the immune response to a viral protein in the course of a natural infection is directed to a small number of antigenic determinants confined to a few parts of the protein molecule; specifically, there is evidence that a protein's immunogenicity (its ability to elicit antibodies) is accounted for by only one site for every 50 amino acids. Moreover, most of these relatively few antigenic determinants seem to be discontinuous ones. To make a peptide simulating such a determinant one would have to make a long stretch of the protein; such a synthesis is technically feasible but inefficient. If a long peptide were synthesized, it seemed unlikely to fold itself into the correct tertiary structure in the absence of the rest of the molecule. Indeed, it was thought the individual side chains in a short peptide would not spontaneously orient themselves to mimic even the simpler secondary structure of a continuous determinant.

To challenge these rather pessimistic assumptions we decided two years ago to synthesize in short segments almost the entire length of a viral protein and see whether antibodies to many of the peptides might not react with the entire protein. The molecule we chose was the hemagglutinin of one strain of the influenza virus. It is a glycoprotein (a protein with sugar chains attached) that forms spikes radiating from the spherical envelope of the virus. The hemagglutinin spikes have important functions. They mediate the binding of the virus to a host cell. As their name indicates, they cause red blood cells to clump together. And they are primarily responsible for eliciting antibodies that neutralize the virus.

Two recent accomplishments made the influenza hemagglutinin an attractive choice for our studies. Walter Fiers of the State University of Ghent in Belgium and his colleagues had determined the nucleotide sequence of the gene encoding the protein, and from it they had predicted the amino acid sequence. Ian A. Wilson and Don C. Wiley of Harvard University and John J. Skehel of the National Institute for Medical Research in London had determined the structure of the protein by X-ray crystallography, so that we would be able to relate each peptide we synthesized to a particular region of the secondary and tertiary structure. A molecule of the hemagglutinin is composed of two protein chains designated *HA*1 and *HA*2; each spike is a trimer consisting of three such molecules. *HA*1 seems to be the primary target of the immune system, and so we turned first to that chain.

We synthesized 20 peptides accounting for some 75 percent of the length of the chain and sampling a wide variety of secondary and tertiary structures. We injected each of the peptides, bound to a large carrier protein, into rabbits and tested serum from each rabbit (by what is called an enzyme-linked immunosorbent assay) for the presence of antibodies that would bind to the injected peptide, to the hemagglutinin molecule and to the intact virus. All 20 of the peptides gave rise to antibodies against themselves; 15 of the antipeptide antibodies reacted with the protein molecule and 17 of them bound to the influenza virus. In other words, a peptide from almost any region of a viral protein can elicit an antibody that will recognize the entire protein, either in purified form or on the surface of the virus; the natural immunogenicity of a protein (the only kind that has been much studied up to now) is less than the potential immunogenicity of its parts. The only requirement seems to be that the synthesized peptide be on the surface of the folded protein, where the antibody can get at it.

There are a number of ways to estimate whether a given peptide is on the surface of a protein—or, to put it more exactly, whether a peptide (or part of it) is accessible to antibody molecules in solution in an animal's blood. One of the most precise ways is to program a computer to make a graphic display of the



NUCLEOTIDE SEQUENCE of a stretch of viral DNA encodes the amino acid sequence of a viral protein (*a*). Once the DNA sequence is known, investigators can predict the amino acid sequence and pick out for synthesis those peptides, or short segments of the protein chain, that seem likely to give rise to antibodies. The chemical structure of part of the peptide shown in *a* is diagrammed (*b*). The main chain is composed of carbon (C) and nitrogen (N) atoms with projecting hydrogen (H) and oxygen (O) atoms. Each amino acid is characterized by a side chain (*color*) that projects from what is called the alpha carbon. In the computer displays of proteins that illustrate this article either the alpha-carbon backbone is shown (*c*) in simplified form (*black*) or the chain is shown in atomic detail but with hydrogen atoms excluded (*gray*).

© 1983 SCIENTIFIC AMERICAN, INC

solvent-accessible regions [*see illustration on page 67*]. First the peptide under study is displayed, complete with its branching side chains, in the conformation it assumes on the intact protein molecule. Then the computer rolls a sphere representing a water molecule around the peptide. The sphere penetrates where it can, given the van der Waals radius (loosely, the zone of influence in terms of electric charge) of the water molecule and of the atoms of the side chains. The sphere leaves a dot at each point of contact with the peptide, thus in effect drawing a map of the solvent-accessible regions of the peptide.

The scheme I have described is effective, but it does require that the structure of a virus and its proteins be known in atomic detail, and such data are not available for many viruses. Fortunately there is a way to predict which parts of a protein are likely to be on the surface if one knows only the amino acid sequence. The 20 amino acids can be characterized for the most part as being either hydrophilic (water-attracting) or hydrophobic (water-repelling). The hydrophobic amino acids are most often buried in the core of the folded protein, whereas the hydrophilic ones are usually on the outside of the molecule. The four most hydrophilic amino acids are lysine, arginine, aspartic acid and glutamic acid. One simply assumes the regions of a protein that are relatively rich in these amino acids are likely to be exposed to an aqueous environment and therefore are on the surface.

Our results with the influenza virus suggested, as I pointed out above, that just about any surface peptide should give rise to antibodies able to react with the entire protein. How is it that a lone peptide can be depended on to assume a conformation similar to its conformation when it is folded into the complete protein molecule? It may be the very lack of constraints on shape that gives rise to the correct shape. An injected small peptide, in solution in an animal's bloodstream, keeps twisting and bending. As it changes shape it presumably continues to encounter lymphocytes, and hence to trigger the secretion of antibodies against each shape in turn. One of those shapes, it now seems clear, is likely to coincide closely with the conformation the peptide adopts as part of a large protein. (It is interesting to note that in many cases small peptides are more effective in eliciting antibodies against a viral protein than the purified protein is. Perhaps the isolated protein locks itself into a folded configuration that is different from the shape it assumes when it is arrayed with many copies of itself on a virus.)

If virtually any surface region of a viral protein can elicit antibodies against that protein, the way is open for designing a vaccine rationally instead of injecting whole virus particles and in effect taking whatever the immune system offers. That is being done in a number of laboratories. The first vaccine we developed was against the virus of foot-and-mouth disease, a historic plague of animals (cattle in particular) that is still a serious problem on every continent except Australia and North America. Some three billion doses of killed-virus vaccine are administered annually in an effort to control it. The virus shell is composed of 240 protein molecules, 60 copies each of the four viral proteins *VP*1, *VP*2, *VP*3 and *VP*4. There have been attempts to make a vaccine based on *VP*1, which is thought to be the major natural target of neutralizing antibody, but the entire protein has proved to be not very immunogenic. In collaboration with Fred Brown and David J. Rowlands of the Animal Virus Research Institute in England, we undertook to synthesize a more immunogenic subregion of the protein.

The sequence of *VP*1's 213 amino acids had already been determined from the nucleotide sequence. The structure of the protein is not yet known in atomic detail. Analysis of the amino acid sequence, however, was enough to draw our attention to the region between amino acids 141 and 160. The region is fairly rich in the hydrophilic side chains of arginine, lysine and aspartic acid, suggesting it might be on the surface. The amino acid proline is often found where a protein makes a sharp bend (which may make for immunological specificity), and the region has two prolines.

Most important, it was known that in this region the amino acid sequence undergoes considerable variation from strain to strain. A virus needs to evade a host's immune-system defenses long enough to invade host cells and produce progeny. It does so by continually changing the antigenic determinants on its outer shell or envelope, with the result that antibodies formed earlier

ANTIGENIC DETERMINANTS are sites on a protein that are recognized by antibodies and to which the antibodies bind. The two kinds of determinants are illustrated by these computer displays, in which part of the main chain of the enzyme lysozyme is shown in yellow, with antigenic determinants (side chains included) picked out in orange. A continuous determinant (*left*) is one formed by an array of contiguous amino acids along one segment of the chain. A discontinuous determinant (*right*) is one composed of several segments that are separated from one another along the chain but are brought together to form a single site by the tertiary folding of the protein.

© 1983 SCIENTIFIC AMERICAN, INC

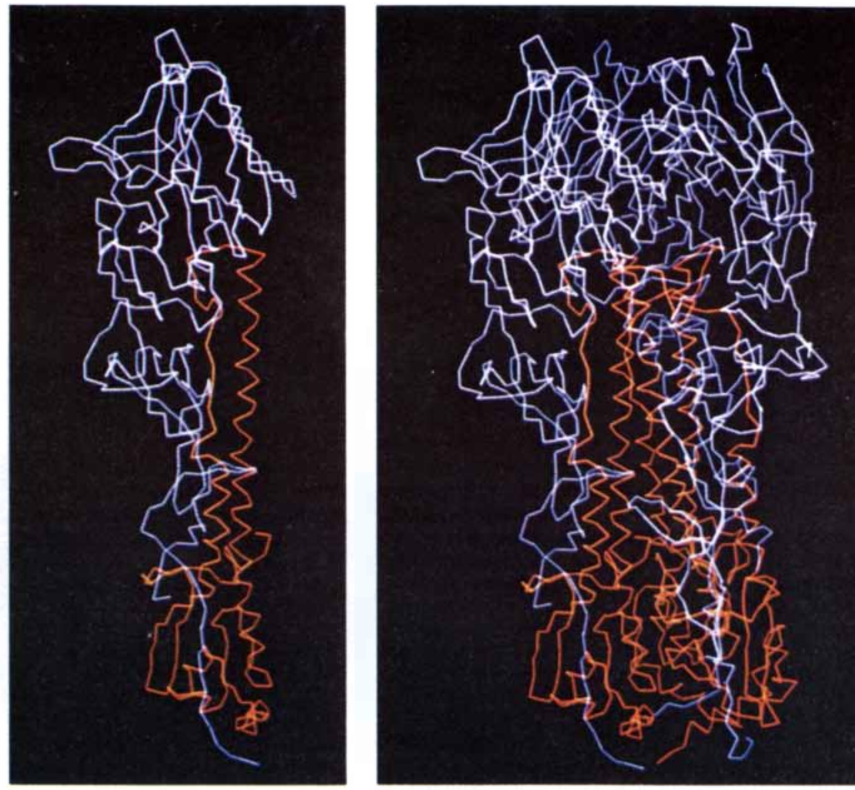
against a parental strain will not bind to the altered virus and neutralize it. In less teleological terms, what happens is that natural selection encourages the survival of a strain in which spontaneous mutation has brought about an immunologically significant variation. A region subject to such mutation is likely to be one that ordinarily elicits antibodies, and so a peptide from that region is a good candidate for a synthetic vaccine.

We synthesized a number of peptides from the *VP*1 of the foot-and-mouth virus, including the peptide spanning positions 141 to 160, coupled each one to a carrier molecule and injected it into rabbits. After determining that the peptides elicited antibodies against themselves, we did a bioassay to test their ability to neutralize the virus. When the rabbit antiserum to peptide 141–160 was injected into mice along with the virus, it effectively neutralized the virus; the mice were given what is called passive immunity by the injected antipeptide antibody. There are three major strains of the foot-and-mouth virus; we had worked with type *O*. When we tested the same antibody's effect on the other two strains, the neutralization was much weaker, supporting the impression that changes in this particular region of *VP*1 allow the virus to elude antibodies made against a different strain.

The real test of any vaccine is its ability to confer not only passive immunity but also active immunity, that is, to cause the immune system to synthesize its own antibodies for protection against a future attack. We immunized guinea pigs with one shot (200 micrograms) of the synthetic-peptide vaccine and after 35 days administered a powerful dose of the virus: 10,000 times the dose normally required to infect half of a group of animals. None of the immunized animals came down with the disease, whereas all the unimmunized control animals did. A similar active-immunity test in cattle is now under way in England. Already we have found that cattle immunized with the peptide secrete antibody at a concentration that should protect them from a virus challenge.

A practical vaccine will, of course, have to protect against all three strains of the virus. We have synthesized peptides from the same region (141 to 160) of the *VP*1 of type-*A* and type-*C* viruses, which have slightly different amino acid sequences. So far antibodies to these peptides have been bioassayed in mice, and they do neutralize viruses of type *A* and type *C* respectively. A synthetic three-strain vaccine against foot-and-mouth disease seems to be within reach.

Given that one can in principle complete the technological chain from nucleic acid sequence to synthetic-peptide vaccine, the next step is to explore certain aspects of vaccine design in more detail. Current research is aimed at finding additional and more sophisticated ways to pinpoint the regions of viral proteins that are most vulnerable to immunological attack and also ways to confer immunity of long duration.



HEMAGGLUTININ MOLECULE of the influenza virus is made up of two protein chains, *HA*1 (*purple*) and *HA*2 (*orange*). The three-dimensional structure of the molecule (*left*) is based on X-ray crystallographic data supplied by Ian A. Wilson and Don C. Wiley of Harvard University and John J. Skehel of the National Institute for Medical Research in London; only the backbone is displayed. Three *HA*1 chains and three *HA*2 chains are associated in the hemagglutinin trimer (*right*). Trimers form spikes on the spherical envelope of the virus particle.

There are essentially two kinds of region in viral proteins. In one kind of region variation is allowed; indeed, it is essential if the virus is to confound the host's immune system. A vaccine that gives rise to antibodies against such a region (*HA*1 of influenza or *VP*1 of foot-and-mouth disease, for example) can be expected to be immunogenic, but it must be tailored to cope with each variant strain. There are other regions in viral proteins that have a constant and essential biological function and whose structure cannot vary. The amino acid sequence of such a region is highly conserved among all strains of the virus. If an essential and conserved region is accessible to antibodies, a vaccine that elicits antibodies against it should be a universal vaccine, one from which the virus cannot escape.

The influenza virus is notorious for its ability to evade the immune system by changing parts of the hemagglutinin on its surface. The changes are in the *HA*1 chain, and it was peptides from *HA*1 that we first tested for immunogenicity in the experiment I described above. The hemagglutinin has at least two other and unvarying functions, however. It forms a socket that binds to sialic acid, a sugar on the surface of cells, thereby initiating the infective process. It also causes the surface membranes of infected cells to fuse together, increasing viral infectivity. Evidence from a number of studies by Purnell W. Choppin of Rockefeller University and his colleagues and Michael Waterfield of the Imperial Cancer Research Fund suggests that a conserved region at the beginning of the *HA*2 chain is at least partly responsible for the fusion.

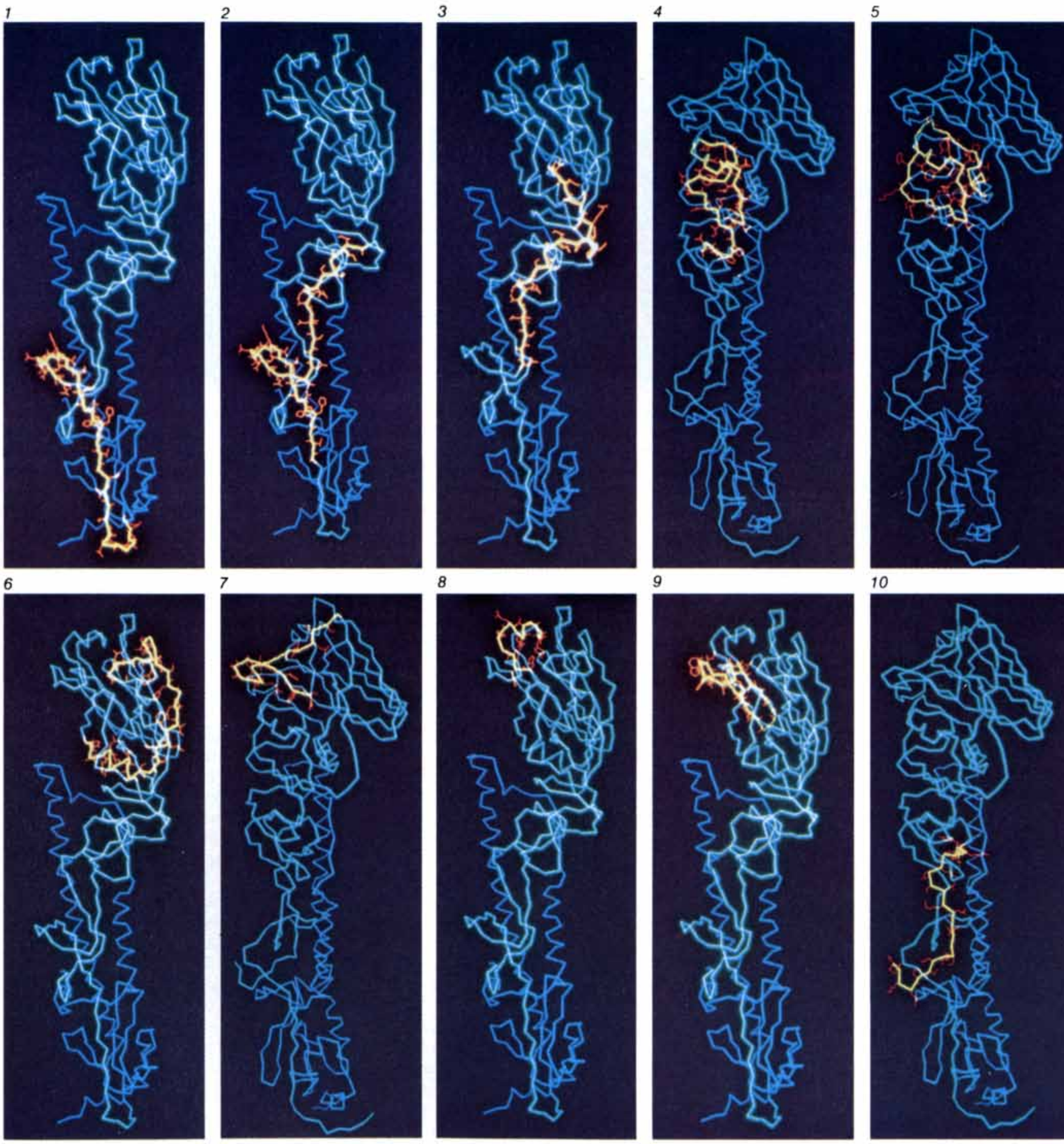
Normally the immune system does not "see" this region; antibodies to it are usually not made in the course of natural infection. If such antibodies were made (if the region were not only highly conserved and biologically essential but also naturally immunogenic), every human being and other animal host would long since have become immune to the influenza virus; there would be no influenza. We decided to see whether a synthetic peptide from the region might elicit antibodies that would neutralize more than one strain of the virus.

Stephen Alexander, Hannah Alexander and Nicola Green synthesized several peptides from the beginning of the

© 1983 SCIENTIFIC AMERICAN, INC

*HA*2 molecule; the amino acid sequence they chose was that of influenza type *A*, subtype H_3. Rabbit antiserums against these peptides effectively neutralized the *A*, H_3 virus when they were tested in a tissue-culture system, showing that even if this region of the *HA*2 molecule is not normally immunogenic, it is accessible to specific antipeptide antibodies. Moreover, the same antiserums neutralized viruses of a different subtype and even a different major type (influenza *B*). In contrast, antibodies elicited against the intact H_3 subtype of the virus did not neutralize other subtypes. Nor, of course, are individuals immunized by the injection of one influenza subtype thereby immunized against infection by a different subtype. Much more study of the anti-*HA*2 antibodies is needed, but these first results suggest that much may be expected of vaccines designed to attack a conserved region having an important biological function.

One way to design a vaccine, then, may be to choose for synthesis a region of a protein that can be expected, on theoretical grounds, to serve as an antigenic determinant. There is another way to proceed. One can try to mimic antigenic determinants that have already been shown to be important by seroepidemiologic studies, in which

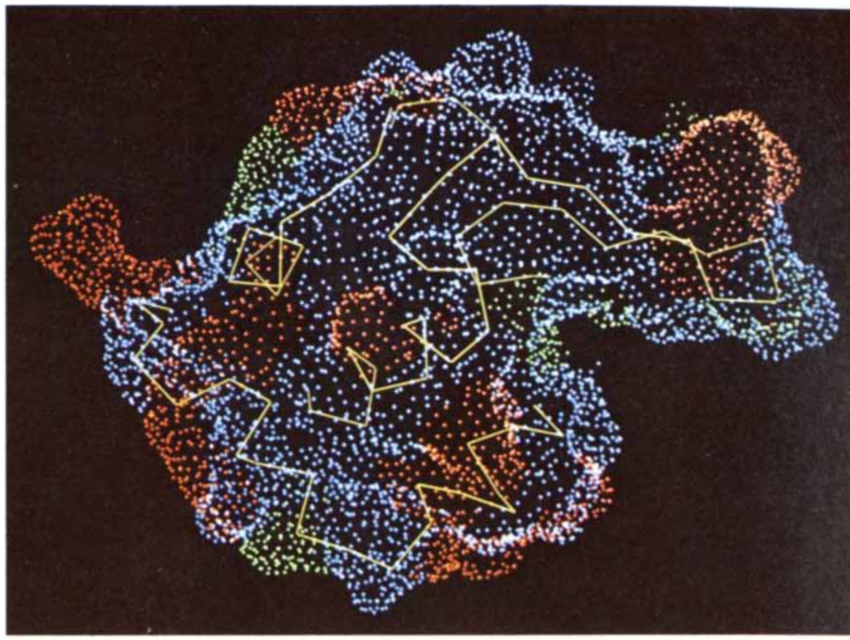


SERIES OF PEPTIDES accounting for some 75 percent of the *HA*1 chain was synthesized and each peptide was tested for its ability to elicit antibodies reacting with the peptide, with the hemagglutinin molecule as a whole and with the influenza virus particle. Ten of the 20 peptides tested are shown in yellow in these displays, with their side chains in orange, the remainder of the *HA*1 chain in light blue and the *HA*2 chain in dark blue; in each case the hemagglutinin monomer has been rotated to show the peptide in question to best advantage. Most of the peptides elicited rabbit antibodies that reacted with the virus, supporting the concept of synthetic-peptide vaccines.

© 1983 SCIENTIFIC AMERICAN, INC

closely related viruses are classified on the basis of the antibodies they elicit. Such studies reveal the existence of different types and subtypes of a virus, such as those I have alluded to in discussing foot-and-mouth disease and influenza.

The types are defined by long-term studies in which viruses responsible for a disease are collected from various areas and at different times. Antibodies taken from patients are tested against the panel of viruses; different antibodies react with different strains of the virus because they "see" different antigenic determinants. By testing many antibodies against a panel of viruses serologists can describe a given virus type in terms of its pattern of determinants and can learn which determinants have been operative in a particular individual or epidemic. Until recently, however, the determinants were abstract "markers" designated by letters or numbers; they were not understood in molecular terms. It was therefore not at all clear that the vast amount of accumulated seroepidemiologic data could help in the design of vaccines. Synthetic peptides might not mimic the structures responsible for serologic markers, for example, if those structures were discontinuous in nature. We decided to try to design a vaccine against the liver disease hepatitis *B* by synthesizing peptides that would be the molecular equivalent of abstract serologic determinants that have been shown to be important.



SOLVENT-ACCESSIBLE SURFACE of the lysozyme molecule is mapped by the method demonstrated in the illustration on page 67, but with the dots color-coded according to the amino acid responsible for each part of the surface: orange for the basic amino acids lysine and arginine, green for the acidic ones aspartic acid and glutamic acid and blue for all others. The basic and acidic amino acids are the most hydrophilic (water-attracting) ones. As the display demonstrates, they tend to be exposed on the surface. If a protein's structure is not known in detail, one can simply assume peptides rich in hydrophilic amino acids will be immunogenic.

The viral component that appears to be the major target of neutralizing antibody, the hepatitis-*B* surface antigen, is a protein on the envelope of the virus particle. Serologic studies show that the several strains of the hepatitis-*B* virus have in common one determinant, which is designated *a*. Each strain also has two other determinants: either *d* or *y* and either *w* or *r*. That means there are four possible types of the virus: *adw, ayw, adr* and *ayr*. Working with John L. Gerin of the Georgetown University School of Medicine and Robert H. Purcell of the National Institute of Allergy and Infectious Diseases, we wanted to synthesize peptides that would specifically elicit antibodies against the *d* and the *y* determinants. The hepatitis-*B* surface antigen, however, is a chain of 226 amino acids. We needed a guide to where to begin. Analysis of the antigen's amino acid sequence in three strains of the virus by several investigators had shown that there are substantial differences among the strains in only one region of the protein, between amino acids 110 and 140. Any difference in determinants must originate with differences in amino acid sequences, and so that region was clearly the place to seek our peptides.

Following one of the three available amino acid sequences, we synthesized several peptides spanning the region and made antibody to each of them. One 13–amino acid peptide (positions 125 to 137) elicited antibody that reacted only with viruses bearing the *y* determinant. This meant that the *y* determinant and its alternative, the *d* determinant, must lie within the 13–amino acid peptide. Having synthesized *y*, we should be able to synthesize *d* by following a different one of the other known sequences for that region. The sequence we chose next differs from the first one at just three positions: two threonines and a tyrosine are replaced respectively by proline, asparagine and phenylalanine. Antibody to the new peptide reacted only with viruses bearing the *d* determinant.

Perhaps the switch from *y* to *d* is brought about by fewer than three amino acid substitutions. By synthesizing peptides in which only one of the three positions is changed at a time we should soon learn whether one, two or all three of the substitutions are required. So far the experiment has taught us two things. One is that very subtle changes in the amino acid composition of a protein can define different strains of a virus. The other is that synthetic peptides can mimic the distinctions revealed by serologic studies; in designing synthetic vaccines one will be able to take advantage of serologic evidence and thereby, for example, attempt to cope with a current epidemic occasioned by a mutant strain.

Before synthetic vaccines can become an alternative to conventional ones ways must be found to increase the magnitude and duration of the immunity they confer. An advantage of attenuated viruses, as I have mentioned, is that they keep proliferating and thus present ever more antigen to the immune system, thereby increasing the antibody level. A synthetic peptide, like a killed virus, does not replicate, and so substances called adjuvants must be added to the vaccine to increase the level and duration of immunity.

One of the most powerful of these, Freund's adjuvant, is an emulsion of mineral oil and water mixed with killed bacteria (mycobacteria, a class that includes the agent of tuberculosis). Just how it works is not known, but it probably does two things. The emulsion retains antigen trapped in it, forming a depot from which antigen is released slowly, simulating the sustained presentation of antigen by a replicating virus. Meanwhile the killed bacteria attract to the depot various cells that have roles in the immune response. To attract such cells is also to serve as an irritant, however, causing soreness at the site of injection and even the formation of an abscess. Powerful adjuvants such as Freund's may not be administered to human beings; instead, alum, or aluminum hydroxide, is commonly included with a killed-virus vaccine. The antigen is adsorbed onto the aluminum hydroxide particles and is released slowly after injection. In some cases, it appears, alum

© 1983 SCIENTIFIC AMERICAN, INC

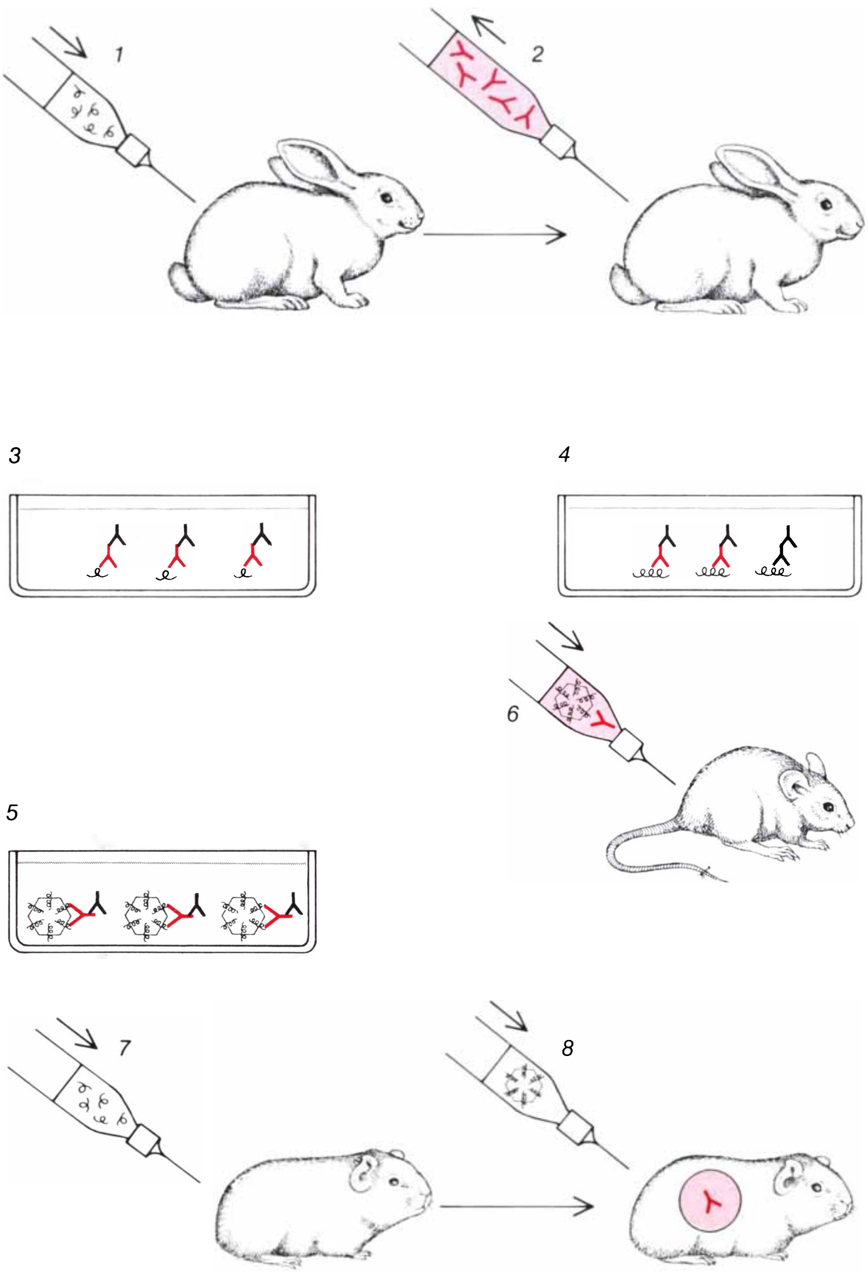
will serve well enough with synthetic vaccines, but most such vaccines will require something better. Much effort is currently aimed at the development of safe and effective adjuvants.

Some hopeful results in this direction have been reported by François Audibert and Louis Chedid of the Pasteur Institute and Ruth Arnon and Michael Sela of the Weizmann Institute of Science, who recently elicited antibodies conferring passive immunity against the diphtheria toxin. They synthesized three peptides corresponding to regions of the toxin molecule. Before injection a synthetic peptide must be coupled to a carrier molecule. (In all our experiments the carrier was keyhole-limpet hemocyanin, a respiratory pigment of a mollusk, which is often used for this purpose.) The Pasteur-Weizmann group devised a carrier-adjuvant combination. They coupled their diphtheria-toxin peptides to a carrier that had been linked to a simple derivative of the cell membrane of mycobacteria; the membrane derivative apparently serves as an adjuvant, significantly increasing the immunogenicity of the peptides.

As I mentioned at the beginning of this article, antipeptide antibodies should have important applications as investigative tools. One application arises from advances in nucleic acid technology. Because of the precision of recombinant-DNA techniques and the remarkable speed of the new methods for determining the sequence of a nucleic acid, it is often the case that one has the complete sequence of a gene before the protein it encodes is known; one has a gene in search of a protein. From the DNA or RNA sequence one can now predict an amino acid sequence and select a small region of it for synthesis. Antibodies elicited by the peptide serve as a probe for finding the protein. By this means, for example, the proteins encoded by some oncogenes, the tumor-inducing genes of certain viruses, have been identified and tracked to particular sites in cells.

The distinguishing feature of antipeptide antibodies is of course their predetermined specificity. They are antibodies not to a protein but to a particular part of the protein that is known to the experimenter. Once they have been manufactured by the immune system of a laboratory animal they can easily be purified by a single passage through an immunoadsorbent column. Then they can be used to explore in detail the structure of a protein and to correlate the structure with function. For example, antibodies to a particular part of a protein will bind to that part and inhibit its function, and so tests with a battery of peptides should make it possible to define the active site of a protein or various sites that are active in different functions. Again, many peptide hormones originate in a long protein chain that is subsequently cleaved by an enzyme into smaller pieces, each one a hormone with a particular activity. Antibodies to peptides serve to sort out the cleaved pieces and correlate each with its activity.

Finally, antibodies of predetermined specificity can reveal just which part of a gene is being expressed to make a protein at a particular time. Mammalian genes are split into a number of the coding regions called exons. In the case of certain variable proteins different exons are expressed at different times. Successive forms of antibody molecules are synthesized, for example, in the course of the immune response. Now one can synthesize the peptide encoded by each exon, make antibody to it and learn which exon is being expressed at a particular stage of the immune response.



SYNTHETIC PEPTIDE is injected into a rabbit (*1*) to test its ability to elicit antibodies. Serum withdrawn from the rabbit (*2*) can first be assayed for the presence of antibodies able to bind to the injected peptide (*3*), to the viral protein as a whole (*4*) and to the virus itself (*5*). The binding of the antipeptide antibody (*color*) is detected by noting the activity of a labeling enzyme coupled to a second antibody (*black*) that binds to the first one. The ability of the antipeptide antibody to neutralize the virus (to confer passive immunity) is tested by injecting mice with the virus and the antiserum (*6*); if the mouse stays well, the virus has been neutralized. The ability of the synthetic peptide to serve as a vaccine conferring active immunity can be tested by injecting the peptide into a guinea pig (*7*) and later injecting the virus (*8*). If animal survives (whereas unimmunized control animals do not), efficacy of vaccine is demonstrated.

© 1983 SCIENTIFIC AMERICAN, INC

1700 pounds of optical perfection.

Perfection is not a term we use lightly. Or frequently. We've been taught for generations that perfection, in lens-optics terms, means 1/4 wavelength. We've bettered this tolerance in the manufacture of an "egg crate" mirror, one-fifth the weight of a solid mirror. Manufactured for the Space Telescope Program, every point on its nearly 8-foot-diameter aspheric surface is within 1/10 wavelength of perfection. So while we're not 100% perfect, we're within several millionths of an inch of that goal.

We put all our experience into play in the manufacture of this massive optical element 97½ inches in diameter and 1700 pounds in weight. And our challenges were many. Since this lens is intended to perform in orbit, where the pull of gravity has been neutralized, we had to solve a gravity-distortion problem during the manufacturing cycle. We solved the problem with an air cushion, a membrane that inflates for testing to provide a uniform support and thus eliminate gravity-induced strain. The support deflates for polishing operations.

An aspheric mirror to gather information from the universe is an example of our large-optics capability. (Optics of this kind are ground, polished, and tested in our own lab with its vibration-isolated, six-story-high interferometric test towers.) But we're big in small things, too. Like lenses for cameras and copiers. We meet customers' needs with pride, precision, and stringent quality control.

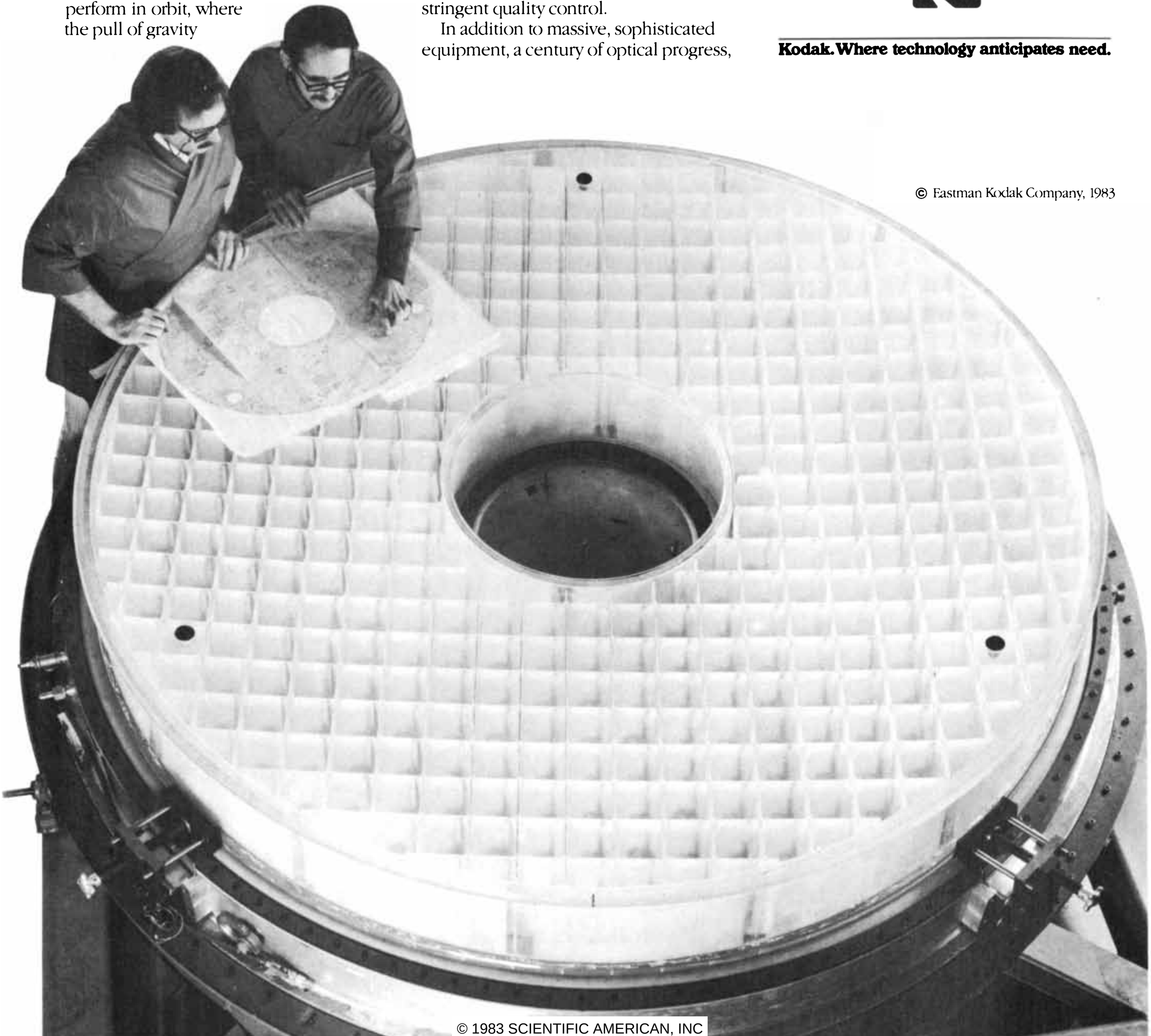
In addition to massive, sophisticated equipment, a century of optical progress, and systems utilizing our ability to the fullest, we offer customers a tradition of realism when predicting costs and delivery.

That's important. Whether you seek to get a clearer picture of your neighbor or of celestial objects.

For more information about Kodak's contribution to the Space Telescope Program, write for the booklet *Mirror for the World's Most Powerful Telescope.* Send your request to: Richard M. Price, Research and Engineering, Kodak Apparatus Division, 901 Elmgrove Road, Rochester, NY 14650.

Kodak. Where technology anticipates need.

© Eastman Kodak Company, 1983



© 1983 SCIENTIFIC AMERICAN, INC

SCIENCE AND THE CITIZEN

Early Warning

The apparent demise of the Reagan Administration's plan for deploying 100 MX missiles in a closely spaced array of "superhardened" missile silos has left strategic military planners in a quandary. The "dense pack" scheme was the latest of more than 30 basing modes for the new missile that have been considered seriously at one time or another by successive administrations; on close scrutiny all of them have been found wanting for a variety of technical, economic and political reasons. At this stage it is beginning to seem to many people, both in and out of the Government, that there simply is no basing mode for the MX that would guarantee its ability to ride out a nuclear attack by increasingly accurate Russian land-based missiles.

Such a conclusion is likely to focus renewed attention on yet another solution to the perceived problem of the vulnerability of the U.S. land-based missile force: the idea is to install the MX missiles in existing Minuteman missile silos and to adopt a "launch on warning" strategy. U.S. surveillance satellites are thought to be capable of providing a 30-minute warning of an attack by Russian land-based missiles aimed at the Minuteman sites. In that interval the U.S. missiles (either Minuteman or MX) could be launched in the opposite direction, leaving only empty holes as targets for the incoming Russian missiles. One attraction of such an approach is economic: it would not require the construction of a costly new basing system for the new generation of missiles. The chief obstacle is political: many Americans would presumably object to the automatic nature of such a response.

The present situation was not entirely unforeseen. Critics of the U.S. role in the nuclear-arms race have argued for some time that the land-based missiles of both the U.S. and the U.S.S.R. would eventually become vulnerable to attack and that greater efforts should therefore be made to phase them out of existence in favor of less vulnerable, and hence less destabilizing, submarine- and aircraft-based strategic weapons. Writing in these pages three administrations ago, Herbert F. York, former director of research and engineering for the Department of Defense, examined the various alternatives that had been advanced by then for protecting land-based missiles. The proposals included the deployment of antimissile defenses, the superhardening of silos, the proliferation of land-based missiles, a "shell game" in which there are more silos than missiles and the use of mobile missiles. York found that none of the plans "holds much promise beyond the next 10 years" (see "Military Technology and National Security," by Herbert F. York; SCIENTIFIC AMERICAN, August, 1969).

According to York, "one and only one technically viable solution seems to have emerged for the long run: Launch on warning." To most people, including him, he added, "such an approach to the problem is politically and morally unacceptable, and if it really is the only approach, then clearly we have been considering the wrong problem" in attempting to preserve the land-based component of the U.S. strategic deterrent force. For the time being, he concluded, submarine- and aircraft-based strategic weapons would appear to provide a satisfactory alternative to land-based missiles. "I expect, however, that as the continuing national debate subjects the whole matter of strategic arms to further public scrutiny we shall learn that these other alternatives also have dangerous flaws, and we shall see confirmed the idea that there is no technical solution to the dilemma of the steady decrease in our national security that has...accompanied the steady increase in our military power."

The Inflationary Universe

According to the standard big-bang cosmology, the universe began as a point of infinite density and expanded to a volume about the size of a softball by the end of the first 10^{-35} second. At this stage the universe was already far too big and had expanded too quickly to have yet been crossed by a signal moving at the speed of light. Causal connections between events, which are limited to the speed of light, were confined to regions about 10^{-25} centimeter across; some 10^{78} such regions would fit in a universe the size of a softball.

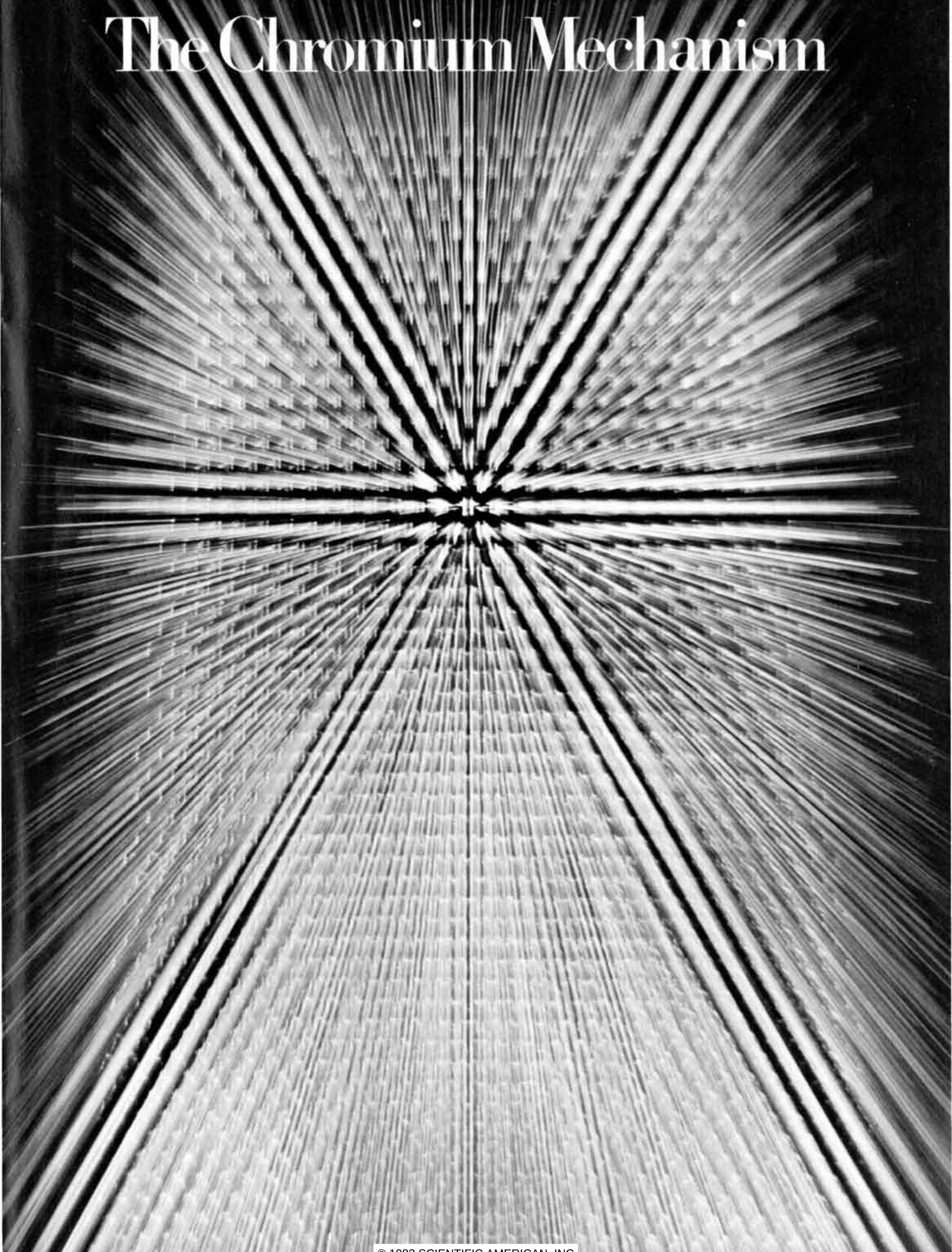
Physical theory now makes it possible to solve several outstanding problems in cosmology by supposing that the universe observable now was once an extraordinarily minute part of such a causally connected region. For a period beginning 10^{-35} second after the big bang and continuing for 10^{-32} second the region may have passed through what is called the inflationary phase. The region blew up by perhaps 100 orders of magnitude, from a trillionth the diameter of a proton to a diameter 10^{47} times that of the current observable universe. (The precise inflation factor is not important as long as it is greater than 10^{25}.) At the end of this titanic expansion the region that was to become the current observable universe must have been the size of a softball, embedded in an unimaginably greater universe that has ever since continued to expand beyond our horizon of cause and effect. At age 10^{-35} plus 10^{-32} second the inflationary phase was replaced by the standard phase, and the softball-size region, as well as the surrounding universe, resumed the rate of expansion that has since governed cosmic evolution.

These striking proposals have been made by a group of physicists and cosmologists who see in the physics of elementary particles a path to a consistent and perhaps even a predictive understanding of what the universe was like only 10^{-35} second after it began. The idea originated in a paper published last year in *Physical Review* by Alan H. Guth of the Massachusetts Institute of Technology. Guth noted certain inconsistencies of the model in his paper, but within a year Andreas Albrecht and Paul Steinhardt of the University of Pennsylvania and, independently, A. D. Linde of the Lebedev Physical Institute in Moscow were able to patch up some of the difficulties. Now these investigators and several others have taken up revisions of the inflationary model and are pursuing its consequences.

Guth's central idea was to apply thermodynamic principles to the interactions of matter and energy in a universe whose temperature is approximately 10^{27} degrees Kelvin. At this enormous temperature a typical particle has an energy at which the distinctions between three of the fundamental forces of nature are thought to break down. At ordinary temperatures there are four such forces: the strong force, which is responsible for nuclear interactions; the weak force, which accounts for the decay of the neutron into a proton; the electromagnetic force, and the gravitational force. One of the most ardent hopes of theoretical physics has been to give a unified account of all four forces, but so far only the first three have been linked by the grand unified theories. Above the critical temperature of 10^{27} degrees K. the theories predict that the strong, the weak and the electromagnetic forces are symmetrical, in the sense that they are indistinguishable from one another. Below the critical temperature the strong force should act differently from the other two forces, and the symmetry is broken. One effect of symmetry breaking, Guth suggests, is that matter and energy enter a new phase.

The phase transition envisioned by Guth is modeled closely on the conversion of a liquid into a solid. The liquid-to-solid transition is accompanied by a rearrangement of the molecules of the liquid into a regular crystalline array. The rearrangement lowers the total en-

© 1983 SCIENTIFIC AMERICAN, INC



© 1983 SCIENTIFIC AMERICAN, INC

The Chromium Mechanism

The first comprehensive explanation of electrochemical activity during the plating of chromium has recently been formulated at the General Motors Research Laboratories. This understanding has aided in transforming chromium plating into a highly efficient, high-speed operation.

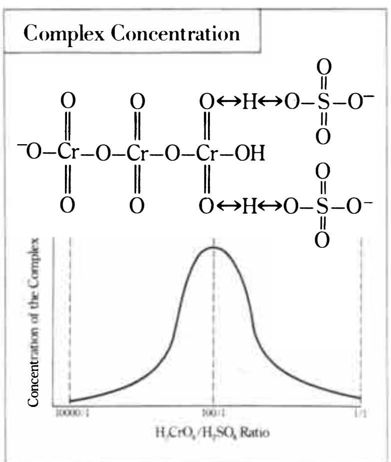


Figure 1: The electroactive complex and a theoretical plot of its concentration as a function of chromic acid to sulfuric acid ratio.

Figure 2: The electroactive complex diffuses from the bulk electrolyte solution (A) through the diffusion layer (B) to the Helmholtz double layer (C) to be discharged as metallic chromium (D) on the cathode (E) surface.

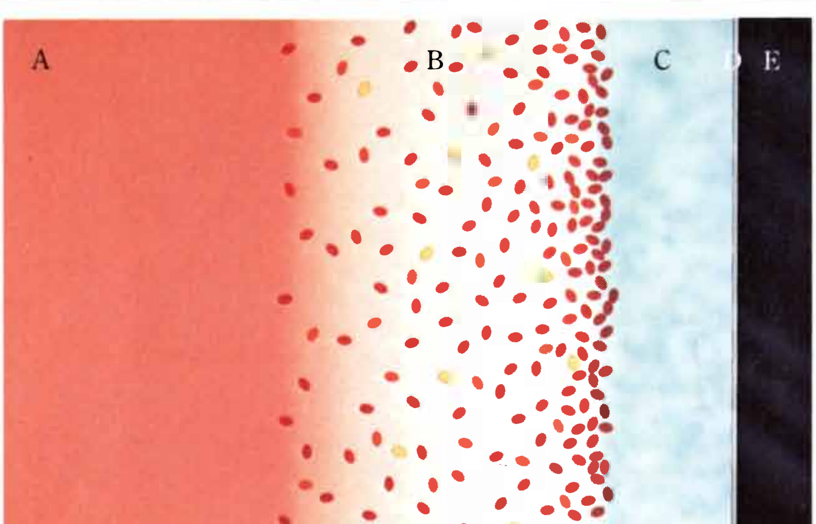
FOR MANY industrial applications, chromium coatings of more than 0.2 mil thickness are required for wear and corrosion resistance. But the conventional method of plating chromium is neither fast nor efficient. Nor, until the recent work of a GM researcher, had the steps involved in the century-old plating process been explained in detail. Through a combination of theory and experiment, Dr. James Hoare has devised the first comprehensive mechanism for chromium plating. This increased understanding has helped electrochemists at the General Motors Research Laboratories develop a system that plates chromium sixty times faster than the conventional method, while improving energy-efficiency by a factor of three.

The electrolyte for plating is a chromic acid solution which contains various chromate ions: chromate, dichromate and trichromate. From a series of steady-state polarization experiments, Dr. Hoare concluded that trichromate is the ion important in chromium deposition.

Sulfuric acid has been recognized as essential to chromium plating and has been assumed by some to be a catalyst for the process. In this strongly acidic solution, sulfate should be mostly present as the bisulfate ion $(HSO_4)^-$. Dr. Hoare found, contrary to expectations, that the addition of sulfuric acid to the plating bath decreased the conductivity of the solution.

Combining these findings with the results of previous investigations, Dr. Hoare concluded that the electroactive species was a trichromate-bisulfate complex (see Figure 1). From equilibrium considerations, he theorized that the maximum concentration of this species occurred at a 100-to-1 chromic acid/sulfuric acid ratio. The observation that the maximum rate of chromium deposition also occurred at this ratio supports the conclusion that this trichromate-bisulfate complex is the electroactive species.

During the plating process, the complex diffuses from the bulk solution toward the cathode (see Figure 2). Electron transport takes place by quantum mechanical tunneling through the potential energy barrier of the Helmholtz double layer and the unprotected chromium in the complex (Cr atom



© 1983 SCIENTIFIC AMERICAN, INC

on the left in Figure 1) loses electons by successive steps, going from Cr^{+6} to Cr^{+2}. Decomposition of the resulting chromous dichromate complex takes place by acid hydrolysis to form a chromous-oxybisulfate complex:

$$^{+}Cr{-}O{\leftrightarrow}H{\leftrightarrow}O{-}\overset{\displaystyle O}{\overset{\|}{\underset{\underset{\displaystyle O}{\|}}{S}}}{-}O^{-}$$

The positive end of this complex is adsorbed onto the cathode surface. Electrons are transferred from the cathode to the adsorbed chromium ion, forming metallic chromium and regenerating the $(HSO_4)^-$ ion. Thus, Dr. Hoare's mechanism explains how sulfuric acid, in the form of the bisulfate ion, participates in the plating process.

IT HAS long been known that chromium cannot be plated from a solution when initially present as Cr^{+3} because of the formation of the stable aquo complex, $[Cr(H_2O)_6]^{+3}$. Yet chromium can be plated when initially present as Cr^{+6} even though it must pass through the Cr^{+3} state before being deposited. Dr. Hoare's mechanism handles this paradox by explaining that the chromium ion being deposited (on the left in Figure 1) is protected by the rest of the complex as it passes through the Cr^{+3} state, so that the stable aquo complex cannot form.

The diffusion of the electroactive complex apparently controls the rate of the process, so that shortening the diffusion path increases the speed of chromium deposition. A high rate of relative motion between the electrolyte and the cathode will shorten the path. This can be accomplished by rapid flow or by agitation of the electrolyte.

Dr. Hoare found that the rate of chromium deposition increased with electrolyte flow until the process was no longer diffusion-controlled. He also found that the use of dilute electrolyte significantly increased plating efficiency.

"This project is an excellent example," says Dr. Hoare, "of how basic research and engineering principles can be combined to develop a new, successful process. Now, we'd like to take on the challenge of plating successfully from Cr^{+3}, which would be an even more efficient way to provide corrosion and wear resistance."

General Motors

MARK OF EXCELLENCE

THE MAN BEHIND THE WORK

Dr. James Hoare is a Research Fellow at the General Motors Research Laboratories. He is a member of the Electrochemistry Department.

Dr. Hoare served as an electronics technician in the U.S. Navy during the Second World War. In 1949, he received his Ph.D. in physical chemistry from the Catholic University of America. After an assistant professorship at Trinity College in Washington, D.C., he joined the US Naval Research Laboratory as a physical chemist. He became a staff member at General Motors in 1960.

Dr. Hoare's sustaining interest has been in electrochemical kinetics and the mechanisms of electrode processes. He is best known to the scientific community for his basic studies of hydrogen and oxygen electrode mechanisms. His book, *The Electrochemistry of Oxygen,* published in 1968, is considered a work of primary importance to the field. In addition to his work on chromium plating, he is responsible for the fundamental research that helped make electrochemical machining a precision process.

© 1983 SCIENTIFIC AMERICAN, INC

ergy of the system, and so during the phase transition a certain quantity of energy called latent heat must be removed. It is possible, however, to cool a liquid below its normal freezing point if there are no sites that can serve as nuclei for the formation of crystals. The result is a new phase called a supercooled liquid; on contact with suitable nuclei such a liquid freezes but the release of latent heat can raise the temperature of the resulting solid.

In a similar way, Guth maintains, the symmetrical phase of matter and energy predicted by grand unified theories may have supercooled during the early stages in the evolution of the universe. The symmetrical phase might have been maintained in certain regions of the early universe at temperatures as low as 10^{22} degrees K. In the supercooled symmetrical phase matter and energy expand exponentially until the regions of symmetrical phase are converted into an asymmetrical phase, a relic of which we now perceive throughout the visible universe. During the phase transition the expanded region of symmetrical phase releases a tremendous quantity of latent heat, and the region is reheated to nearly 10^{27} degrees.

Although there remain several theoretical problems for the inflationary model, the model resolves at least three cosmological puzzles. One of them is the homogeneity of the cosmic microwave background radiation, which varies with direction by less than one part in 10,000. If the standard big-bang model is correct, regions of the sky separated by more than two degrees of arc have not been in contact throughout the entire history of the universe. No causal influence would have had time to propagate between them. Because the uniformity of the radiation cannot be accounted for by any physical interaction between distant regions, it must be introduced as an ad hoc hypothesis in the standard model. In the inflationary model such homogeneity is readily explained because the precursor of the present observable universe is much smaller than the one predicted by the standard model. Every part of the observable universe was at one time causally connected to every other part.

The second puzzle is called the flatness problem. A basic question in cosmology is whether the universe is open or closed: will it continue expanding forever, or is the density of matter sufficiently high for gravitational forces to halt the expansion and drag the universe back into a single point? If the density is exactly equal to the critical density that allows the expansion to continue forever, the universe is said to be flat: the geometry of space-time is not curved but instead approaches the Euclidean geometry of a plane.

The actual density of the universe has not been determined precisely, but it is known to be within one or two orders of magnitude of the critical density. For understanding the evolution of the early universe, however, the problem is not whether the present universe is open or closed but rather how the observed density came to be so close to the critical value. In the standard big-bang model the density of the early universe must be equal to the critical density to an accuracy of one part in 10^{49} in order to be consistent with the present range of values. In the inflationary model such precision is not necessary. During the exponential expansion of the universe the region occupied by the observable universe becomes increasingly flat, and the mass density naturally converges to the critical density. The observed universe is flat for much the same reason that the geometry of a small patch on the surface of a balloon approximates the geometry of a plane.

The third puzzle is a rather technical one: How can the production of the exotic particles called magnetic monopoles, which the grand unified theories say ought to exist, be reconciled with the big bang? Monopoles are generated during the phase transition from the symmetrical to the asymmetrical phase, and the number of monopoles generated is proportional to the rate of the phase transition. In the standard big-bang model the transition is so fast that monopoles are produced copiously; indeed, there are enough to imply an unreasonably high density of matter in the universe. In the inflationary model the phase transition is slower and the monopoles are not nearly as abundant.

The most serious difficulty for the inflationary model is accounting for the clumps of matter that are the most salient feature of the observable universe: galaxies, stars and planets. Small fluctuations in the density of the early universe either disappear entirely or collapse far too quickly to account for the observed distribution of matter. Galactic clumping, in the view of several investigators, may have come about later.

Autosplicer

In virtually all organisms genetic information is encoded in the nucleic acid DNA and is transcribed into three species of the similar nucleic acid RNA: messenger RNA, which is translated into protein, and ribosomal RNA and transfer RNA, which have roles in the process of translation. The various steps in transcription and translation are mediated by enzymes, the protein catalysts of biochemical reactions. There are enzymes called nucleases, for example, that cleave a nucleic acid into pieces and ligases that connect the pieces or form a single piece into a circular molecule. It has been taken for granted that for every such reaction there is an enzyme; nucleic acids do not cleave and ligate and circularize themselves by their own bootstraps. Or do they? Workers at the University of Colorado at Boulder have now found firm evidence that in the one-celled protozoan *Tetrahymena thermophila* there is an RNA that does just that.

In several species of *Tetrahymena* the genes coding for ribosomal RNA are split, that is, the coding region is interrupted by an intervening sequence (or "intron"). The entire gene is transcribed into a precursor RNA, which is then spliced: the intervening sequence is cut out of the precursor and the two long flanking sequences ("exons") it had interrupted are ligated, or joined by a covalent chemical bond, to form the mature ribosomal RNA. So far so good: since 1977 it has been known that in organisms higher than bacteria such intervening sequences interrupt many genes for messenger, transfer and ribosomal RNA. When Thomas R. Cech and his colleagues at Colorado studied the splicing process in *T. thermophila,* however, they found some surprises. One surprise was that the intervening sequence excised from the precursor RNA as a linear molecule is soon converted into a circular form. Moreover, sequence analysis revealed an extra guanosine nucleotide, one of the subunits of RNA, at the beginning (what is called the 5′, or upstream, end) of the linear molecule. It is extra in the sense that it is not encoded in the DNA from which the RNA is transcribed.

Strangest of all, when Cech and his co-workers tried to follow in detail the transcription from DNA to RNA, they found the precursor RNA was short-lived. Even in a laboratory system that should not have promoted splicing, the intervening sequence was quickly excised and circularized. No protein seemed to be involved in the process. By manipulating the concentration of salts in their preparation the investigators were eventually able to separate transcription and splicing, so that unspliced precursor RNA accumulated. When they added a supply of guanosine triphosphate (GTP), excision and circularization followed, even when proteins were carefully excluded from the preparation. The supplied GTP was shown to be mediating the excision by inserting itself at the junction between the upstream exon and the beginning of the intervening sequence. Another guanosine nucleotide at the downstream end of the intervening sequence then inserts itself near the 5′ end to circularize the excised molecule.

The Colorado group has now reported more rigorous evidence of autoexcision and autocircularization. Writing in *Cell,* Kelly Kruger, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling and Cech report on experi-

© 1983 SCIENTIFIC AMERICAN, INC

Pocket Big Savings On This Radio Shack TRS-80® Hand-Held Computer System

TRS-80 Pocket Computer Model PC-2

Cut $80 **199.95**

Was $279.95 in Cat. RSC-8

New Low Price! Now the powerful Radio Shack PC-2 (Cat. No. 26-3601) is more affordable than ever! Engineers, students, scientists—anyone with tough problems to solve—can put this TRS-80 computing system to work.

Unexpected Power. The PC-2's RAM is 2640 bytes big—expandable to over 10K. Our phenomenal Extended Pocket BASIC programming language gives you comprehensive handling of string variables up to 80 characters long, 2-dimension arrays, and 10 digit accuracy. It includes 42 statements, 34 functions, and 6 commands.

Superior Speed. The PC-2's 8-bit CPU operates at a clock speed of 1.3MHz—as fast as many desktop computers. Its advanced design allows direct memory access, and maskable and timer interrupts.

Additional Features. PC-2 has a built-in real-time quartz clock, accessible from either BASIC or the keyboard. The 26-character Liquid Crystal Display can produce full upper and lower case alphanumerics, special engineering and scientific characters or 7 × 156 dot matrix graphics. There's even a programmable beeper.

Printer/Plotter Cassette Interface

Cut $20 **219.95**

Was $239.95 in Cat. RSC-8

Flexibility With Peripherals. Add our unique Printer/Plotter/Dual Cassette Interface (26-3605)—also at a new low price—and plot full X/Y/Z axis graphics in four colors! Add two cassette recorders for automatic overlays, chaining and data storage under program control.

See It Today. Visit your nearest Radio Shack Computer Center, store, or participating dealer—where Pocket Computers start at just $149.95.

Radio Shack®

The biggest name in little computers®

A DIVISION OF TANDY CORPORATION

Send me your free TRS-80 Computer Catalog today!

Mail To: Radio Shack, Dept. 83-A-273
300 One Tandy Center, Fort Worth, Texas 76102

NAME ______________________

ADDRESS ______________________

CITY ______________ STATE ______ ZIP ______

PHONE ______________________

Retail prices may vary at individual stores and dealers.

© 1983 SCIENTIFIC AMERICAN, INC

ments in which the DNA of the intervening sequence and parts of its adjacent exons were inserted into a plasmid (a small circle of bacterial DNA). The purified plasmid DNA was transcribed with a bacterial polymerase, the enzyme that assembles RNA nucleotides on a DNA template. Then the polymerase was removed, so that no protein remained in the purified preparation. Again the intervening sequence excised itself and formed itself into a circle. Cech and his colleagues were also able to show that the two exons are ligated in the course of the same reaction that excises the intervening sequence. Several lines of evidence suggest that in many respects the *T. thermophila* intervening sequence functions as an enzyme. Cech and his colleagues call it a ribozyme.

How prevalent are ribozymes? No one knows. Few splicing enzymes have been identified and it could turn out that autocatalysis is widespread (although it does not seem likely to be responsible for the splicing of messenger-RNA precursors, which seems not to depend on the integrity of the intervening sequences). The discovery of just one instance of self-splicing is enough to stimulate new speculation about RNA, however. Unlike the double helix of DNA, a single strand of RNA can fold readily into a variety of shapes, and a typical enzyme is characterized primarily by a highly specific, complex shape. For several years the suspicion has been growing that some RNA's may have an active rather than a purely structural role in translation and some other cellular reactions. In addition to transfer RNA and ribosomal RNA there are many small RNA-protein complexes in cells, some of whose functions are just beginning to be glimpsed. Many people who think about the origin of life favor the hypothesis that RNA was the first genetic material, antedating DNA. The ability of primitive RNA's to catalyze their own rearrangement would presumably have given the early genetic material a large capacity for variation, the primary requirement for evolution.

Toward the Silicon Foundry

According to Gordon E. Moore of the Intel Corporation, the microelectronics industry obeys two laws of growth: the number of circuit elements that can be fitted on a chip of silicon doubles each year, and the cost of designing such a chip increases almost as fast as its complexity. A large-scale integrated circuit such as a microprocessor can have more than 450,000 circuit elements and can take two years or more to develop. The cost of designing the device can exceed that of fabricating all the chips made from the design.

In response to rising costs new and more structured approaches to integrated-circuit design are being developed. In addition computer-aided design systems are being introduced to relieve the designer of the need to describe an integrated circuit at its most primitive and detailed level. The result may be a major change in the semiconductor industry. System designers, who now build electronic equipment from the chips designed by commercial semiconductor manufacturers, may become designers of their own integrated circuits. And a proliferation of custom chips, tailor-made for specific low-volume applications, may to some extent replace general-purpose, high-volume chips.

At present electronic systems are customized largely by programming. A typical product consists of a few large-scale, programmable chips, such as microprocessors and memories, and many simpler chips needed to support the operation of the more complex ones; all the parts are assembled and interconnected on a printed-circuit board. The builder of the system adapts it to a specific purpose by choosing the parts and the interconnection scheme, but the system's function is determined primarily by programming the microprocessors and memories. What the system designer has not been able to do is alter the internal structure of a chip.

In the past five years a variety of semicustom technologies have begun to enable the system builder to intervene in the design of some types of integrated circuit. One semicustom device is the gate array, a chip on which thousands of logic gates (combinations of a few circuit elements that perform a simple logical function) are arranged in regular rows. The gates themselves are the same on all the chips, but the layers of metallic conductors that determine which gates are used and how they are connected can vary from one chip to another. The system designer defines the function of the chip by specifying the pattern of connections.

About 40 companies are now making gate arrays. Some of them provide computer-aided design systems that help to route the connections in a way that exploits the resources of the chip efficiently. A typical design system includes a library of the connection patterns corresponding to elementary units of electronic logic. The designer enters a description of the logical functions the device is to carry out and the system translates the description into a pattern of connections. The pattern serves as a template for the deposition of the metallic layers.

The building blocks of a gate array are logic elements. The building blocks of another semicustom design technology, called the standard cell, include logic elements, memories and central processing units. Some eight semiconductor manufacturers offer standard cells. The manufacturer defines the functions of a cell and creates a data base giving its characteristics, such as the time needed for a signal to propagate through it. Computer-aided systems assist the designer in combining various standard cells on a single chip. The designer breaks the system down into blocks corresponding to cells and creates network lists that specify the cells and how they are to be connected. The computer-aided design system positions the cells and routes the connections between them. Ultimately the system generates the data from which all the layers of the chip will be made.

Standard cells allow more intervention in the design of a chip than gate arrays do, but the system designer still cannot control the internal layout of the cells themselves. Carver A. Mead of the California Institute of Technology and Lynn A. Conway of the Xerox Palo Alto Research Center have created a structured approach to the custom design of integrated circuits that completely frees the system builder from dependence on the design expertise of semiconductor manufacturers. An automated design system supporting this approach would generate the entire layout of each chip rather than retrieving blocks of the layout from a data base.

Mead calls an automated design system of this kind a silicon compiler by analogy to the compiler program that supports a high-level computer language. The silicon compiler accepts a high-level functional definition of a device and translates it automatically into a chip layout in a standard format, much as a computer-language compiler translates programs written in a high-level language into instructions that can be executed by a computer. Several universities have developed silicon compilers and the first commercial one is to be introduced soon.

Silicon compilers could make it possible for a large population of engineers to design their own integrated circuits. As a result, Mead notes, a new kind of semiconductor manufacturer could begin to flourish: the silicon foundry. A silicon foundry would handle only the fabrication of a chip, leaving the design to its customers, much as a metal foundry casts parts designed by customers.

By processing many low-volume chip designs a silicon foundry would be able to compete with a commercial manufacturer designing and processing a few high-volume chips. Some "foundry" business is now done by commercial manufacturers, which interleave production of customer-designed chips with production of their own designs. True silicon foundries are also beginning to appear. As both silicon compilers and silicon foundries make custom design more economical, electronic systems for a variety of applications may

© 1983 SCIENTIFIC AMERICAN, INC

be systems on a single chip rather than systems assembled from many chips on a printed-circuit board.

Detumescence

When earth movements of any kind are reported in fault-checkered southern California, everyone listens. The citizens of Palmdale, a small agricultural community astride the San Andreas fault zone 35 miles northeast of Los Angeles, have been particularly attentive. In 1976 three geologists reported in *Science* that the town was nearly a foot higher than it had been in 1955. One of the authors of the report, Robert O. Castle of the U.S. Geological Survey, later revised his estimate of the degree of uplift to more than a foot. Half of the uplift, Castle noted, seemed to have taken place between 1961 and 1965. The movement could not be related to any local seismic activity, but the possibility was considered that it might be a sign of an impending earthquake.

Speculation about the cause of the Palmdale "bulge" has continued since then, but no convincing explanation has been given. Now Sandford R. Holdahl of the National Oceanic and Atmospheric Administration has suggested that no explanation is needed because in fact there has been no significant uplift. To account for the earlier findings he cites a bit of surveyors' wisdom propounded in 1981 by a fellow geophysicist, W. E. Strange: Do not conduct leveling surveys along railroad tracks in the southwestern U.S., where solar radiation levels are high.

Writing in *Journal of Geophysical Research,* Holdahl reconstructs the story as follows. Until 1964 the instrument used by surveyors in California for measuring elevation was good for sighting over long distances. It was customary for surveyors to make their measurements along railroad tracks, where the gradual slope allowed such long sightings. Surveyors even traveled from sighting to sighting on handcars. It had been known since the 1930's that the leveling procedure is susceptible to refraction errors: a temperature gradient in the air along the sight line can bend the path of light and give an inaccurate measure of elevation. Nevertheless, surveys in the U.S. were not corrected for refraction because the magnitude of the errors was thought to be small.

In most instances the errors probably are small. Both railroad ballast and the rails themselves, however, are efficient heat sinks; air-temperature gradients in their vicinity are large, particularly in the sunny Southwest. Surveys along tracks were thus exposed to significant refraction error, and the problem was compounded by the long sightings.

The introduction of new level instruments in 1964–65 resulted in a reduction of sighting distances and so of refraction error, even when surveys were done along rail rights-of-way. Holdahl recalculated the leveling measurements that had led to the proposal of a bulge at Palmdale, correcting for refraction error. The rise proved to be not a foot but somewhat less than three inches (plus or minus 1.6 inches). Even the slight remaining apparent uplift might be due at least in part to residual error.

Topological Perception

Current attempts to understand vision often begin by affirming that the brain analyzes images for their content of rather abstract features. A notable example is the detection of edges by nerve cells in the visual cortex. Another example may be the detection of topological properties such as connectedness. Writing in *Science,* Lin Chen, who holds positions at the China University of Science and Technology in Hofei and at the University of California at San Diego, suggests that the detection of topological properties may turn out to be "a primitive and general function of the visual system."

Chen reasoned from experiments showing that people readily distinguish a figure from its background even when "the detailed structure of the stimulus remains vague and amorphous." The distinction must have depended only on the overall properties of the stimulus. Such properties, Chen notes, "can be described mathematically as topological properties, such as connectivity." In topology a circle, a triangle and a square are equivalent. Any two points in such a figure can be connected by a straight line that stays entirely inside the figure. On the other hand, a circle, a triangle and a square are quite different from a ring or any other figure that has a hole in it.

In one experiment Chen had subjects press a button in order to see a stimulus consisting of a circle and a square, a circle and a triangle or a circle and a ring. Each stimulus appeared for only five milliseconds; then the subject was asked whether the two figures in it were the same or different. Under these circumstances the visual system proved to be most sensitive to the difference between the circle and the ring; they were "reported as 'different' on 64.5 percent of the presentations." The circle and the square were "different" on 43.5 percent and the circle and the triangle on only 38.5 percent.

Another experiment presented subjects with two glimpses of a short line segment. The position and the slope of the segment were changed from the first glimpse to the second; then the subjects were asked which segment was first in the sequence. They were correct on 59.2 percent of the trials. In some additional trials the stimuli included a stationary circle placed so that one line segment would be inside it whereas the other segment would be outside. Since the additional circle stayed "in the same position, this context gave no clue about which of the two target lines was presented." Nevertheless, the subjects were correct about the sequence of segments in 79.1 percent of the trials. Evidently a topological property (the placement of the segment inside or outside a closed curve) had facilitated the identification.

"We do not know," Chen observes, "how this resonance to [topology] is generated." Yet questions of that nature lie "at the core of issues in perception."

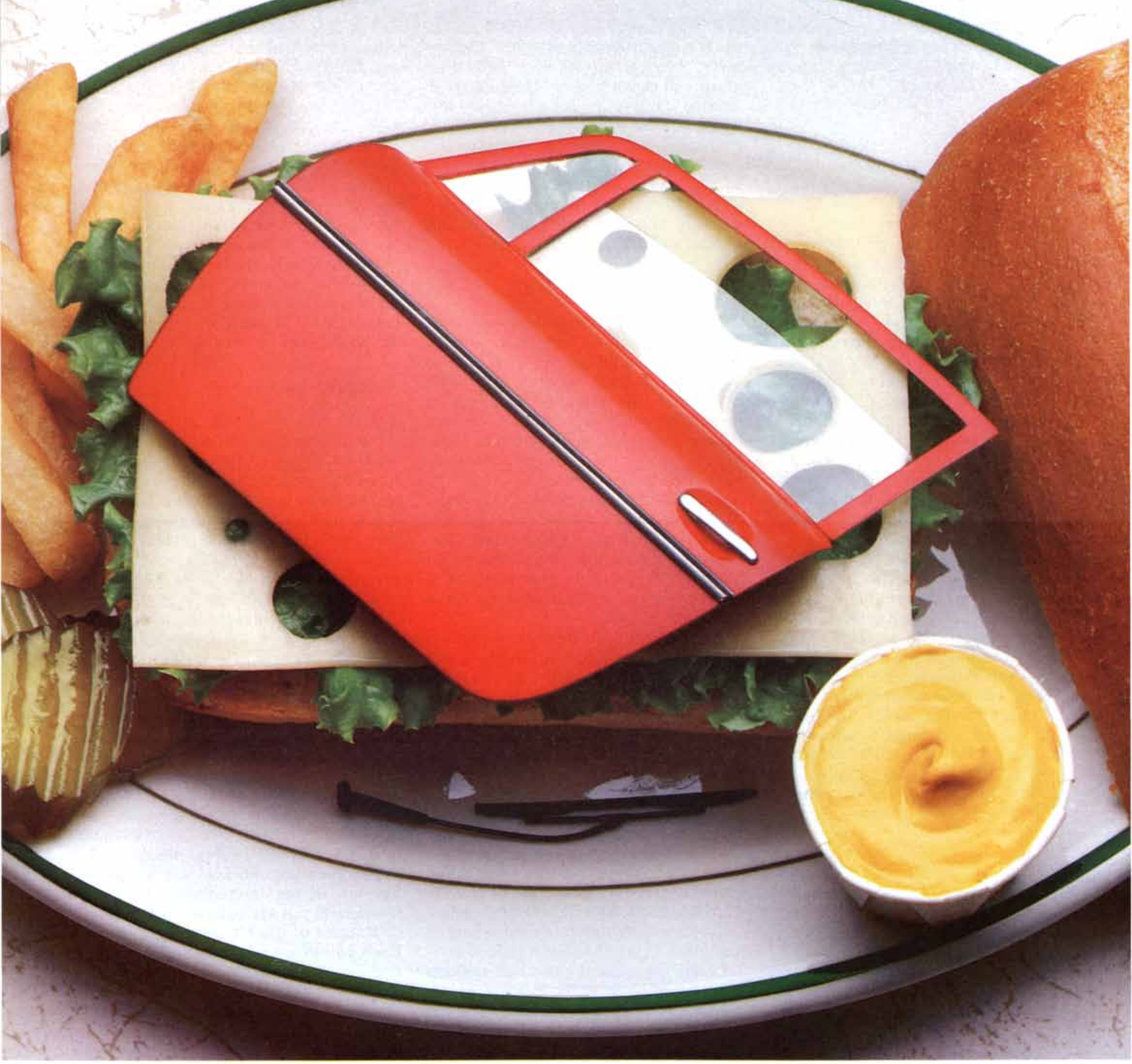
A Bigger Pi

Pi, the ratio of a circle's circumference to its diameter, can be expressed approximately as 3.14 and more precisely to an indefinite number of decimal places. By the end of the 16th century pi was known to 30 decimal places. Since then many people have extended the calculation. For almost a century the record was held by the British mathematician William Shanks, who over a period of 20 years laboriously calculated by hand until he had 707 decimal places of pi; his results were published in *Proceedings of the Royal Society* in 1873–74. (In 1945 it was found that Shanks had made a mistake at the 527th place.) With the advent of the electronic computer the number of known decimal places rose rapidly, to 100,000 in 1962 and to a million in 1973. Now within a few weeks two workers in Japan have calculated pi to 2^{21} (2,097,152), to 2^{22} (4,194,304) and to 2^{23} (8,388,608) decimal places.

The Japanese workers are Yoshiaki Tamura of the International Latitude Observatory at Mizusawa and Yasumasa Kanada of the University of Tokyo. Each employed a fast digital computer and an algorithm devised some years ago by Eugene Salamin of Stanford, Calif. Tamura calculated pi to more than two million decimal places in 7.2 hours of computer time. Kanada and Tamura together, using a faster computer, calculated it to more than four million decimal places in 2.9 hours of computer time. They went on to make the calculation to 2^{23} places, an effort that took 6.8 hours on the faster computer, a HITAC M-28OH. Kanada and Tamura plan to continue the calculation to 2^{24} (16,777,216) places by the middle of this year.

The chief spur to such calculations is the search for departures from randomness in pi. It is widely believed the sequence of digits in the decimal expression of pi has no pattern, but the conjecture has not been proved. Even if pi is random overall, small-scale patterns can be expected. For example, does the sequence 123456789 appear in pi? If it does, where?

© 1983 SCIENTIFIC AMERICAN, INC



How we keep cars from becoming a meal for the scrapyard.

At 3M we make a small army of products to assemble, fix, refinish and protect cars and help them live longer, healthier lives.

We make everything from structural adhesives that firmly anchor windows in place to abrasives that are specially designed to speed body repairs.

Also there's Scotchgard Protector for carpets and seat cushions inside. And Scotchcal Films to decorate autos, trucks and other vehicles outside.

In fact, by listening to people's needs, we've pioneered over 900 products to serve the needs of the transportation industry. We developed everything from electrical connectors to composite materials that form helicopter rotor blades.

All because at 3M, we're in the business of hearing. So let us hear from you.

3M hears you...

For your free 3M Transportation Brochure, write: Department 020702/ 3M, P.O. Box 4039, St. Paul, MN 55104.

Name________________

Address________________

City________ State____ Zip____

Or call toll-free: **1-800-323-1718,** Operator 362. (Illinois residents call 1-800-942-8881.)

3M

© 1983 SCIENTIFIC AMERICAN, INC

The Optical Computer

A computer based on beams of light rather than electric currents might be capable of a trillion operations per second. The crucial component, an optical analogue of the transistor, has been built

by Eitan Abraham, Colin T. Seaton and S. Desmond Smith

The development of the digital computer in the past 40 years is so closely associated with the development of electronic technology that people tend to think of the computer as an inherently electronic device. Actually the operations carried out by a computer are logical and arithmetic ones, and they could be done by any of several means. Since the mid-1970's it has become apparent that the potential exists for constructing a computing device in which signals are transmitted by beams of laser radiation rather than electric currents. There is a powerful incentive for developing such an optical computer: it might operate 1,000 times faster than an electronic computer.

The fundamental components of any digital computer are switches capable of two different states of transmission. The speed of the computer in its calculations is limited by the time required for a switch to change states, among other factors. In an electronic computer virtually all the switches are transistors, and even the fastest transistors now in use cannot be made to change states in less than about a nanosecond, or a billionth of a second. An optical device analogous to the transistor could switch from one transmission state to the other in about a picosecond, or a thousandth of a billionth of a second.

We have developed an experimental version of the optical transistor that is switched by a small change in the intensity of an incident beam of laser radiation. The optical transistor, which we call a transphasor, is based on a property of certain crystals. An increase in the intensity of light causes a change in the crystal's refractive index, a measure of the extent to which light is slowed as it passes through the material. With the right combination of crystalline material and laser wavelength it is possible to utilize the change in refractive index to make a device in which the intensity of the transmitted radiation changes greatly with a very small change in the intensity of the incident beam. With our experimental devices we have already achieved switching times that appear to be approximately a few picoseconds.

The optical transistor could be employed to build computers that would process information in much the same way as electronic machines do but faster. In the more distant future the capabilities of the optical transistor could transform the organization of the computer itself. Each optical transistor could be the site of many simultaneous switching operations carried out on parallel laser beams; an electronic device operates on one signal at a time. Moreover, the crystals we work with can switch to successively higher levels of transmitted power with successive increases in the incident beam; in contrast, an electronic transistor of the kind employed in computers has only two output states. The adoption of devices with more than two stable states might lead to a new system of computer logic.

Constructing an optical computer would require a variety of circuit elements in addition to the optical transistor. In our laboratory at Heriot-Watt University we have demonstrated in experimental form many of the elements needed to make an integrated optical circuit. Formidable technical difficulties must be overcome before such a circuit could be manufactured. Nevertheless, the optical computer is an intriguing prospect for the relatively near future.

The Functions of a Switch

In current practice the three basic functions of a computer—arithmetic operations, logical operations and the storage of information, or memory—are all done by devices that have two stable states. In arithmetic operations the two states represent the numerals 0 and 1 of the binary number system. In the evaluation of logical propositions the two states stand for true and false. The memory of the computer stores the results of arithmetic and logical operations in devices that occupy one of two states.

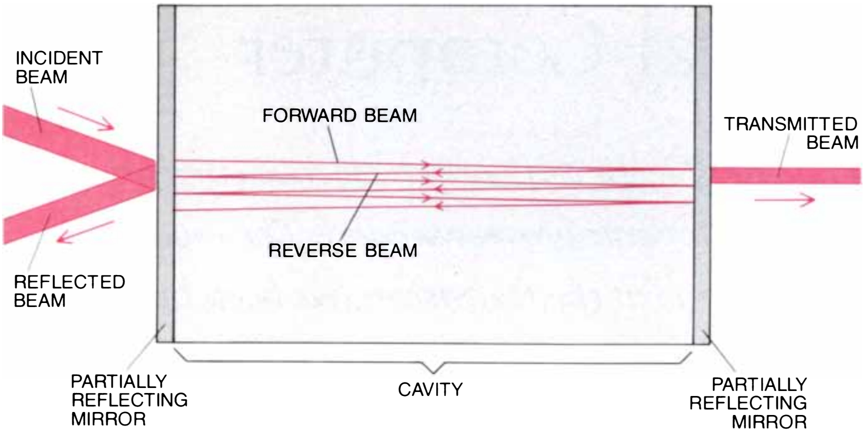
With the binary algebraic system a computer can evaluate the truth of propositions by making use of just three logical functions, which are usually referred to as the AND function, the OR function and the NOT function. In the AND function a statement is taken to be true if all its components are true. In the OR function a statement is taken to be true if any of its components is true. In the NOT function the truth value of a statement is reversed. More elaborate logical operations can be built up out of the three basic functions, and so can arithmetic operations such as addition.

Thus a computer requires a device that can represent the values 0 and 1 or true and false in physical form and can be assembled into larger-scale devices that perform the three logical functions. The device obviously must have two output states that can be readily distinguished. If the computer is to operate at high speed, the time needed for switching between the output states must be short. Other considerations suggest the most valuable switch is the one that is smallest, simplest to fabricate and requires the least power.

The first practical electronic switch was the vacuum tube. It was slow and large, and its operation generated a substantial amount of heat, which limited the operating life of the tube. As a result computers based on vacuum tubes were large in physical dimensions but small in computing capacity. More powerful machines would have required continuous replacement of defective tubes.

The invention of the transistor in 1947 provided a smaller, faster and more efficient switch. In its simplest form the transistor is a three-layered sandwich of semiconductor materials. (A semiconductor is a material whose electrical conductivity is between that of a conductor and that of an insulator; the electrical properties of a semiconductor are readily amenable to alteration.) The outer layers are called the emitter and the collector; the middle layer is the base. The electrical properties of the layers are chosen so that a slight change in the current flowing from the base to the collector causes a much larger change

© 1983 SCIENTIFIC AMERICAN, INC



FABRY-PEROT INTERFEROMETER, shown in schematic form, is the basis of optical switches constructed by the authors. The device consists of two plane mirrors placed parallel to each other with a transparent material inserted into the space between them. The space is called the cavity. Each mirror reflects part of the light that strikes it and transmits the rest. Thus when a beam is trained on the front mirror, some of the light is reflected and some passes into the cavity. The transmitted "forward" beam travels through the cavity and strikes the rear mirror, where some light passes out of the device and some is reflected as a "reverse" beam. The beam continues to rebound until all the light leaks out of the cavity. Changing the length of the cavity, the material in it or the wavelength of the light can cause substantial changes in the transmission of the interferometer. In an optical switch the medium in the cavity has properties that allow the transmission to be controlled by the intensity of the incident beam.

in the current from the emitter to the collector.

A large current from the collector can stand for a 1 and a small current can stand for a 0. By combining transistors and other circuit elements, structures that carry out the AND, OR and NOT functions can be assembled; the circuits are called logic gates. The AND gate yields a large current if all its inputs are large. The OR gate yields a large current if any of its inputs is large. The NOT gate yields a large current with a small input and a small current with a large input.

Limitations of the Transistor

Depending on their size, material, design and switching power, semiconductor logic gates require from about a nanosecond to 1,000 nanoseconds to operate. Ideally, then, the fastest gates could do as many as a billion logical operations per second; in practice computer processing speeds are considerably lower. Fundamental physical considerations suggest that the present minimum of about a nanosecond is close to the shortest switching time attainable with a semiconductor switching device.

For a transistor to switch, a current of electrons must traverse the base. There are well-defined limits to the speed with which electrons can move in a semiconductor. Reducing the width of the base can reduce the switching time, but there are also limits to how narrow the base can be made. The limits have probably been reached. Therefore any substantial reductions in switching time will result not from small improvements in transistor design but from the adoption of a new switching technology.

One possibility is to replace the current of electrons with a signal of another kind to carry information through the circuit. The highest velocity a signal that carries information can have is the speed of light, and so light (or electromagnetic radiation in another part of the spectrum) is a suitable candidate for a new information carrier. Following this line of reasoning we have utilized the coherent radiation from a laser to operate switching devices with many of the capabilities of the transistor. The devices can switch between two readily distinguishable output states; they can serve as memory elements, and by a suitable choice of materials and laser beams fast, simple logic gates can be assembled.

The starting point of the optical transistor is an ingenious and widely used piece of optical apparatus known as the Fabry-Perot interferometer, invented by the French physicists Charles Fabry and Alfred Perot in 1896. Fabry and Perot employed the interferometer to measure the wavelength of various colors of light; it has since been put to many other uses. The Fabry-Perot apparatus in its simplest form consists of two plane mirrors placed parallel to each other and separated by a space. A material that transmits radiation with the wavelength of interest can be inserted into the space, which is called the cavity.

Each of the mirrors partially reflects and partially transmits the light that falls on it. Partially reflecting mirrors are commonplace. Indeed, a plate-glass display window can act as such a mirror: the reflected light provides an image of the observer and the transmitted light provides an image of the goods behind the window. The strength of the images depends on the proportion of the incident light that is reflected and the proportion transmitted.

Ignoring for the moment the material in the cavity, consider what happens when a beam of light strikes the mirror that forms the front face of the interferometer. Assume that the mirrors reflect 90 percent of the incident light and transmit 10 percent. (Such proportions are quite close to those in some of our work.) Therefore when the beam of light strikes the front mirror, 90 percent of it is reflected and 10 percent passes into the interior of the interferometer.

The light, now with a tenth of its original intensity, travels to the rear mirror as the "forward" beam. The properties of the rear mirror are the same as those of the front mirror. Thus at the rear mirror 90 percent of the forward beam is reflected back into the cavity as the "reverse" beam and 10 percent is transmitted from the interferometer. Because the forward beam is a tenth as strong as the original incident beam the transmitted light has an intensity a hundredth as great as that of the original beam. The reverse beam, of course, travels back to the inner surface of the front mirror, where it is divided again into a transmitted part and a reflected part. The attenuated beam continues to rebound within the cavity, becoming progressively weaker, until all the incident light leaks out of the interferometer. Each passage makes a contribution to the transmitted beam, but the contributions after the first one are quite small; as an approximation it is reasonable to think of the transmitted light as being about a hundredth of the incident beam.

Interference in the Cavity

If the account given above captured all that goes on in the interferometer, the device would be of little interest as a switch for an optical computer. The intensity of the transmitted beam could be increased only by altering the properties of the mirrors. There is, however, an additional consideration. The forward and reverse beams cannot be separated in reality as neatly as we have separated them in our schematic description. In an actual interferometer the interaction of the beams influences the intensity of the light in the cavity, which in turn determines the transmission.

As is well known, light has some of the characteristics of a wave and some of those of a particle. Sometimes one description is more convenient and sometimes the other; in this context the wave description is more appropriate. The forward beam and the reverse beam along with the successive weaker rebounds can be thought of as waves propagating through the cavity. The incident beam enters the interferometer perpen-

© 1983 SCIENTIFIC AMERICAN, INC

dicular to the front face; therefore all the beams travel along the same path.

Keeping in mind that there are many reflected beams in the cavity, we can consider the interaction of just two of them: the forward beam and the reverse beam. The outcome of the interaction depends on the phase of the waves, that is, on the relative positions of the crests and the troughs. Crests can be aligned with crests and troughs with troughs in the relation known as constructive interference. Alternatively, the crests of one wave can be aligned with the troughs of the other in destructive interference. Of course, every phase between these extremes is also possible.

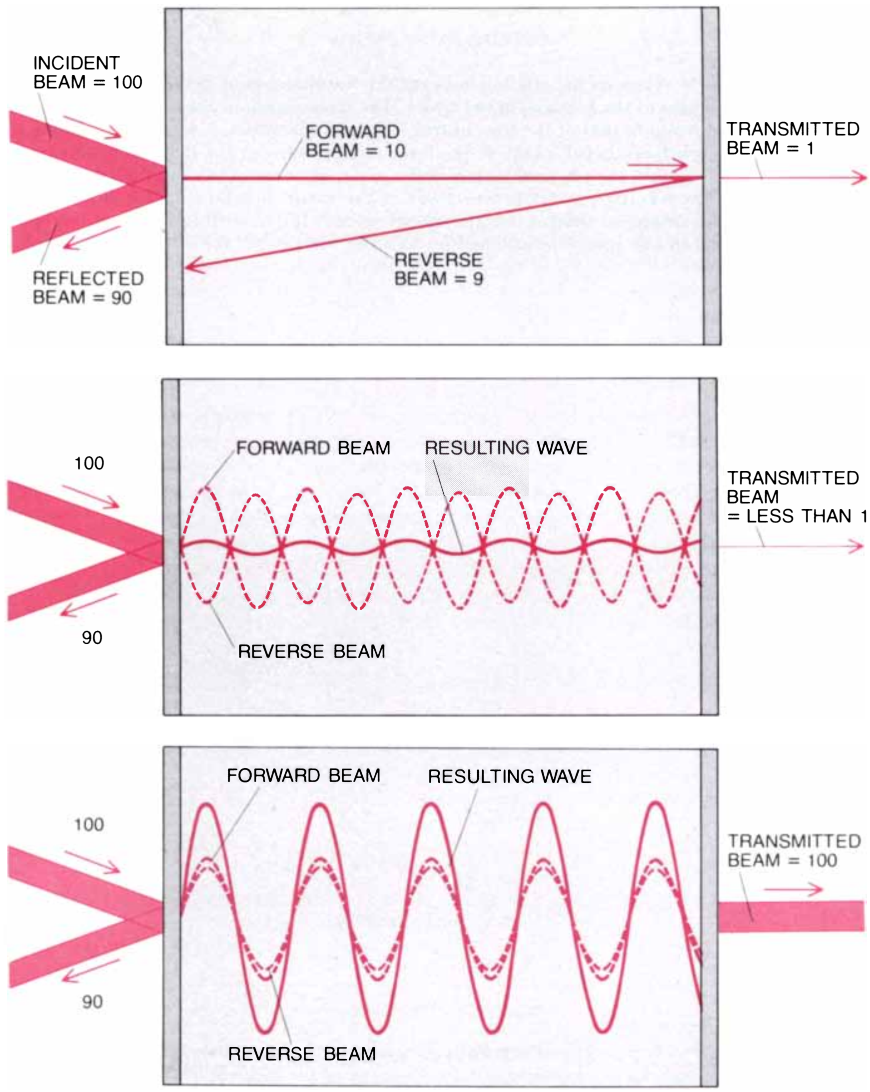
Through interference the two overlapping waves combine to form a third wave. The amplitude of the resulting wave at each point in the cavity is the sum of the amplitudes of the components at that point. If the interference is constructive, the amplitude is large; if the interference is destructive, the amplitude is small. The effect of interference on the transmission of the Fabry-Perot interferometer is far from trivial. At full destructive interference the intensity of the light in the cavity is almost zero and transmission is negligible. At full constructive interference the many forward and reverse beams reinforce one another and yield a light intensity in the cavity that is as much as 10 times the intensity of the incident beam. Since the rear mirror transmits 10 percent of the light reaching it, at full constructive interference the interferometer transmits a beam equal to the incident beam.

Effect of the Refractive Index

The phase of the forward and reverse beams, and hence the transmission of the interferometer, can be adjusted by changing the length of the cavity or the wavelength of the incident light. Neither method is of much value for the optical transistor. What is more important, the phase relations can also be changed by altering the optical properties of the material in the cavity, which up to this point we have passed over.

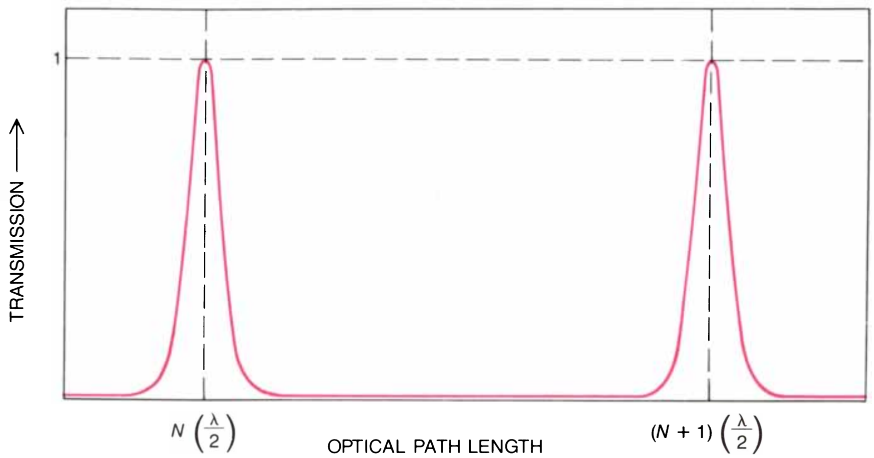
Light attains its maximum speed of propagation in a vacuum. In a transparent medium other than a vacuum the light is slowed by an amount that varies with the medium. It is the slowing of the light wave that gives rise to the familiar phenomenon of refraction; the refractive index of a material expresses the ratio of the speed of light in a vacuum to its speed in the material. For example, if the refractive index of a certain kind of glass is 1.5, then whereas light travels at about 300,000 kilometers per second in a vacuum, its speed is about 200,000 kilometers per second in the glass. The slowing of the light is accompanied by a reduction in the wavelength, but the frequency does not change. It can readily be seen that even without a change in the length of the cavity or the wavelength of the incident light a change in the refractive index of the material in the cavity will alter the phase relations of the beams in it by altering the wavelength. By selecting a material with the right refractive index the forward and reverse beams can be made to interfere either constructively or destructively.

An important quantity in analyzing an optical system is what is termed the optical path length, or simply the optical length. The optical length of the medium in the interferometer cavity is its physical length multiplied by its refractive index. It can be shown that full constructive interference results when the optical length is equal to the distance represented by an integral number of half wavelengths of the incident light. (A half wavelength, of course, is the distance from a crest to a trough.) There are thus many possible optical lengths that yield full constructive interference. When the optical length is halfway between distances corresponding to integral multiples of the half wavelength,

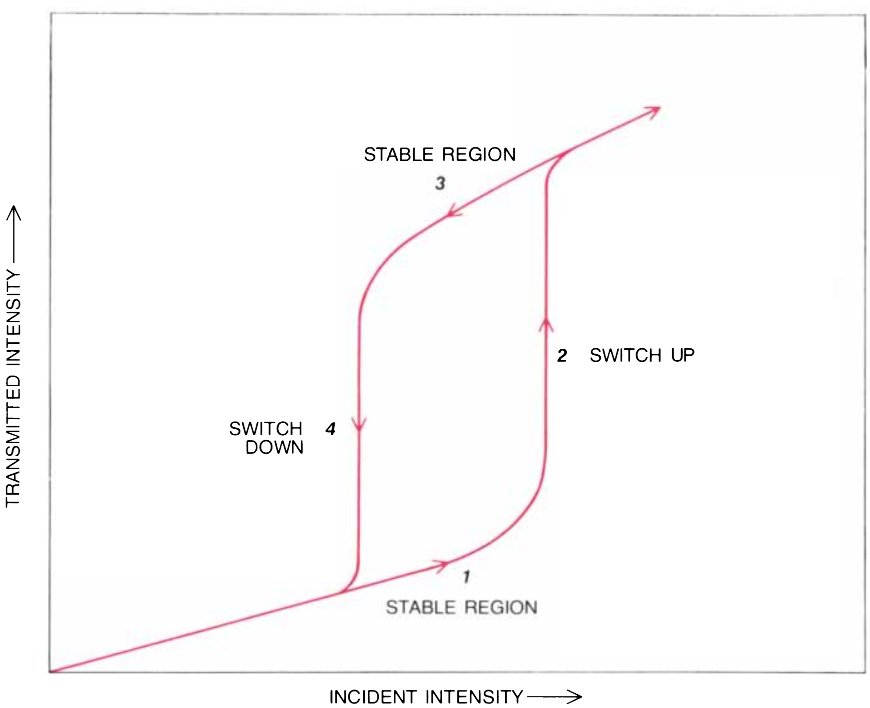


INTERFERENCE of the forward and reverse beams in the Fabry-Perot cavity can greatly change the light intensity within it and hence the transmission of the device. The front and rear mirrors can be chosen so that they reflect 90 percent of the light striking them and transmit 10 percent. Without considering interference the situation is as follows (*top*). Ten percent of the incident beam passes into the cavity. When the attenuated beam strikes the rear mirror, a beam with 1 percent of the original intensity is transmitted; the reverse beam is equal to 9 percent of the original one. Actually the forward and reverse beams interact as waves to yield a wave equal at each point to the sum of the two components. If the crests of one beam are aligned with the troughs of the other in destructive interference, the resulting wave is small (*middle*). Light intensity in the cavity is negligible and transmission can be less than 1 percent of the incident beam. If the two waves are aligned in constructive interference, the resulting wave is large (*bottom*). Because there are many forward and reverse beams, the intensity in the cavity can be as great as 10 times that of the incident beam. Thus when the interference is fully constructive, the transmitted beam can be equal in intensity to the original incident beam.

© 1983 SCIENTIFIC AMERICAN, INC



AIRY FUNCTION expresses the relation between the transmission of the interferometer and the optical path length of the material in the cavity. The transmission is the ratio of the intensity of the incident beam to that of the transmitted beam; its maximum is 1. Transmission is at the maximum when there is full constructive interference between the beams in the cavity. The optical path length is the physical length of the material in the cavity multiplied by its refractive index. (The refractive index is a measure of the extent to which light is slowed as it passes through a transparent substance.) The interference is fully constructive when the optical length is equal to any integer multiplied by half the wavelength (λ) of the incident light.



HYSTERESIS CYCLE represents the way the intensity of the beam transmitted by the interferometer changes in response to changes in the intensity of the incident beam. In the cavity is a material with a nonlinear refractive index: one that varies according to the intensity of the light in the material. Changes in the incident beam alter the refractive index and the optical length and hence the transmission. If the intensity of the incident beam is increased slowly, the transmitted intensity initially shows a gradual increase (*1*). The change in refractive index, however, soon brings the optical path length near the value corresponding to constructive interference, thereby increasing the light intensity in the cavity; the change in intensity further alters the refractive index. At a certain point this positive feedback becomes quite strong. A threshold is reached and transmission increases rapidly (*2*). If the incident intensity is then reduced somewhat, transmission does not fall to the original level because there is still enough light in the cavity to keep the refractive index and the optical path length near the values corresponding to maximum transmission; transmission falls gradually (*3*). At a second threshold, however, further dimming of the incident beam changes the refractive index enough to cause a greater loss of intensity in the cavity; transmission then falls rapidly to the initial low level (*4*). An optical device that has a hysteresis cycle is said to be optically bistable. Regions *1* and *3* of the curve are the stable regions, where transmission changes little with changes in intensity.

the result is full destructive interference.

Since the intensity of the light transmitted by the interferometer varies according to the phase relations of the beams, it follows that the transmitted intensity varies with the optical length. The function that describes the relation between the optical length and the transmission of the interferometer (defined as the ratio of the strength of the transmitted beam to that of the incident beam) is called an Airy function and is quite important in the construction of an optical transistor. In the Fabry-Perot devices we have been building the function has a steep peak at each point where the optical length corresponds to an integral number of half wavelengths; between the peaks the value of the function is small, and it changes only gradually. Therefore as the optical length approaches a value corresponding to peak transmission, the transmission increases slowly until a threshold is reached and then increases very rapidly.

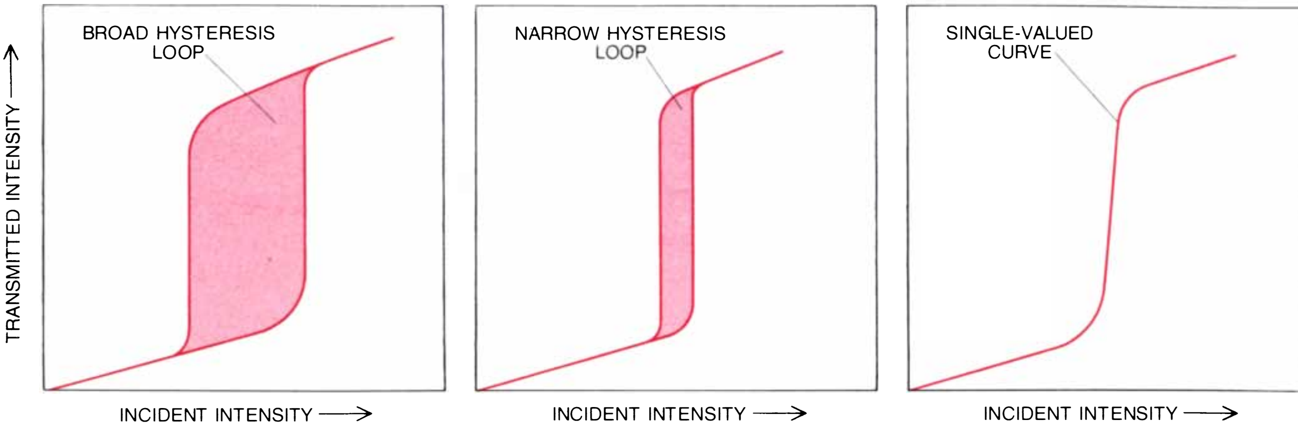
The effect of the refractive index on the optical length implies that the transmission of a cavity can be set at some fixed level simply by inserting into the cavity a material with the appropriate refractive index. Moreover, exchanging one material for another would alter the transmission. Of course, such a method of switching would be quite clumsy in a computer. In the 19th century, however, it would have been considered the only means of altering the refractive index. James Clerk Maxwell, in his fundamental work on electromagnetic radiation, postulated that refraction and other interactions of light with matter are independent of the intensity of the light. If this were always true, the refractive index would be a constant and the output of the interferometer would be directly proportional to the intensity of the incident beam. Under the appropriate conditions many substances yield such a result; they are said to be linear with respect to the refractive index.

Nonlinear Refraction

When the development of lasers made available a powerful source of coherent radiation (that is, radiation in which all crests and troughs are aligned), it became apparent that not all materials have a linear refractive index. When laser radiation is focused on certain substances, the refraction varies with the intensity of the beam; the materials are nonlinear with respect to the refractive index. With a nonlinear substance in the Fabry-Perot cavity the refractive index can be altered merely by changing the intensity of the incident beam.

The effect is subtle enough to be considered in detail. Imagine a cavity of fixed length with a nonlinear material in it; a beam of laser radiation whose intensity can be adjusted is directed into the

© 1983 SCIENTIFIC AMERICAN, INC



HYSTERESIS LOOP CAN BE ALTERED in an optically bistable device by changing the length of the interferometer, the wavelength of the incident beam or the material in the cavity. In the basic hysteresis cycle the stable regions are wide and the loop is broad (*left*). A given level of incident intensity can correspond to either one of two levels of transmitted intensity. By the appropriate manipulation of the interferometer the stable regions can be reduced and the loop narrowed (*middle*). The limit of the process is a single-valued curve, in which each level of incident intensity corresponds to only one level of transmitted intensity (*right*). The curve is similar to the transfer-characteristic curve of an electronic transistor; an interferometer with such a transmission curve is utilized in the optical transistor.

cavity. Initially the intensity of the beam is such that the refractive index yields an optical length in the region of the Airy curve between peaks; transmission is low. If the intensity of the beam is slowly increased, the refractive index and the optical length gradually change; transmission increases slowly. At some point the refractive index and the light intensity in the cavity form a positive-feedback loop, that is, they become mutually reinforcing. The changing refractive index brings the beams closer to constructive interference, thereby augmenting the intensity in the cavity; the increased intensity further changes the refractive index. At a certain critical level of incident intensity the feedback effect becomes quite strong. The steep region of the Airy curve is reached and transmission increases very rapidly indeed.

At the peak of the Airy function the ratio of transmitted intensity to incident intensity is 1. When the intensity of the laser beam is reduced somewhat from the peak value, the transmission of the cavity does not fall to its initial low level. The reason is that the light intensity in the cavity is sufficient to keep the refractive index and the optical length near the values that yield maximum transmission. As the intensity of the incident beam is reduced further, transmission still decreases only gradually. At a certain point, however, the refractive index and the intensity in the cavity become mutually attenuating. A small decrease in the incident intensity changes the refractive index enough to reduce the intensity in the cavity substantially. Hence there is a sudden sharp decrease in transmission.

If the intensity of the transmitted radiation is plotted against the intensity of the incident beam, the resulting curve is a loop with four sections [*see bottom illustration on opposite page*]. Beginning at a low incident intensity the curve rises gradually: even a large increase in incident intensity causes only a small gain in transmission. At the threshold the curve becomes steep: a small increase in intensity brings a sharp gain in transmission. The return to the starting point follows a different path. Initially the curve slopes downward only slightly, and even if the intensity is reduced, the transmission remains high. When a second threshold is reached, however, the curve falls precipitously: a small loss of intensity greatly attenuates the transmission.

The loop constitutes what is called a hysteresis cycle. An optical device that exhibits such a cycle is said to be optically bistable because it has two stable regions where the transmitted intensity varies little with changes in the incident intensity. It should be noted that in a bistable device a particular value of incident intensity can be associated with two levels of transmitted intensity; the actual transmission depends on the previous condition of the cavity.

Bistable Optical Devices

The existence of optical bistability was predicted by Abraham Szöke and his colleagues at the Massachusetts Institute of Technology in 1969. It was first observed by Hyatt M. Gibbs, Samuel L. McCall and Thirumalai N. C. Venkatesan of Bell Laboratories in 1976. Since then bistable optical devices have been the subject of work in many laboratories.

In making an optical bistable device it is important to choose a material with a substantial nonlinearity in its refractive index. In our work we have utilized the compound semiconductor indium antimonide, which has a highly nonlinear response. Like most semiconductors, indium antimonide is opaque to radiation in the visible part of the spectrum, but it is transparent to infrared radiation of some wavelengths. Our source of radiation is a carbon monoxide laser whose wavelength can be adjusted over a narrow region in the infrared.

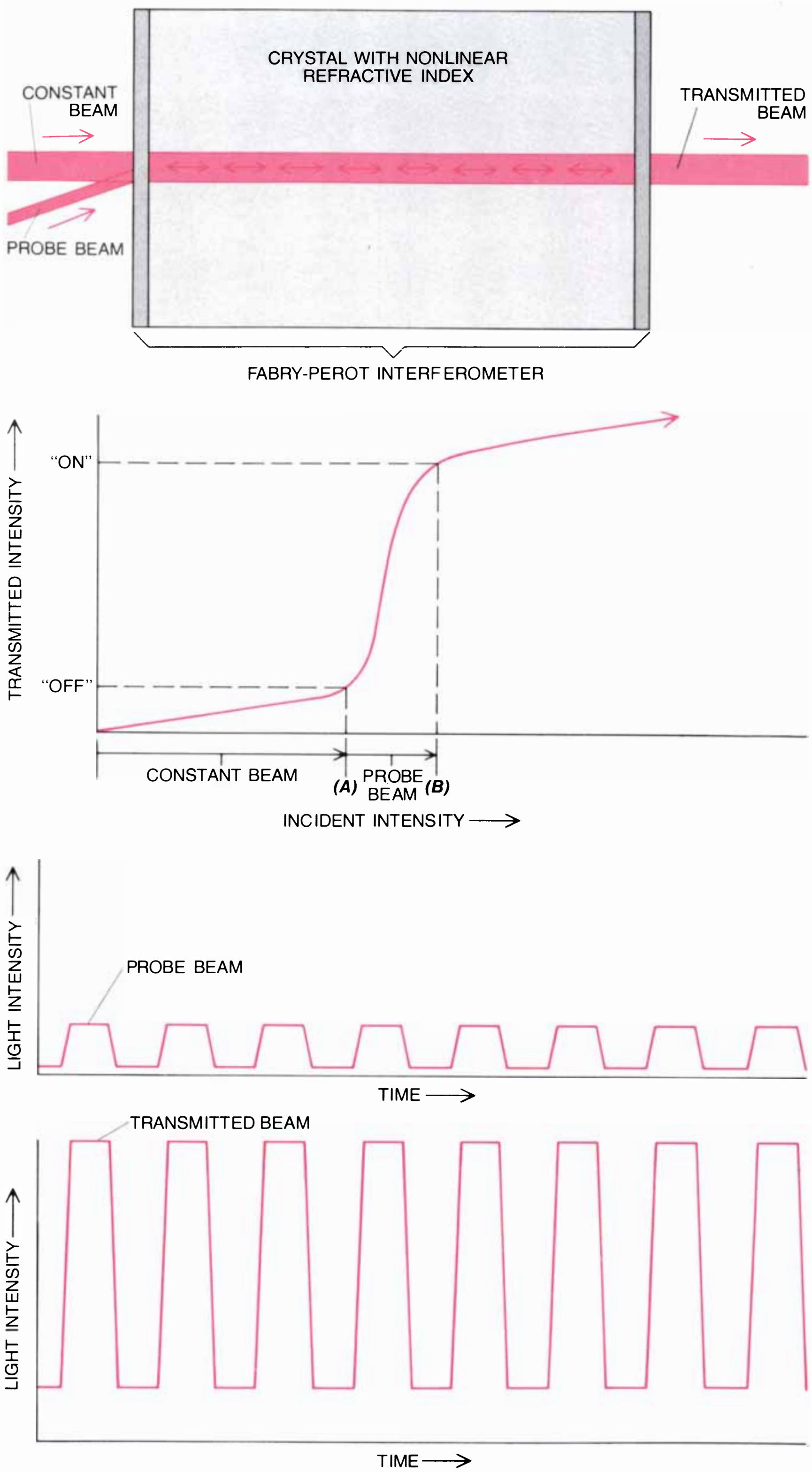
The interferometer we employ is a small rectangular crystal of indium antimonide. Generally each edge of the crystal is a few millimeters long; we have also built devices out of much smaller crystals. In some cases the front and rear faces of the crystal are coated with a reflective substance that serves as the mirror, but the coating is not essential: the polished face of the crystal itself can serve as the mirror. A device beyond the rear mirror measures the intensity of the transmitted beam.

On the atomic scale there are many mechanisms that can change the refractive index of a semiconductor. The mechanism in indium antimonide is complex and indeed is still being worked out. Some aspects of the process are understood; the following is a simplified account of what is known.

The interaction of electromagnetic radiation with a solid material is mediated almost entirely by the electrons in the solid; at the wavelengths of visible light or infrared radiation it is mainly the outermost electrons of each atom that participate. In a good conductor of electricity, such as a metal, the outermost electrons are free to roam throughout the volume of the material; in a good insulator each electron is bound tightly to a particular atom and can be dislodged only at a high cost in energy. The situation in a semiconductor is intermediate between these extremes: each of the outermost electrons is bound to an atom but only a small input of energy is needed to set an electron free.

These differences in electronic configuration, which determine the optical as

© 1983 SCIENTIFIC AMERICAN, INC



OPTICAL TRANSISTOR devised by the authors relies on two precisely adjusted laser beams focused on the front face of a Fabry-Perot interferometer containing a substance with a nonlinear refractive index (*top*). The "constant" beam is strong and unvarying. The "probe" beam is weaker and can be modulated. The intensity of the constant beam (*A*) is chosen to yield a low transmitted intensity that is just below the steep region of the single-valued transmission curve (*middle*). The intensity of the probe beam (*B*) is chosen so that when it is added to the intensity of the constant beam, the transmission increases to the high level at the top of the steep region. The low-transmission or "off" state can stand for the value 0 in the binary logic of the computer; the high-transmission or "on" state can stand for 1. Because of the steepness of the transmission curve, small changes in the probe beam can cause large changes in the transmitted intensity (*bottom*). In this respect the optical transistor is analogous to the electronic transistor.

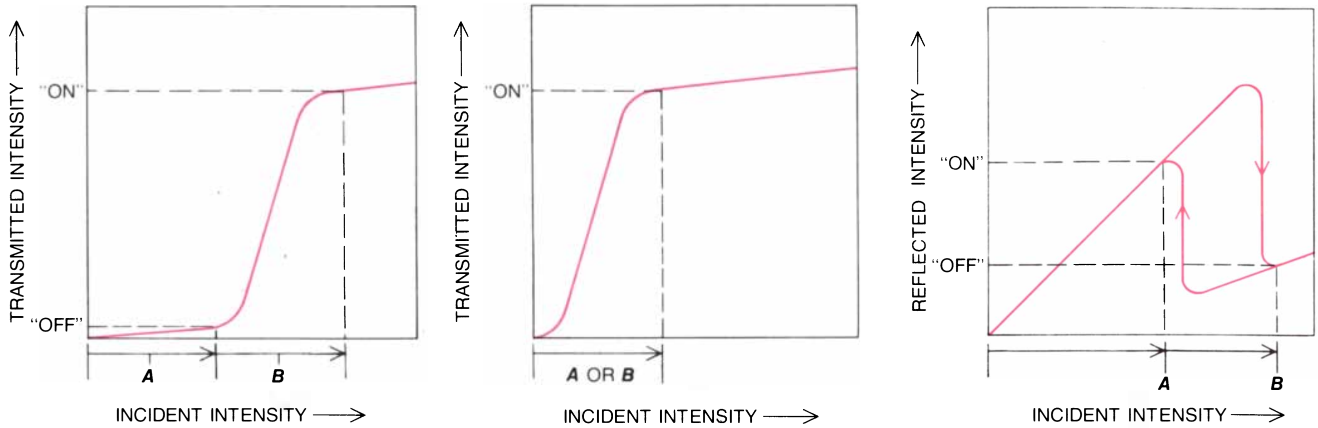
well as the electrical properties of the material, can be understood by considering the energy levels available to each electron. In an isolated atom the levels are sharply defined, but in a solid the levels associated with the many atoms stack up, so to speak, to form broad bands of allowed energies separated by forbidden band gaps, representing energies no electron can have. If a band is completely filled, meaning that every level in it is occupied, each electron in the band is confined to the vicinity of an atom. The reason is that a little extra energy is needed to set an electron in motion, and there is no vacant level in the band to which an electron with extra energy could move. If a band is only partly occupied, an electron that acquires extra energy can readily move to a slightly higher level in the band and is thereby set free.

Semiconductor Band Structure

A metal has a partly occupied band. In an insulator the highest occupied band (the valence band) is filled exactly, and it is separated by a wide gap from the next higher band (the conduction band), which is empty. An electron could be set in motion in an insulator only if it absorbed enough energy to cross the gap, an unlikely event. A semiconductor also has a filled valence band and an empty conduction band, but the gap between them is smaller. There are several ways an electron in a semiconductor can gain the rather small amount of energy needed to go from one of the higher levels of the valence band to one of the lower levels of the conduction band. When an electron is promoted to the conduction band, it leaves behind it a "hole" in the valence band, which in some respects acts as an entity with a positive charge equal in magnitude to the negative charge of the electron. When the promoted electron returns to the conduction band, it joins a hole and the hole is eliminated; the process is called recombination.

The energy required for an electron to traverse the band gap varies with temperature. Much of our work with indium antimonide has been done at 77 degrees Kelvin. At this temperature the band-gap energy of indium antimonide is about .2 electron volt. (An electron volt, abbreviated eV, is the energy an electron gains when it is accelerated through a potential difference of one volt.) Most other semiconductors have a much wider band gap. The nonlinearity of the refractive index increases in inverse proportion to the square of the band-gap energy. Because of its narrow gap indium antimonide has a nonlinearity between 100 and 1,000 times as great as that of gallium arsenide, another semiconductor that has been utilized in bistable optical devices. Although a

© 1983 SCIENTIFIC AMERICAN, INC



OPTICAL LOGIC GATES have been constructed from optical transistors. A logic gate is a device employed in a digital computer for arithmetic operations and for evaluating the truth of propositions. Three kinds of gate are required; the principles according to which optical versions of them have been constructed are shown. The AND gate yields a high output only when all its inputs are present (*left*). If the two incident laser beams are chosen so that each one yields a transmitted intensity just below the steep part of the curve, the result is an AND gate: both incident beams must be present for high transmission. The OR gate yields a high output if any of its inputs is high (*middle*). If the incident beams are adjusted so that each one yields a transmission at the top of the steep region, the result is an OR gate: either incident beam can yield high transmission. The NOT gate has a high output when the input is low and a low output when the input is high. The optical NOT gate exploits the light reflected from the interferometer (*right*). The reflected beam is associated with the hysteresis loop shown. Because of the shape of the loop, a high input (*B*) yields a low output whereas a low input (*A*) yields a high output.

narrow band gap is desirable, if the gap is too narrow, operation of the optical switch becomes impractical.

One way for an electron to acquire the energy needed for promotion to the conduction band is by absorbing incident radiation. Here it is convenient to adopt the description of light as a stream of particles. The particle, or quantum, of light is the photon. The energy of a photon is inversely proportional to the corresponding wavelength. When a photon is absorbed by the crystal, the photon's energy can be added to that of an electron. If the photon carries the right amount of energy, the electron is promoted to the conduction band. The wavelength equivalent to .2 eV is about five micrometers, which is in the infrared part of the spectrum and within the range of our carbon monoxide laser.

Adjusting the laser to a wavelength of about five micrometers therefore leads to the promotion of many electrons into the conduction band. Once the electrons reach the lower levels of the conduction band the thermal energy of the crystal lattice causes them to scatter to nearby energy levels. The result is an irregular distribution of electrons in the lower reaches of the conduction band.

It is the promotion and subsequent scattering of the electrons that causes the refractive index of indium antimonide to change. In the quantum-mechanical account of how light interacts with matter, refraction and absorption are closely connected. Indeed, each entails the other: if a beam of radiation is absorbed, it is also refracted, and vice versa, although the two phenomena may affect radiation of different wavelengths within the beam. Preventing absorption from taking place can thus simultaneously prevent refraction from taking place and thereby alter the refractive index. How is absorption to be prevented? There is a limit to the number of electrons that can be promoted to the lower levels of the conduction band. Once the limit has been reached no additional photons with the precise energy needed to accomplish those promotions can be absorbed; the absorption of such photons is said to be saturated. The saturation of the absorption forestalls the accompanying refraction and the refractive index changes. Very soon after the laser beam is turned off the promoted electrons return to the valence band and recombine with holes. Absorption then becomes possible again and the refractive index reverts to its original value.

Optical Logic Gates

We have taken advantage of the extreme optical properties of indium antimonide that result from this mechanism to make components for an optical computer. The hysteresis cycle described above has an obvious application as a memory element. The stable high-transmission state of the interferometer can represent one logical or arithmetic value and the stable low-transmission state the other value. Either state can be maintained indefinitely by a beam of intermediate intensity.

For an optical switch to be used in processing information rather than storing it hysteresis is not needed. By changing the optical setup slightly the hysteresis loop can be narrowed or eliminated altogether. The result is a single-valued transmission curve, in which each level of incident intensity is associated with only one level of transmitted intensity. The shape of the single-valued curve can itself be manipulated. In one of the most useful shapes transmission is low and almost constant at low incident intensity, then at a threshold the transmission rises steeply, reaching a high level that again remains almost constant as the incident intensity increases further. The curve resembles the one that describes the current from the collector of an electronic transistor. Indeed, the curve is the basis of the optical transistor. We call the device the transphasor because its operation is based on controlling the phase of the light within it.

In the transphasor two precisely adjusted laser beams are focused on the same spot on the front face of a crystal of a nonlinear material such as indium antimonide. The "constant" beam has an intensity that is large and unvarying. The "probe" beam has an intensity that is much smaller and can be modulated. The strength of the constant beam is chosen so that it yields a transmitted intensity just below the steep region of the transmission curve. The intensity of the probe beam is such that when it is added to the intensity of the constant beam, the steep part of the curve is traversed and the transmission peak is reached.

Because the transmission curve of the transphasor is steep a very small modulation of the probe beam controls a very large increase in the transmitted intensity of the device. The analogy with the electronic transistor can be extended. The constant beam is analogous to the constant "bias" current that flows from the emitter to the collector of the transistor. The probe beam is analogous to the smaller current from the base to the collector. Just as a small change in the base current enables the transistor to conduct a much larger collector current, so a small change in the probe beam enables the transphasor to transmit a

© 1983 SCIENTIFIC AMERICAN, INC

larger fraction of the incident radiation.

Like the transistor, the transphasor can be switched between two readily distinguishable states. The switching is extremely fast. Although we have not been able to measure the switching time directly, the indirect evidence suggests it is a few picoseconds.

Once an optical transistor has been constructed, the assembly of the prototype of a logic gate is a straightforward procedure. Indeed, a single transphasor could serve as either an AND gate or an OR gate depending on the beams supplied to it. If the two incident beams are selected so that neither beam alone is sufficient to switch the device but together they are strong enough to do so, the result is an optical AND gate. Transmission is high only when both incident beams are present. If the incident beams are chosen so that either one is capable of switching the device, the result is an optical OR gate. The transmitted intensity is high when either beam is present.

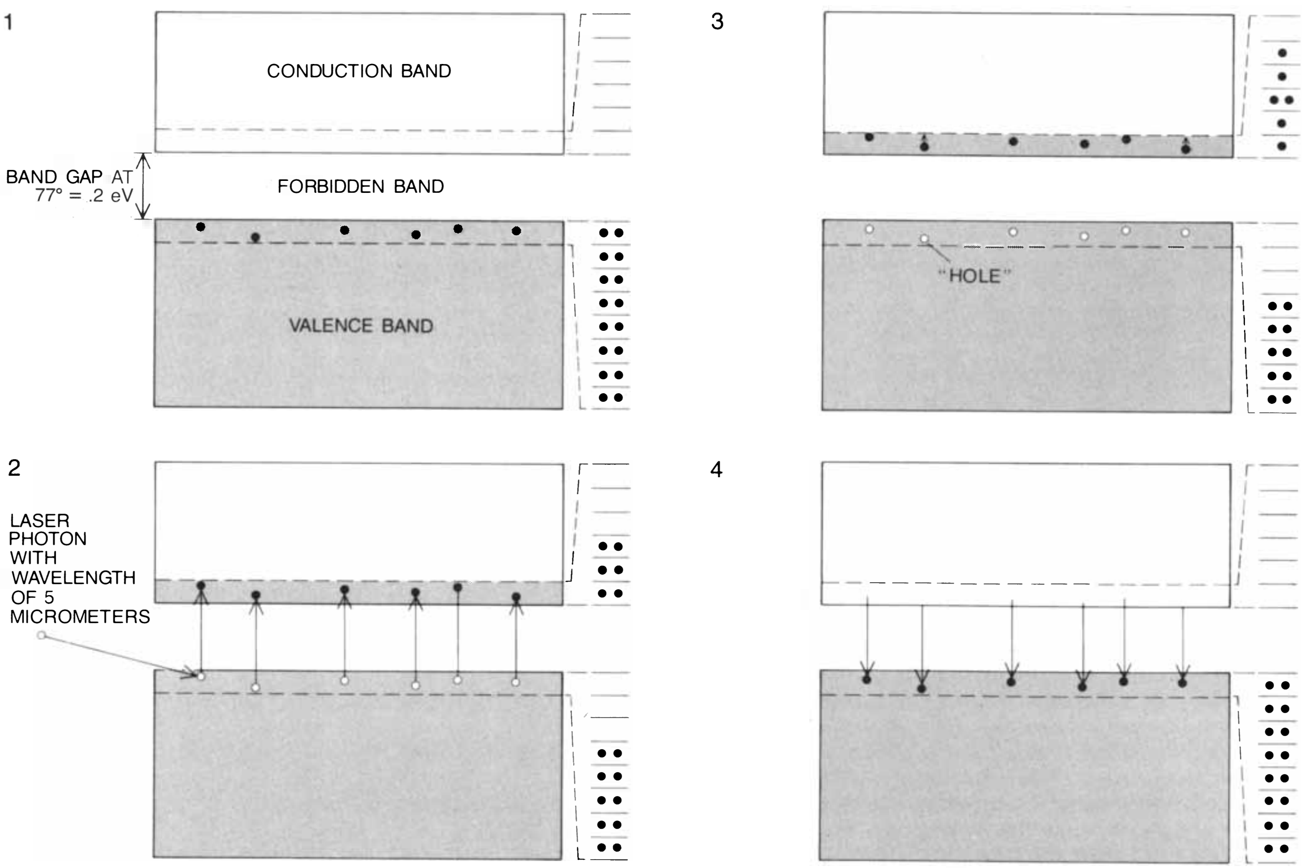
A NOT gate can be constructed by utilizing the reflected beam as the output. Since the reflected beam is the inverse of the transmitted one, increasing the incident intensity will yield low output. Decreasing the incident beam will yield high output.

Toward Photonic Computing

To build an optical computer many optical circuit elements would have to be assembled and interconnected. In the 1970's considerable work was done with thin films through which laser radiation can be made to propagate. The waveguides could in principle form the connecting elements in integrated optical circuits, although the fabrication of such circuits has not yet been attempted.

One application of the new optical technology would be the construction of an optical computer with an architecture, or logical organization, much like that of the standard electronic machine. The chief value of a computer with a conventional architecture but optical components would be its great speed. Optical logic gates could carry out 1,000 billion logical operations per second compared with a maximum of one billion for electronic switching devices.

Optical switches have properties that could also make them suitable for new computer designs. Whereas several currents in a single transistor inevitably become mixed, multiple beams can be passed through an interferometer and still remain separate. It would therefore be possible to utilize a single crystal for several separate switching operations. There could be, say, five laser paths through a crystal, each path forming the site of one operation. With the selection of the appropriate beam intensities each operation could be different. For example, two paths might function as AND



REFRACTIVE INDEX of indium antimonide changes in response to changes in the intensity of radiation passing through it. Indium antimonide is the semiconductor utilized by the authors as the nonlinear material in the optical transistor. In a semiconductor crystal the energy levels available to an electron are arranged in bands; within a band there are many closely spaced levels, each of which can be occupied by two electrons, but energies in the gaps between the bands are forbidden. Here two bands are shown: at low temperature the lower band (called the valence band) is full and the upper band (the conduction band) is empty (*1*). To go from the valence band to the conduction band in the process called promotion an electron must acquire energy equivalent to that represented by the band gap; in indium antimonide the energy is about .2 electron volt. The energy can be acquired by absorption of a photon with a wavelength of five micrometers, which is in the infrared. The authors employed a carbon monoxide laser with an adjustable wavelength to supply such photons (*2*). When an electron absorbs a photon and is promoted to the conduction band, it leaves behind it in the valence band a "hole" that acts like a positively charged particle. Thermal and vibrational energy quickly scatter electrons among the lower levels of the conduction band (*3*). When these levels are filled, no more photons with a wavelength of five micrometers can be absorbed. The blocking of the absorption leads to a concomitant change in the refractive index of the crystal. When the incident beam is switched off, the electrons return to the valence band and recombine with holes, which are thereby eliminated (*4*). The refractive index returns to its original value.

© 1983 SCIENTIFIC AMERICAN, INC

gates, two as OR gates and one path might work in the basic transistor mode.

If the modulated beams for each of the five paths were derived from a single signal that had been divided into five parts, the original signal could be subjected to five logical operations at the same time. The resulting signals could then be processed further, either separately or together. Such a mode of operation would require computer circuits quite different from those of an electronic machine.

Optical switching could lead to other modifications of computer design and logic. One possibility, as we noted above, is the introduction of devices with more than two logical states. Some of the crystals we work with have the capacity to yield successively higher levels of transmitted intensity with successive increases in the intensity of the incident beam. The levels of transmitted intensity might represent the states of a many-valued logic system.

Optical Data Processing

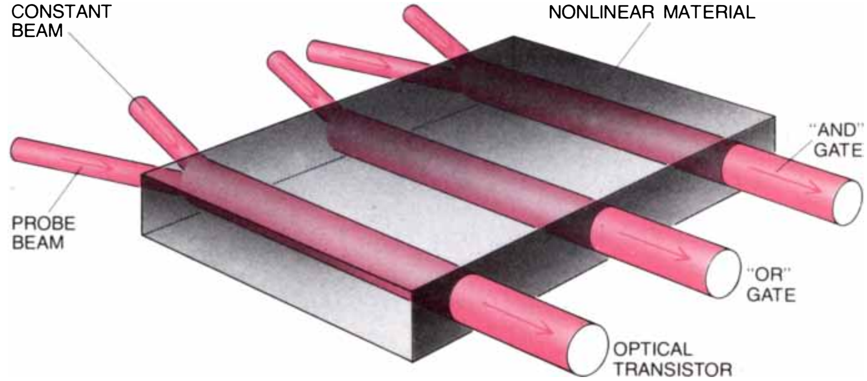
Optical switching systems need not be thought of merely as rapid substitutes for electronic devices. On the contrary, the greatest benefits of optical switches could come from applications that cannot be duplicated by other means. Optical fibers, which have the capacity to carry prodigious amounts of information, are being utilized increasingly for communications, including communications among computers. Optical switching is a natural candidate to mediate between electronic systems and optical ones. On the other hand, if the computation is done optically to begin with, optical fibers could be employed as direct links between computing systems.

As we have seen, optical switches can be made in the form of thin crystals. Current techniques of crystal growing or thin-film fabrication make it possible to fabricate very large thin sheets. If an optical image were projected directly through a large sheet in which each small area served as an optical switch, the transmission from the switches would serve as a digital record of the image. Many kinds of processing and enhancement of the image could be done while the record was being made.

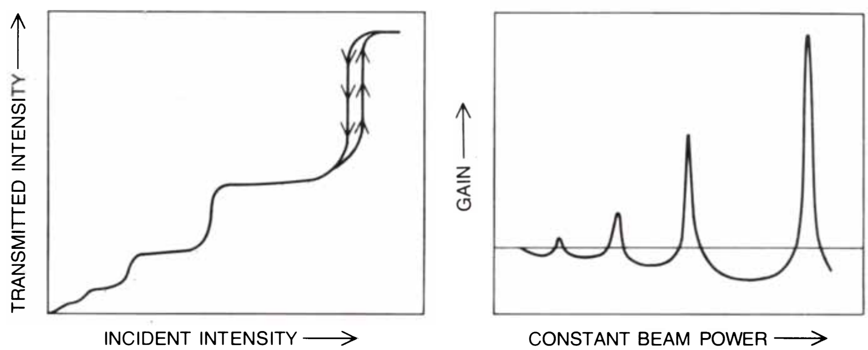
We have expressed considerable enthusiasm about the potential of optical switching devices. It might be well to consider a few of the difficulties that stand between current experimental devices and practical commercial ones. In our own investigations there has been the problem of working at low temperatures. Operation at or near room temperature would obviously make an optical computer more practical. As discussed above, the band-gap energy of indium antimonide is small, and it decreases with increasing temperature. By employing a carbon monoxide laser with a wavelength of 10 micrometers, we have recently been able to achieve optical switching at room temperature.

New lasers and new nonlinear materials will undoubtedly make it possible to improve the speed and efficiency of optical switches. The question of appropriate materials is at the center of the practical problems of building an optical computing machine. In general it is desirable that a switch consume as little power as possible. The switching power is the product of the light intensity incident on the crystal and the area of the crystal face. Reducing the size of the device can therefore reduce the switching power. The device cannot be made too small, however, because there is a limit to how finely a laser beam can be focused. In addition, as the size of the device decreases overheating becomes a problem.

Balanced against the general desirability of achieving a low switching power is the fact that within certain limits the switching time can be reduced by applying more power. The size of the device, the switching speed and the switching power must all be considered in relation to one other, to the optical and thermal properties of the nonlinear medium and to the function of the switch. The selection of the right medium or the discovery of new nonlinear materials is probably the most fundamental task in the development of optical switching. In spite of such formidable difficulties we think the optical computer is a realistic and exciting prospect.



PARALLEL PROCESSING is a capacity of optical switches that could lead to new designs and new capabilities in the computer. Multiple laser beams can be focused so that they remain separate as they pass through a crystal of nonlinear material. Each radiation path could serve as the site of a separate operation, and the operations carried out on the various beams could be different. For example, if there were three paths, one could be an AND gate, one an OR gate and one a transistor. If three incident beams came from one beam that had been split, three operations at once could be done on the original signal. Such a capacity would require a new form of information processing; in electronic computers one operation at a time is done on a signal.



MULTIPLE BISTABILITY is a property of optical switches that could lead to the development of a new system of computer logic. For a given interferometer there are many optical path lengths equal to an integral multiple of the half wavelength of the radiation. Each such optical length corresponds to a peak in the Airy function and so yields maximum transmission. Increasing the intensity of the incident beam causes the optical length of the nonlinear crystal in the cavity to reach successive peaks. The graph of transmitted intensity plotted against incident intensity thus shows a series of nearly vertical steps, each step corresponding to a different peak in the Airy curve (*left*). If the intensity of the constant beam in the optical transistor is increased, successive operations of the smaller probe beam take the transmitted power up successive vertical steps in the curve of transmitted intensity. Since the vertical steps in the curve at the left are progressively greater, the gain, or amplification, of the output by the operation of the probe beam increases with each switching operation (*right*). The different transmission levels could be made to stand for the values of a logic system that takes the place of binary logic.

© 1983 SCIENTIFIC AMERICAN, INC

Hidden Visual Processes

Vision is usually regarded as being a single sense. Experiments show, however, that the visual system includes subsystems whose operation is normally hidden from the awareness of the perceiver

by Jeremy M. Wolfe

From the standpoint of a person contemplating his abilities to perceive the world, vision seems to be a single sense. The images impinging on the two retinas give rise to a single awareness of the objects in the world: their sizes, shapes, colors, textures and positions. This view, however, is mistaken. The visual system (the brain mechanism that processes data from the eyes) is actually a set of specialized subsystems each of which acts more or less independently on some subset of visual data. Furthermore, some of the visual subsystems have an output that cannot be seen. They contribute to brain function and even to our awareness of the world, but no amount of introspection can make us aware of the subsystems themselves. They perform hidden visual processes.

How is a hidden process revealed? One way is to examine the abilities of people who have suffered brain injury. Consider the pupil of the eye, which constricts in response to an increase in the intensity of light falling on the retina. If someone suffers an injury that destroys the visual cortex (the part of the cerebral cortex first in line to get data from the eyes), he is rendered perceptually blind. That is, his awareness of a loss of vision is the same as that of someone who has lost the use of the eyes themselves. Nevertheless, the pupils continue to constrict in response to light. Even more strikingly, Ernest C. Poppel, Richard Held and Douglas Frost of the Massachusetts Institute of Technology have found that when people who are perceptually blind because of an injury to the cerebral cortex are asked to direct their eyes toward a spot of light, they do surprisingly well. The subjects report that they cannot see the spot, and so they think they are guessing, but they look in roughly the right direction more often than chance would allow.

Brain-injured people thus show evidence of multiple visual processes: some that are damaged, and hidden ones that remain functional. What such studies cannot confirm is that similar hidden processes operate in people whose visual system is intact. For that, special experimental strategies are required. Here I shall describe three sets of experiments each of which reveals a visual process that in normal people is hidden from introspection. In this regard all of us are like the brain-injured patient who can look at a spot of light although he cannot perceive it. We too are unaware of the abilities of certain parts of the visual system even though our behavior is often based on their output.

One of the functions of the visual system is to control the muscles that focus the eye on objects at various distances by changing the shape of the lens. The closer the object is, the more nearly spherical the lens must be. The process is called visual accommodation. One's impression is that one can accommodate for anything one can see. It is natural to assume, therefore, that accommodation and visual perception have access to the same set of stimuli.

Do they? By means of a number of experimental methods the accommodative status of the eye can be measured while the subject looks at a stimulus placed a certain distance from him. In one such method the subject views the stimulus through polarizing filters, and a flash of light from behind a slit briefly superposes on the stimulus a bright horizontal bar. The filters have no effect on the stimulus, but they ensure that light from the left half of the bar will enter only the top of the lens of the subject's eye and light from the right half will enter only the bottom.

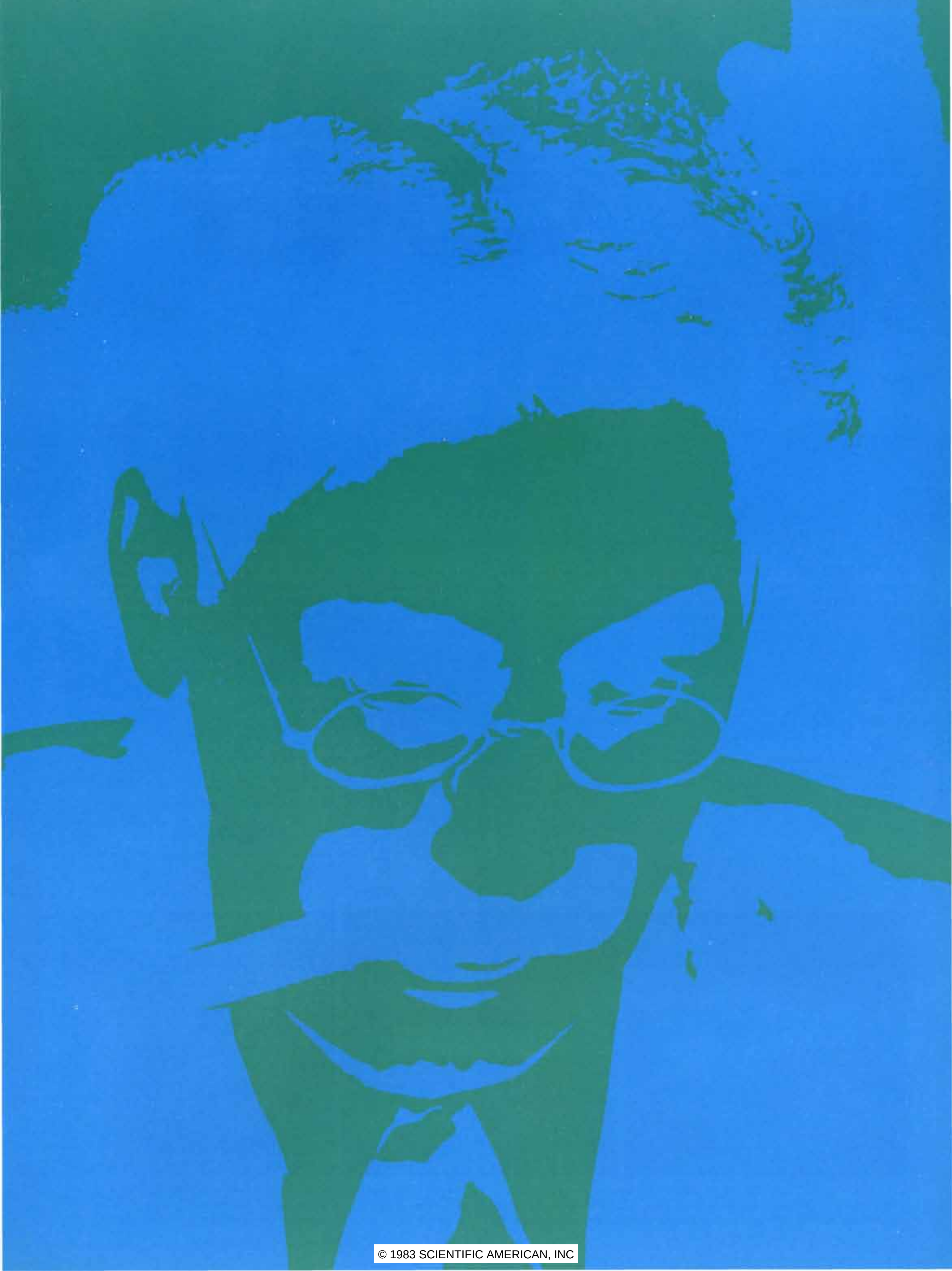
Suppose a subject who is looking at the stimulus is accommodating for a distance greater or less than the distance to the slit. If the bar is flashed at that instant, the two halves of it will be misaligned on the retina and the subject will see them as being offset. Hence the experimenter need only have the subject look at the stimulus while the bar repeatedly appears and ask the subject whether or not its halves are aligned. The slit is moved to various distances until the subject reports that they are aligned. The distance from the slit to the eye is then the distance for which the subject is accommodating.

The lens of the eye hardens with age, and so it is best to study accommodation in subjects no older than their 30's. In the laboratory such subjects can be shown stimuli that change only in distance, not in brightness and size. This eliminates all perceptual clues to the distance of the stimulus except the fact that a certain accommodative state brings the stimulus to a focus. Under these circumstances the typical subject's accommodation is about 90 percent of perfect. (If it were perfect, a stimulus at a distance of 25 centimeters would cause the eye to accommodate for a distance of 25 centimeters and so on.)

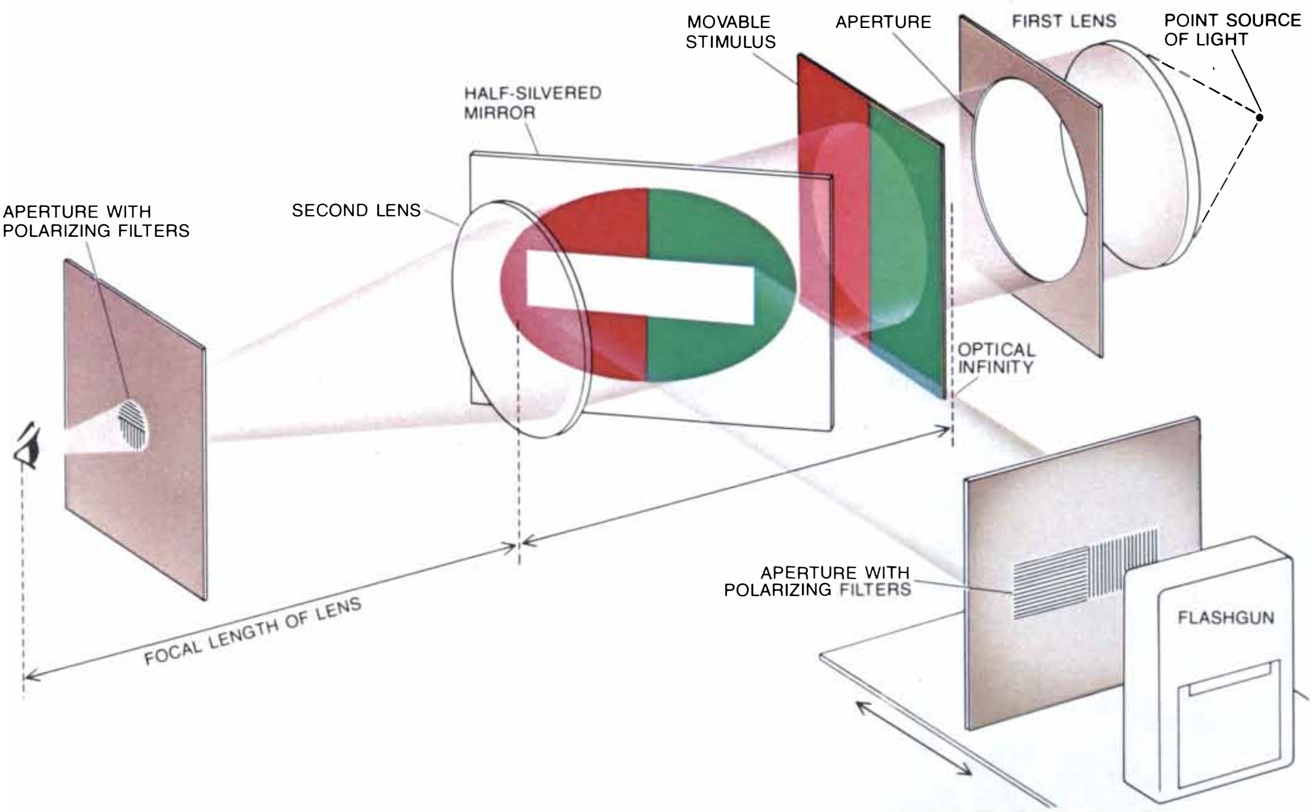
How does the eye respond when the stimulus has no features for the eye to focus on? One such stimulus would be a blank screen surrounding the subject; the experience is like being inside a giant ping-pong ball. Another stimulus would be a smaller blank screen viewed through a lens that makes the screen appear too close (in optical, or apparent, distance) for the eye to focus

ISOLUMINANT STIMULUS is an image whose edges are defined only by a change in color, not by a change in brightness. The stimulus here is imperfect: the blue parts and the green parts of the image are only as nearly equal in brightness as they can be on the printed page. Moreover, the change in brightness beyond the edge of the page is apparent, and so is the fact that the reader is holding the magazine at reading distance. When such cues are removed under laboratory conditions, subjects faced with an isoluminant stimulus prove unable to bring its edges into focus. This deficiency contributes to making a familiar face hard to recognize. The experiment indicates that the brain process underlying visual accommodation (the focusing of the eyes) cannot "see" color; it is a hidden process distinct from the processes that lead to perception. The image shows Groucho Marx as he appeared in the motion picture *Horse Feathers*.

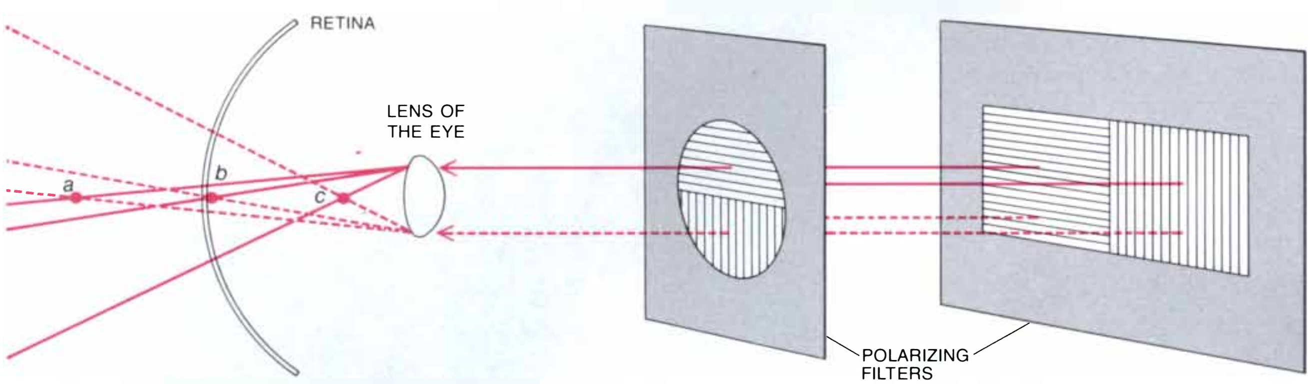
© 1983 SCIENTIFIC AMERICAN, INC

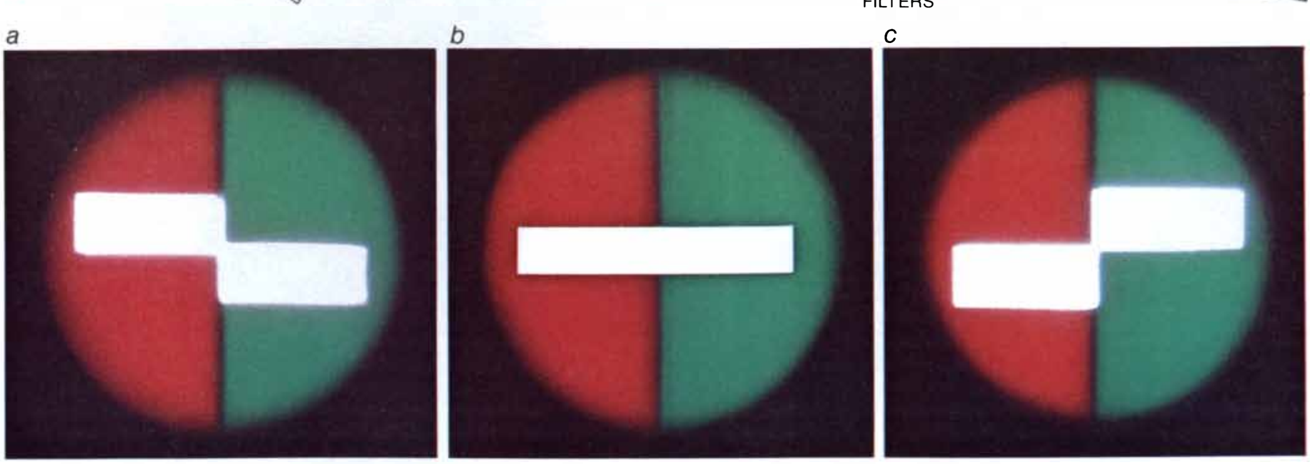


© 1983 SCIENTIFIC AMERICAN, INC



EXPERIMENTAL APPARATUS tests the ability of the eye to accommodate for an edge in an isoluminant stimulus. The lens toward the upper right in the illustration collimates a beam of light. The beam passes through an aperture, then through the stimulus, which has a vertical edge defined only by the colors red and green. A second lens directs light into the subject's eye. A further arrangement consisting of a flashgun, a movable slit and a half-silvered mirror briefly superposes on the isoluminant stimulus a horizontal bar of light.





ACCOMMODATION IS MEASURED by having people describe the flashing horizontal bar. A series of polarizing filters ensures that light from the left side of the bar enters only the top of the lens of the subject's eye; light from the right side enters only the bottom. The two sides of the bar will line up on the retina (*b*) only if the subject happens to be accommodating for the distance to the movable slit.

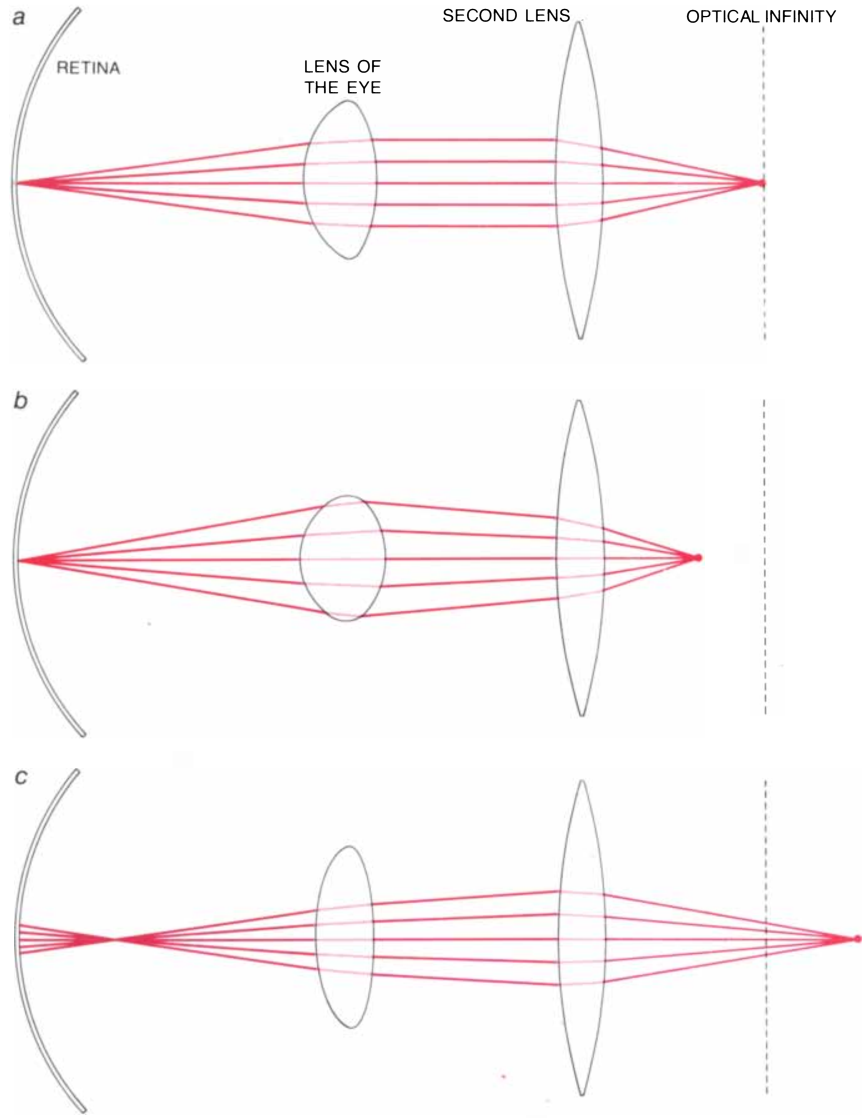
© 1983 SCIENTIFIC AMERICAN, INC

on its edges. Another would be a featureless sky; still another would be complete darkness. Herschel W. Leibowitz and D. Alfred Owens of Pennsylvania State University have found that in any of these circumstances the lens of the eye assumes a rather stable resting curvature for a particular focal distance called the dark focus. Each individual has a characteristic dark focus that is usually about one meter, or about an arm's length from the eye.

The same thing happens when the stimulus is a grating of black and white lines so fine that they give the impression of a gray field. Here again the experimenter's lens can place the stimulus at a variety of optical distances. When the stimulus is at an optical distance of 25 centimeters, the lens of the subject's eye assumes its resting state. When the stimulus is moved to an optical distance of one meter, the subject's accommodative state shows no change. The conclusion is that accommodation can shape the lens of the eye to keep an object in focus only if the system responsible for accommodation is presented with something it can "see."

Armed with the knowledge that the accommodation system is blind to certain aspects of the visual world, Owens and I asked whether the system can "see" color. We measured accommodation while our subjects looked at a simple stimulus: a circular field divided vertically in half so that it included a single vertical edge. An optical system known as Maxwellian-view optics ensured that the perimeter of the circle was beyond optical infinity. That is, the rays of light from each point on the perimeter of the circle were made to enter the eye in such a way that they could never be brought into focus whatever shape the lens of the eye assumed. As a result the vertical edge between the two halves of the field was really the only edge in the stimulus for which the eye could accommodate. The optical system was also designed to ensure that the image of the stimulus would keep the same size on the retina when we changed its optical distance.

The edge itself could be created by a difference in either color or brightness between the two half fields. In the real world most contours arise from a difference in both. We chose to make one half field red, the other green. Then we varied the brightness of one half field or the other. In this way we created stimuli ranging from an edge between red and black to an edge between black and green. At each extreme the contrast in brightness across the edge between the color and the black was 100 percent. In the precisely intermediate case, however, the red and the green were equally bright. They formed an isoluminant stimulus whose single edge was defined only by color. The contrast in brightness was zero.



OPTICAL DISTANCE of a stimulus is the apparent distance given the stimulus by the second lens in the experimental apparatus. If the stimulus is positioned at the focal length of the lens (*a*), it will be at optical infinity, that is, the light rays from each part of the stimulus will approach the eye in parallel, as if they came from a source infinitely far away. The lens of the eye will assume a curvature that brings the rays to a point on the retina. If the stimulus is closer than optical infinity (*b*), the rays will diverge as they approach the eye. The lens of the eye will then assume a more nearly spherical shape to bring the rays to a focus. If the stimulus is farther away than optical infinity (*c*), the rays will converge as they approach the eye. The lens of the eye will be unable to reduce its curvature to the required extent; thus the rays will come to a point in front of the retina. A stimulus beyond optical infinity cannot be brought into focus.

When we tested people's ability to accommodate for each of these stimuli, we found the ability declined as the contrast in brightness declined. Thus our subjects could readily tell that the isoluminant stimulus was an edge between red and green, but it was impossible for them to bring the edge into focus. In one experiment we had subjects look at a black *E* on a white background. The subjects accommodated quite well for optical distances from infinity down to 22 centimeters. Then we had them look at edges between red and black or between black and green. Their accommodation was about 80 percent as good as the best they had done for the *E*. Finally we showed them the red-green edge that has no contrast in brightness. Their accommodation was only 19 percent as good. When we repeated the experiment with pairs of colors such as red and orange or blue and green, the subjects' performance was equally poor or worse.

Like any other optical system, the eye has a chromatic aberration, which brings different colors to a focus at slightly different focal lengths. The aberration therefore tends to shift the retinal image of one color half field with respect to the image of the other. This can give rise to a bright or dark contour

© 1983 SCIENTIFIC AMERICAN, INC

Rover's family tree begins 200,000,000 years ago.

The award-winning creator of *Archosauria* and *Synapsida* now explores the natural history, variety, and adaptability of the dog, from the prehistoric ancestors he shared with man to present-day wild canids and *Canis familiaris*. "Readers are sure to emerge with a deeper understanding of just how and why this adaptive creature became man's best friend.... McLaughlin's most broadly appealing book yet."—*Kirkus Reviews*

87 illustrations

THE CANINE CLAN

A New Look at Man's Best Friend

JOHN C. McLOUGHLIN

THE VIKING PRESS Dept. ATE-SA
40 West 23rd St., New York 10010
Send ___copies of THE CANINE CLAN @ $15.75. Also send ___copies of *Archosauria* and ___copies of *Synapsida* @ $14.95 each. I enclose $_______ total. (Add sales tax where applicable plus $1.75 per copy for shipping.)

Name____________

Address____________

City________ State______ Zip______

Garden Camera CALCULATORS

Hewlett Packard

HP 67	$288.50
HP 97	558.50
HP 32E	48.50
HP 37E	57.95
HP 33C	63.50
HP 34C	74.50
HP 38C	74.50
HP 41C	148.50
HP 41CV	207.50
Accessories	Call
Printer	Call
Card Reader	Call
HP 11C	74.50
HP 12C	111.50
HP 15C	99.50
HP 16C	114.50
HP 10C	59.50

Texas Instruments

TI 58C	78.95
TI 59	168.50
PC 100C	149.50
TI MBA	49.50
Criterion Telescope	499.00

master charge VISA

Call for low prices on Nikon, Minolta, Olympus and all Major Brand Cameras

Prices subject to change without notice

Call 1 (800) 223-0595

Or Send postage and handling to

GARDEN CAMERA

135 West 29 Street, N.Y., N.Y. 10001
Or Visit Our Store: 345 Seventh Avenue
New York, Alaska & Hawaii Call:
Tel: (212) 868-1420—Open Weekdays 8:30-6:00
OPEN SUNDAYS 10-4 p.m. Closed Saturdays

Speak German like a Diplomat!

What sort of people need to learn a foreign language as quickly and effectively as possible? *Foreign service personnel*, that's who.

Now you can learn to speak German with the same materials used by the U.S. State Department — the Foreign Service Institute's *Programmed Introduction to German*.

The FSI spent thousands of dollars and many years developing these materials for use by members of the United States diplomatic corps. Today people in all walks of life who need to learn to speak a foreign language are turning to this outstanding audio program.

The FSI's German Course is by far the most effective way to learn German at your own convenience and pace. It consists of a series of cassettes, recorded by native German speakers, and accompanying textbook. You simply follow the spoken and written instructions, listening and learning. By the end of the course you'll find yourself learning and speaking entirely in German!

This course turns your cassette player into a "teaching machine." With its unique "programmed" learning method, you set your own pace—testing yourself, correcting errors, reinforcing accurate responses.

This Programmed Course comes in two volumes, each shipped in a handsome library binder. You may order one or both:

☐ **Volume I. Programmed Introduction to German,** 9 cassettes (12 hr.), and 647-p. text, $115.

☐ **Volume II. Basic Course Continued,** 8 cassettes (8½ hr.), and 333-p. text, $98.

(Conn. and N.Y. residents add sales tax.)

TO ORDER BY PHONE, PLEASE CALL TOLL-FREE NUMBER: **1-800-243-1234.**

To order by mail, clip this ad and send with your name and address, and a check or money order — or charge to your credit card (AmEx, VISA, MasterCard, Diners) by enclosing card number, expiration date, and your signature.

The Foreign Service Institute's German course is unconditionally guaranteed. Try it for three weeks. If you're not convinced it's the fastest, easiest, most painless way to learn German, return it and we'll refund every penny you paid. Order Today!

81 courses in 26 other languages also available. Write us for free catalog. Our 10th year.

Audio-Forum
Suite 23-D
On-the-Green,
Guilford, CT. 06437
(203) 453-9794

AUDIO-FORUM®

Or visit our New York sales office: 145 E. 49th St., New York, N.Y. 10017 (212) 753-1783

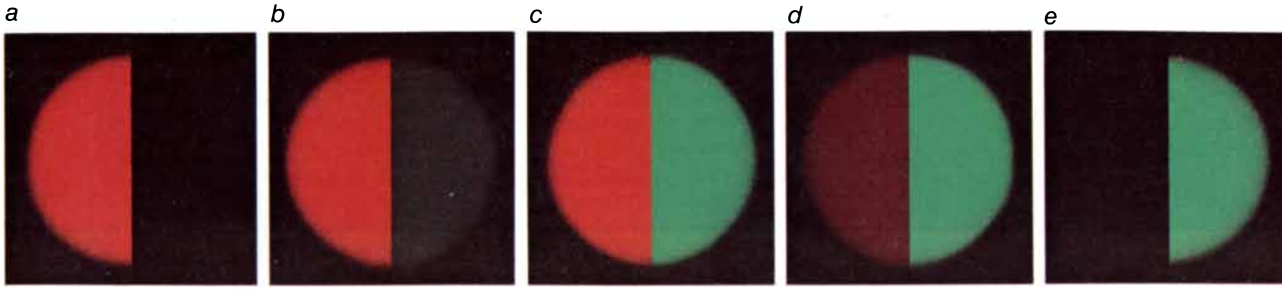
depending on whether the images separate or overlap. As a result it is surprisingly difficult to create a stimulus that has absolutely no clues to the presence of an edge except the difference in color itself. We suspect our subjects' performance would have been even worse than it was if chromatic aberration could have been eliminated. Still, the conclusion seems clear: accommodation is colorblind, or at least it "sees" color very badly. It is a visual subsystem remarkably independent of visual perception. Here, then, is a finding that no amount of introspection could have suggested about vision.

A second strategy for exploring the division of the visual system into subsystems is to consider the ways a particular visual function is performed by different parts of the system. One such function is binocular vision. A normal person has two eyes, and the brain makes every effort to integrate the data from them. It is becoming clear, however, that the brain does more than simply combine the inputs from the eyes in one grand binocular process that leads to visual perception. Instead several special-purpose mechanisms combine the inputs in their own way to meet their own particular needs. Just as no amount of introspection will reveal that accommodation is colorblind, so no amount of introspection will reveal these multiple binocular processes. Nevertheless, the processes do exist and can be revealed by experimentation.

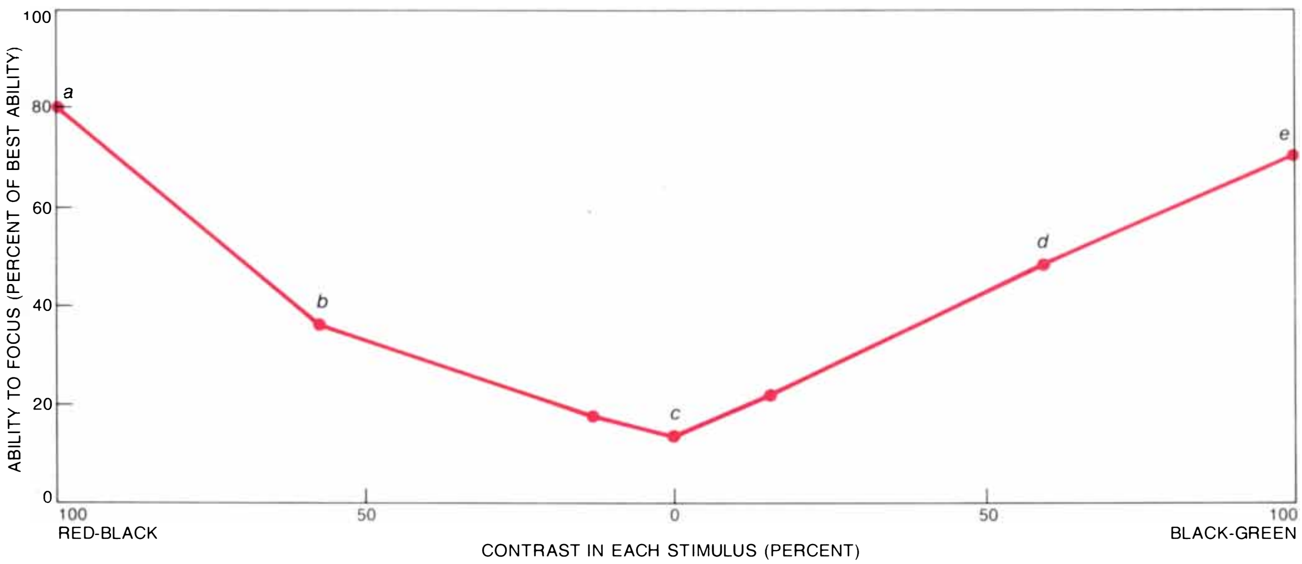
Consider a visual illusion that everyone has experienced in one form or another. You are sitting in a train, waiting for it to pull out of the station. Looking out the window, you see another train motionless on the adjacent track. The other train's image starts to slide backward, and you distinctly feel that your train is moving forward. Then you see it is the other train that has moved; your train is still in the station. The motion of the other train somehow deluded you into believing you were moving. You have experienced an illusion of motion created solely by visual stimulation. The illusion is known as vection.

Several investigators have sought to examine the brain mechanism responsible for vection. Held and I became interested in a rather different question. We wanted to know what the vection process can "see." In particular we wanted to know if the process can make use of people's ability to look at the world through two eyes. One might think the answer could be found quite easily by comparing the vection experienced when both eyes are open with the vection experienced when only one eye is open. Normally, however, the sensation of vection is already at its maximum when only one eye is open; the illusion of self-motion cannot be more pronounced. One

© 1983 SCIENTIFIC AMERICAN, INC



STIMULI FOR THE EXPERIMENT on the ability to accommodate vary from an edge between red and black (*a*) to an edge between black and green (*e*). The precisely intermediate stimulus (*c*) is isoluminant. The circular perimeter of each stimulus is defined by an aperture placed well beyond optical infinity in the experimental apparatus; hence the subject's eye cannot bring the perimeter into focus.



ABILITY TO FOCUS on the stimuli turns out to depend on their contrast in brightness. Here the average performance of four subjects is charted as a percentage of the ability they had when they looked at black letters on an eye-examination chart. The subjects' performance grows poorer with decreasing contrast. It is worst (19 percent of the best performance) when the stimulus is isoluminant.

needs, therefore, to create a stimulus whose binocular effect could conceivably be greater than its monocular effect. Such a stimulus must have a purely binocular component: there must be something about the stimulus that is invisible to each eye by itself.

A component that meets this requirement is called a cyclopean stimulus, after the Cyclops, the Homeric one-eyed creature encountered by Odysseus. The name is apt because a cyclopean stimulus is evident only if the brain combines the input from the eyes. In effect the brain must act as a single, cyclopean eye. One example of a cyclopean stimulus is the small difference in position between the image of an object in the world on each of the retinas. The difference can easily be seen. Stretch out one of your arms and look at the tip of a finger first with one eye closed and then with the other eye closed. You will notice that the two views are slightly different. The brain exploits such differences in the processes that lead to the perception of three-dimensional depth. Clearly the brain must draw on data from both eyes for the stimulus to exist.

Our cyclopean stimulus for vection capitalized on the well-known fact that objects seen in a motion picture seem to move smoothly even though they are presented in a succession of still photographs. Our subjects sat inside a cylinder three feet in diameter and five feet high. The inside surface of the cylinder was white and was covered with a random pattern of black dots an inch in diameter. The cylinder rotated about the subject at the rate of 30 degrees of angle per second. When the inner surface of the cylinder was illuminated by ordinary lighting, the subject reported a sensation of rotation in the opposite direction. For our experiments we illuminated the surface with the periodic flashes of a stroboscopic lamp. Each flash produced the equivalent of a frame from a movie of black dots. The subject reported the same sensation of vection.

To create a cyclopean stimulus two strobe lamps were needed. One was covered with a red filter, the other with a green filter. The subjects wore goggles that placed a red filter in front of one eye and a green filter in front of the other. No red light could pass through the green filter; no green light could pass through the red. Hence the light from one strobe lamp was seen by one eye and the light from the other strobe lamp was seen by the other eye.

Suppose each strobe lamp is flashing at 10 hertz, or 10 times per second. If the two lamps are flashing in phase (that is, in synchrony), the subject should notice no important difference in looking with two eyes rather than one. In either case he will see in effect a 10-frame-per-second movie of moving dots. Suppose, however, the strobe lamps are exactly out of phase, so that they flash in alternation. Now there is a difference. A subject who has only one eye open will see again a 10-frame-per-second movie. With both eyes open he will see an additional movie. It will appear at a rate of 20 frames per second, and it will consist of frames presented in alternation first to one eye, then to the other, then to the first eye again.

The 20-hertz interocular movie is thus

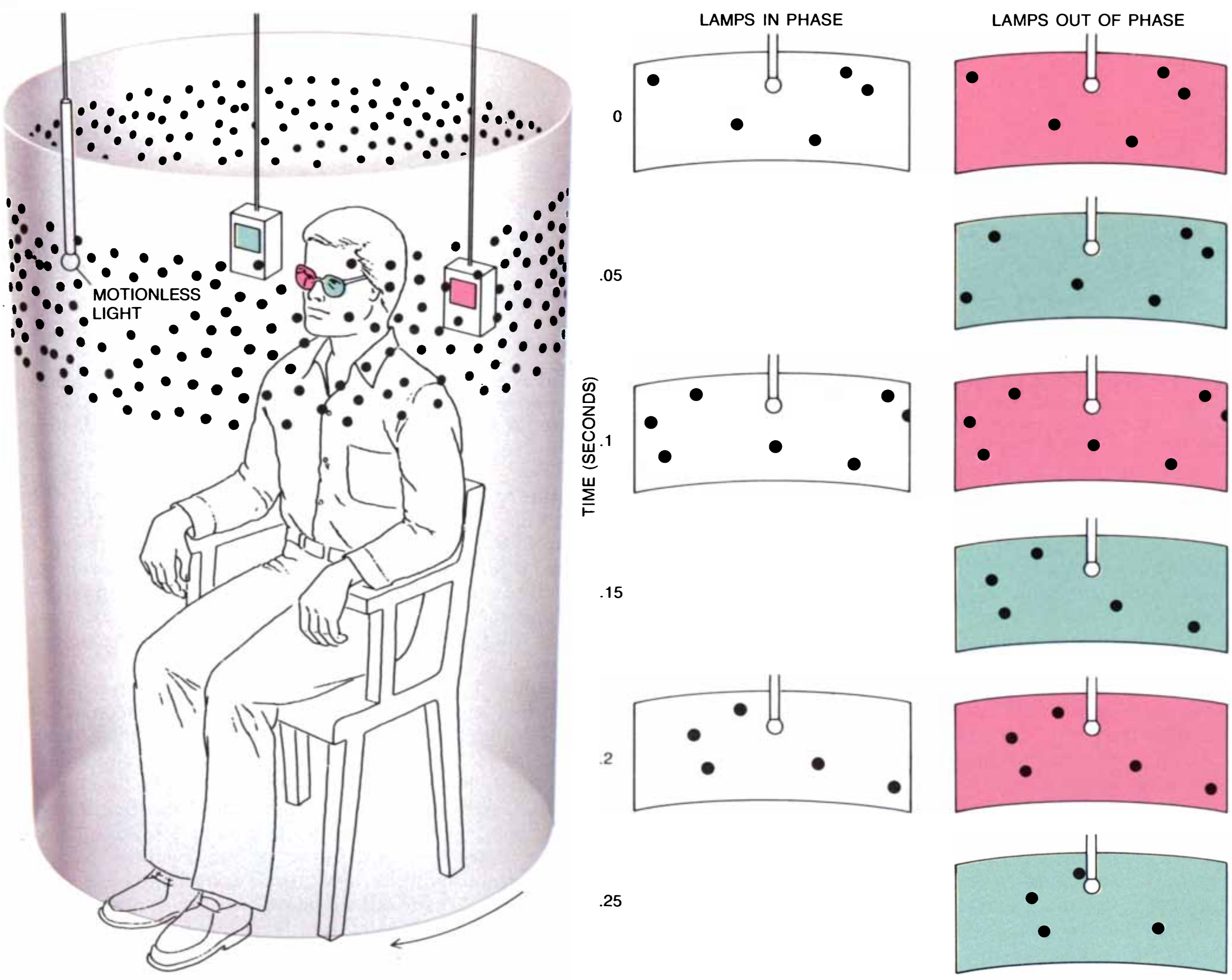
© 1983 SCIENTIFIC AMERICAN, INC

a cyclopean stimulus: it cannot be seen by either eye alone. It does give rise to the appearance of motion. The question is whether the vection system can "see" it. If it can, the experiment where the strobe lamps are out of phase could produce a greater sensation of vection than the experiment where the lamps are in phase. If it cannot, the two experiments should have the same result.

With our apparatus in place we prepared subjects for the experiments. In particular we taught them a method known in experimental psychology as magnitude estimation. We asked them to give a rating of 10 to a compelling sensation of self-motion and one of zero to no sensation of motion. Ratings between 0 and 10 were to be given to sensations between these two extremes. Unlikely as it may seem, some 25 years of research in experimental psychology, notably the work of S. Smith Stevens of Harvard University, has shown that people are quite good at assigning numbers to their sensations. They can do it repeatedly, systematically and reliably.

One thing did concern us: there could be no difference between the results of the two experiments if the sensation of vection yielded the maximum rating of 10 when the lamps were in phase. If the frames of a movie are shown at a progressively lower rate, the illusion of motion becomes progressively less compelling. In the same way the flicker rate turns out to be important to the sensation of self-motion. We found that any rate higher than about 2.5 hertz produced some degree of vection, but only rates higher than about 15 hertz consistently earned ratings of 10 from our subjects. Therefore flicker rates between 2.5 hertz and 15 hertz would suit our purpose. The cyclopean stimulus created (at twice the flicker rate) by flashing the lamps out of phase would be capable in principle of producing a more compelling sensation of vection than the stimulus produced by leaving the lamps in phase.

Our findings were straightforward. For any flicker rate between 2.5 hertz and 15 hertz the two strobes out of phase always elicited the greater magnitude estimates. Evidently the vection system can "see" the purely binocular stimulus. Does this result establish that the brain has multiple binocular processes? Not in itself. As I have noted, the slight difference in position between the images of an object on each retina is



EXPERIMENT WITH VECTION (the sensation of bodily motion arising from the motion of the visual world) requires that subjects be seated inside a rotating cylinder whose inner surface is white and is covered with a random pattern of black dots. Two stroboscopic lamps illuminate the surface in flashes of red and green light; a pair of goggles allows the light of each color to enter one eye only. If the lamps flash in phase (that is, synchronously) at a rate of 10 flashes per second, a subject with one eye open sees what a subject with both eyes open sees: in effect a 10-frame-per-second motion picture of moving dots. If the lamps flash out of phase, a subject with one eye open sees the identical movie, but a subject with both eyes open sees an additional movie. It appears at a rate of 20 frames per second, and it is interocular: its successive frames are seen by the two eyes in alternation. A small stationary light serves as a reference point.

© 1983 SCIENTIFIC AMERICAN, INC

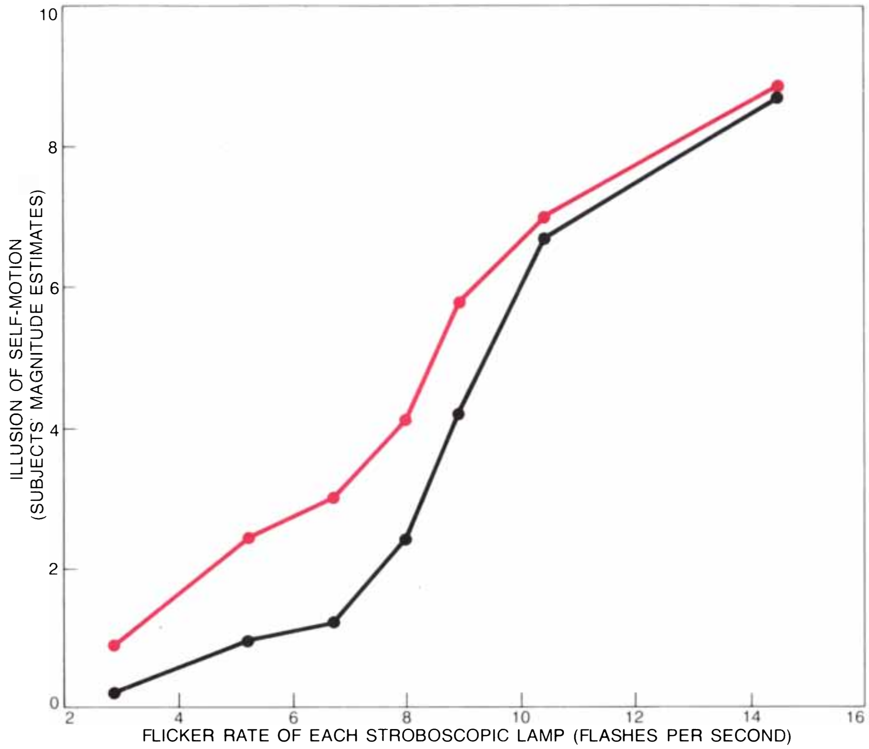
a cyclopean clue to three-dimensional depth. It is conceivable that the binocular mechanism serving this aspect of visual perception also serves the binocular contribution to vection.

Some people, however, cannot perceive depth on the basis of cyclopean stimuli. For example, they cannot get the illusion of depth when they look at an image through a stereoscope or when they go to a "3-D" movie. They are stereoblind. Some of them are born that way, just as some people are born colorblind. Others lose the ability because of defects that develop in infancy in the ability to align the eyes. Stereoblindness is somewhat rarer than colorblindness; it seems to affect only a few percent of the U.S. population. A person who is stereoblind can still perceive depth on the basis of visual cues such as the apparent size of familiar objects or the fact that some objects in the visual field block others that are farther away. These monocular cues to three-dimensional depth are quite good; many people who are stereoblind do not know they are until a test reveals it.

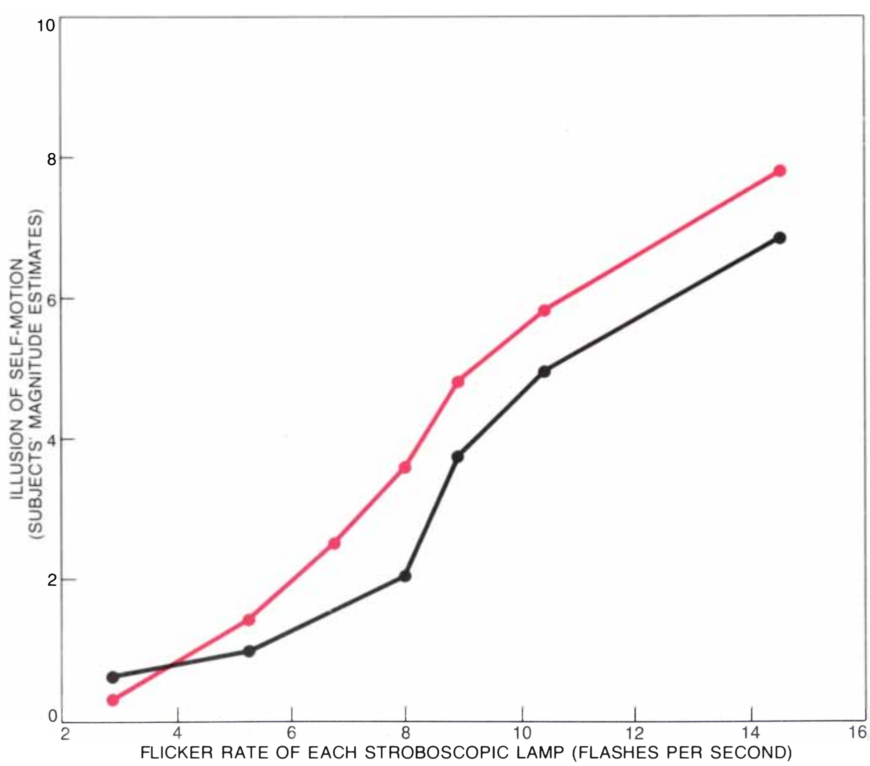
Held and I asked four people who had been established as stereoblind to be subjects in our vection experiments. Their magnitude estimations were much like those of our normal subjects. They too got an increased sensation of self-motion from the purely binocular stimulus. Here, then, are people with a defective binocular mechanism who nonetheless prove to be perfectly normal in experiments requiring that their brain employ a binocular mechanism. Thus the defective mechanism cannot be the only binocular mechanism in the brain. There must be at least two binocular mechanisms, one mediating stereoscopic depth perception, the other involved in the production of vection. The experiments again show that visual data feed more than one processing system.

The two hidden visual processes I have now described (the colorblind visual process that controls accommodation and the binocular visual process that contributes to vection) appear to play no direct role in visual perception. Other hidden visual processes do take part in perception.

For an example I shall return to the binocular visual processes. Our experiments with vection revealed two binocular processes. The experiments I shall now describe show more. Indeed, there emerges a remarkable assortment of processes. Between our two eyes and our single perception of the visual world lie processes that can "see" out of one eye, processes that can "see" out of either eye and, most surprising, a purely binocular process: a process that can "see" only out of both eyes. In fact, it can "see" only when both eyes are looking at the same stimulus. If your visual

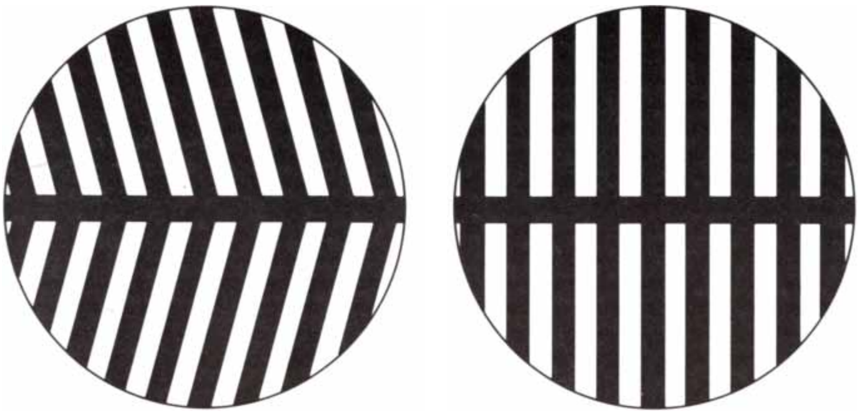


SENSATION OF VECTION is greater if the lamps are out of phase (*color*) than it is if the lamps are in phase (*black*), showing that the brain process responsible for vection can "see" a movie created by stimulating the eyes in alternation. It is thus a binocular process. The data in the chart were collected by asking six subjects to assign ratings in which 10 signified the most compelling illusion of bodily motion whereas zero signified the absence of such an illusion.



RATINGS BY STEREOBLIND SUBJECTS show that they too have a greater sensation of vection if the lamps are out of phase. People who are stereoblind lack the binocular visual process that compares the images from the eyes to aid in the perception of three-dimensional depth. Hence the process responsible for vection must be a different binocular process. The data in the chart were collected from four stereoblind subjects, who rated their sensations.

© 1983 SCIENTIFIC AMERICAN, INC



TILT AFTEREFFECT is an illusion that serves the study of hidden visual processes on the pathways leading to visual perception. To experience the illusion let your eyes scan back and forth along the horizontal black bar crossing the chevron at the left side of the illustration. You should then find that the two halves of the pattern at the right are briefly not colinear. The tilt aftereffect is measured by making the pattern at the right adjustable and asking subjects to adjust it so that it seems to be colinear immediately after they have stared at the chevron.

system had only the purely binocular process, you would be unable to see if one eye was closed. You would be unable to see when you tried, for example, to peek out from behind a tree and one eye surveyed only tree bark while the other surveyed a house.

The ability to perceive a purely binocular stimulus such as the cyclopean stimulus that served our study of vection does not ensure that the brain has a purely binocular visual process. The cyclopean stimulus could have been "seen" by a visual process that accepts input from both eyes and also from either eye alone. How, then, is it known that a purely binocular process is there? Again one resorts to an indirect method, in this case based on a temporary distortion of vision called the tilt aftereffect.

Look at the pattern of stripes at the right in the illustration above. The top half of the pattern should appear to be aligned with the bottom half. Now look at the chevron pattern at the left in the illustration. Let your eyes scan back and forth for one or two minutes along the horizontal black bar that crosses the center of the chevron, then quickly shift your gaze to the pattern at the right. You should find that the two halves no longer seem aligned. They should appear to be bent in the direction opposite to the direction of the bend in the chevron.

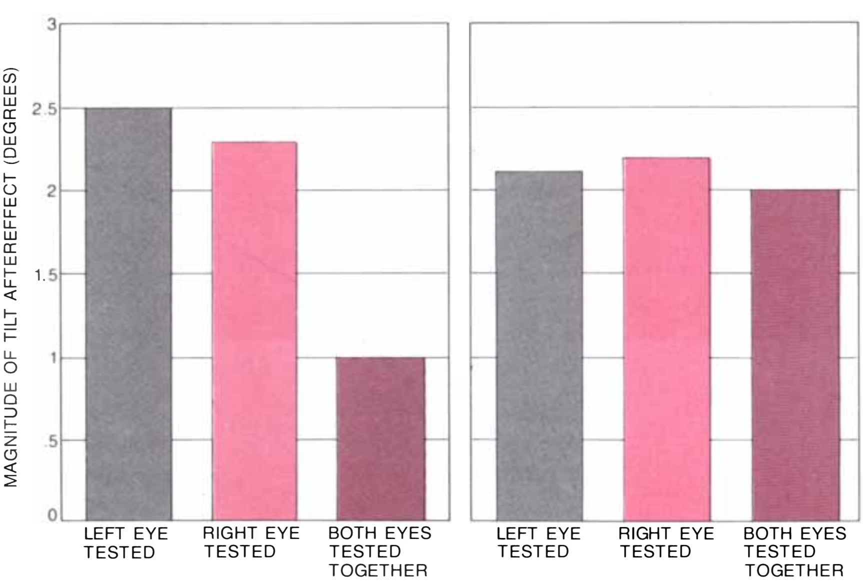
This is the tilt aftereffect. It can be measured by making the two halves of the pattern at the right adjustable and asking people to make the two halves look aligned. Before people view the chevron they make settings quite close to colinearity. After viewing the chevron their settings are systematically displaced by about two degrees.

Suppose a subject looks at the chevron with only his right eye open and then manipulates the adjustable pattern with only his left eye open. The left eye never sees the chevron, yet an aftereffect is detected. In short, there is interocular transfer. It is evidence for a binocular process but not a purely binocular one. After all, the process was activated when the right eye was exposed to the chevron and it was activated when the left eye was tested with the adjustable pattern. Apparently it can respond to either the left eye or the right.

As it happens, the aftereffect is smaller when it is tested with the eye that was not exposed to the chevron than it is when it is tested with the eye that was exposed. Only 70 to 80 percent of the aftereffect transfers. Randolph Blake and his colleagues at Northwestern University conclude from this finding that at least two processes are involved when the tilt aftereffect is tested with one eye. One of them is the binocular process; the other one is monocular. Thus the exposure of the right eye to the chevron activates both the binocular process and the right eye's monocular process. The testing of the right eye activates both of them again. The result is a strong tilt aftereffect. Suppose the left eye is tested. The binocular process, which was exposed to the stimulus, becomes active again. The left eye's monocular process also becomes active, but it never "saw" the chevron. Its output dilutes the magnitude of the aftereffect.

By means, then, of the tilt aftereffect Blake demonstrated the existence of both monocular and binocular processes. Held and I exploited the aftereffect to show that a purely binocular process exists as well. We had subjects look at the chevron with each eye in alternation: one minute with the left eye, one minute with the right eye, one minute with the left eye again and one minute with the right eye again. When we measured the aftereffect, we had them keep either the left eye open, the right eye open or both eyes open.

We reasoned that the alternating monocular viewing of the stimulus would expose every visual process that gets input from the left eye, every proc-



PURELY BINOCULAR PROCESS is revealed by comparing the results of two experiments that employ the tilt aftereffect. In one of the experiments (*left*) subjects stared at the chevron with each eye open in alternation; thus the stimulus was available to every visual process that gets input from one eye and every process that gets input from either eye. On the other hand, the stimulus was unavailable to a purely binocular process: one that gets input only from both eyes. The subjects then manipulated the adjustable pattern with one eye or with both eyes open. The binocular aftereffect proved to be less than the monocular aftereffect. The second experiment (*right*) differed only in that the subjects stared at the chevron with both eyes open. The binocular aftereffect now proved to equal the monocular aftereffect. Evidently the binocular viewing of the stimulus in the second experiment exposed the purely binocular process.

© 1983 SCIENTIFIC AMERICAN, INC

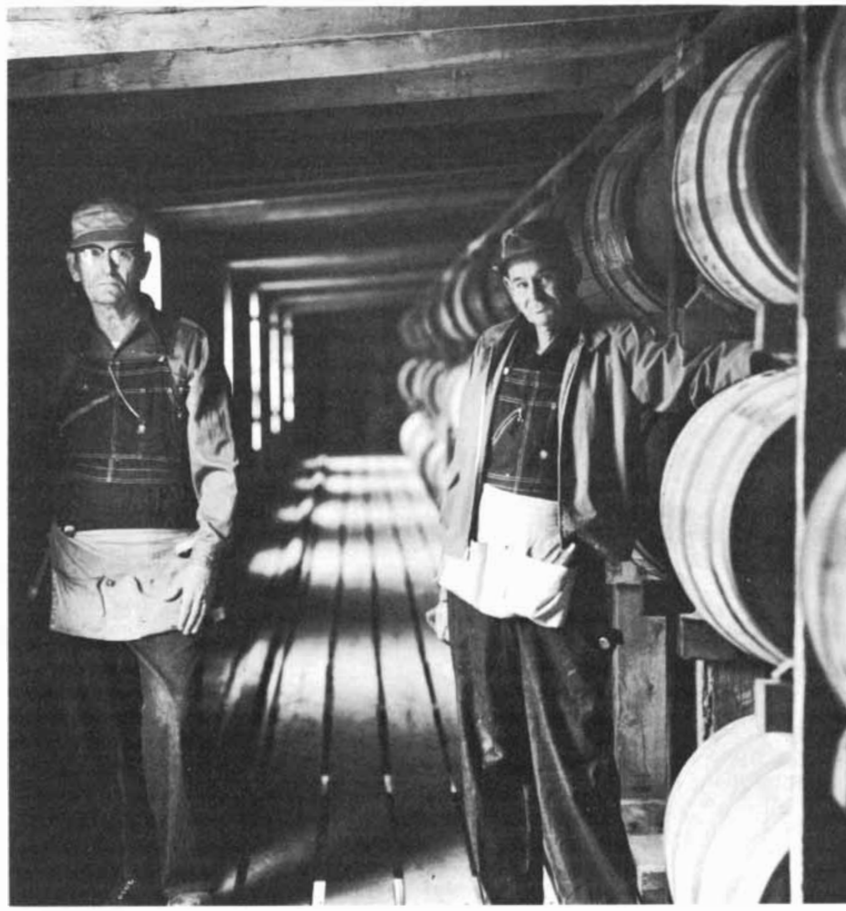
ess that gets input from the right eye and every process that gets input from either eye. If these were the only visual processes, they would constitute the entire visual system. Hence it would make no difference whether the subject had one eye open or both eyes open when he manipulated the adjustable pattern; the tilt aftereffect would be the same. If, however, there were a visual process that is active only when the left eye and the right eye are stimulated simultaneously, the results would be different. The alternating monocular viewing of the stimulus would leave this process unexposed, but then if both eyes were opened for the test of the aftereffect, the process would be activated, and it would dilute the aftereffect.

That is in fact what we found. The magnitude of the aftereffect was much less with both eyes open than it was with one eye closed. In a further experiment we had subjects look at the chevron with both eyes open. We expected this would expose every visual process, and we were right: the aftereffect was now much the same whether the subjects had both eyes open or one eye open when they manipulated the adjustable pattern.

In both experiments the left eye and the right eye were both exposed to the chevron. The only difference is that in the first experiment our subjects never saw the chevron with both eyes open at the same time. In the second experiment they did. A process was left unexposed by the first experiment and then was exposed in the second. That process must "see" only with both eyes. It is a purely binocular process.

What could a purely binocular process do in assistance of vision? In order to perceive three-dimensional depth the visual system seeks matches and slight mismatches between features in the images in each eye. The purely binocular process could serve in such a search. Visual perception cannot, however, be based on the output of a purely binocular process alone. If it were, all the unmatched parts of the two retinal images would vanish. In fact what we cannot see by means of a purely binocular process we can see by means of a monocular process or a process that responds to either eye.

From ancient times people have spoken of five senses. It is becoming ever clearer, however, that five is much too small a number. Senses such as touch seem to be divided into a variety of submodalities; the visual system is also divisible. Perhaps vision is best regarded not as a single sense but as a set of systems, each one a sense in its own right. It is likely that the full range of human visual senses remains to be discovered. The senses are like an old and crowded attic in that one finds the unexpected in each new corner one explores.


If you'd like to know more about Jack Daniel's Whiskey, drop us a line.

THESE MEN can tell exactly what's happening inside every barrel in a Jack Daniel's warehouse.

In the heat of summer the whiskey is expanding into the charred inner wood of the barrel. Come Halloween, it's starting to cool. And inching its way back toward the center. Over the aging period, this gentle circulation of whiskey is going on constantly. Of course, it can't be perceived by the human eye. But after a sip of Jack Daniel's, we believe you'll recognize its importance.

CHARCOAL MELLOWED DROP BY DROP

Tennessee Whiskey • 90 Proof • Distilled and Bottled by Jack Daniel Distillery
Lem Motlow, Prop., Inc., Lynchburg (Pop. 361), Tennessee 37352
Placed in the National Register of Historic Places by the United States Government.

© 1983 SCIENTIFIC AMERICAN, INC

The Active Solar Corona

Once visible only during eclipses, the pearly surround of the sun is under constant surveillance. The observations show dynamic activity reflecting the interaction of matter and magnetic fields

by Richard Wolfson

One of nature's rarest and most dramatic spectacles is the solar corona: the thin outer atmosphere of the sun that becomes visible during a total eclipse. Although a total eclipse of the sun occurs somewhere on the earth every year or so, most people have the chance to witness only one or two in a lifetime. Light from the solar corona is normally overwhelmed by light from the sun's luminous surface: the photosphere. That the corona is visible at all is a result of the remarkable coincidence that the moon's apparent size is almost exactly the same as the sun's. If the moon were any smaller, eclipses would never be total. If the moon were any larger, the intricate inner corona would be hidden from view.

In the past decade what is known about the corona and indeed about the sun itself has been vastly enlarged by instruments both on the ground and on spacecraft, unmanned and manned. The corona is observed to be in a continuous state of flux, punctuated by periodic explosive outbursts. A "wind" of coronal particles streams outward through the solar system; fluctuations in the wind reflect coronal activity. This view of the corona and its manifestations transforms the sun from a compact object at the center of the solar system to a single entity of giant dimensions, reaching beyond the most distant planets. The corona is a huge and yet nearby laboratory for studying astrophysical processes. Understanding the corona can help to advance knowledge of such diverse phenomena as fusion reactors here on the earth and supernova explosions in distant stars.

In ordinary photographs the corona appears as a pearly halo surrounding the eclipsed sun. Most of the light from the halo does not originate in the corona but is sunlight scattered from coronal electrons. For every million photons of light leaving the sun's photosphere only one is scattered in the corona. As a result the corona is a million times dimmer than the photosphere. The corona's transparency indicates that the coronal gas is exceedingly rarefied. Indeed, by laboratory standards on the earth the corona is an excellent vacuum; the maximum density of the coronal gas is between 10^8 and 10^9 particles per cubic centimeter.

Spectroscopic studies suggest that the corona is like the visible sun in composition, being dominated by hydrogen and helium, with only traces of the heavier elements. Such studies also offer evidence of the corona's temperature. When gases are heated, their atoms become ionized, or shorn of one or more electrons. The coronal gas turns out to be much more highly ionized than the gas in the photosphere, whose temperature is some 6,000 degrees Kelvin (degrees Celsius above absolute zero). In the corona hydrogen and helium are essentially stripped of electrons, and the heavier elements have lost many of their electrons. The high state of ionization signifies that the corona is extremely hot. Moreover, the lines of the coronal spectrum are broadened, indicating that coronal particles are in random motion at high speed. On the basis of these spectroscopic observations the temperature of the corona is estimated to lie between one and two million degrees K.

Accounting for the high temperature of the corona long presented theorists with a major puzzle. The second law of thermodynamics states that heat cannot flow spontaneously from a cooler object to a hotter one. For this reason the thermal energy of the sun, enormous though it is, cannot heat the corona beyond 6,000 degrees K. So fundamental is this objection that for decades solar physicists were reluctant to accept the spectroscopic evidence at face value. Various ad hoc hypotheses were put forward, suggesting, for example, that the ions in the corona might originate somehow deep within the sun or that they might be created directly in the corona by nuclear fission. Finally in the 1940's the concept of a truly hot corona became widely accepted. Once theorists were convinced of the reality of high coronal temperatures they soon made rapid progress in accounting for the corona's gross physical structure.

If the second law of thermodynamics rules out the transfer of heat energy from the photosphere to the corona, what is the heating mechanism? The most likely candidate is the direct transfer of mechanical energy. According to present theories, mechanical energy that originates in the convective layer of the sun below the photosphere is carried outward to the corona in the form of either intense sound waves or magnetohydrodynamic waves.

Sound waves are simply mechanical disturbances of a gas that manifest themselves as pressure variations. Magnetohydrodynamic waves, which occur in ionized gases, are disturbances arising from the interaction of ionized matter and magnetic fields. Although the general concept of mechanical heating by waves is well established, the elaboration of detailed models consistent with observation is engaging the efforts of many investigators. The work promises to link the physical properties of the corona with those of the underlying layers, and it has the goal of achieving a unified understanding of the entire sun.

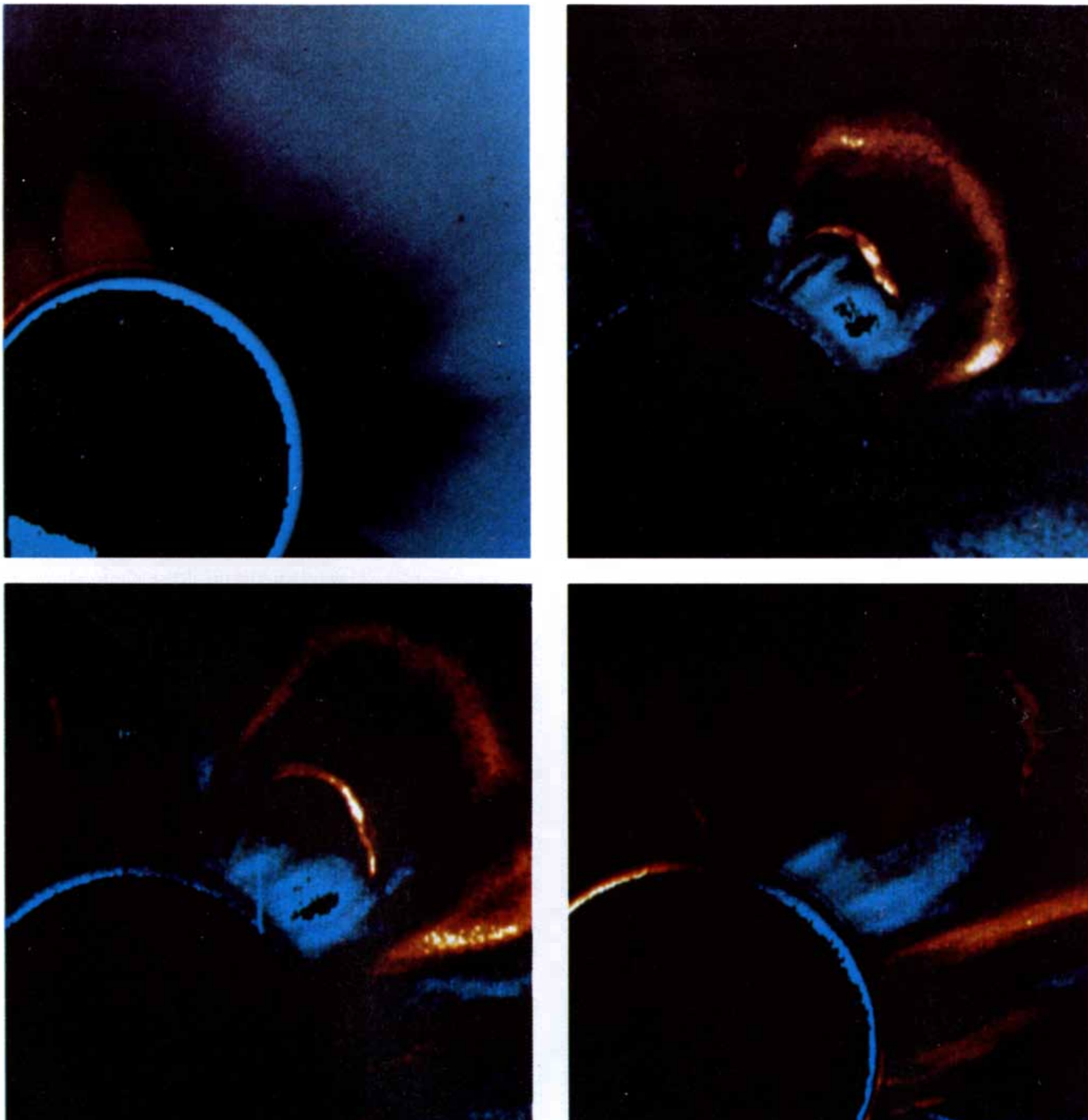
The corona, then, is a region of high temperature and low density that surrounds the cooler, denser and more familiar photosphere. (A thin intermediate layer, the chromosphere, separates the photosphere from the corona.) Most ordinary photographs show that the corona is far from symmetrical, often appearing dimmer and less extended near the solar poles. Away from the poles prominent bulges suggest a complexity of structure. Such deviations from symmetry vary from eclipse to eclipse. The most symmetrical coronas coincide with times of maximum sunspot activity in the 11-year solar cycle.

Coronal structure hinted at in ordinary photographs is largely washed out by overexposure of the bright inner corona. The most successful method for overcoming this limitation is to photograph the corona through a special fil-

© 1983 SCIENTIFIC AMERICAN, INC

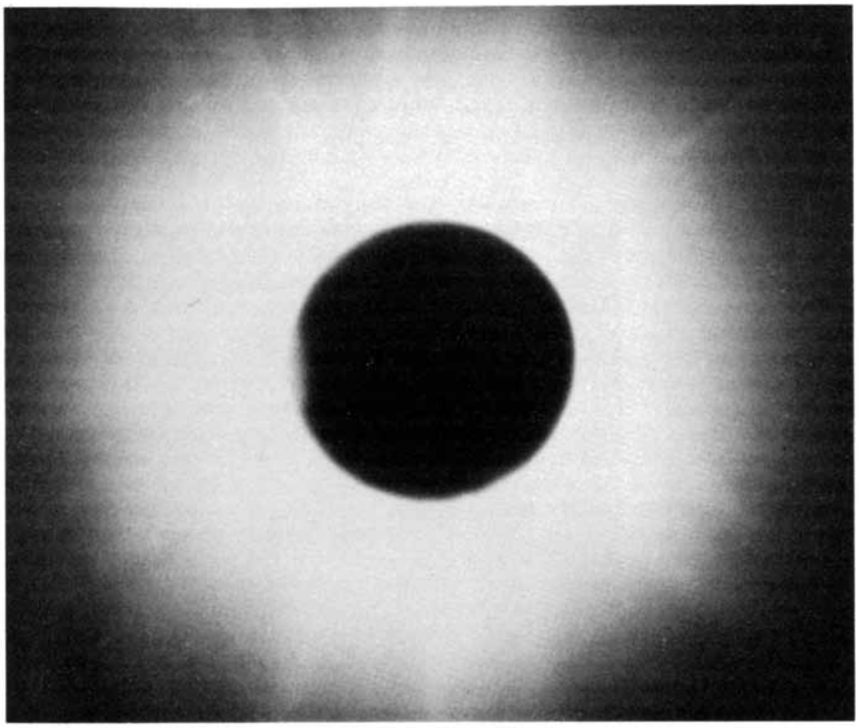
ter whose density is radially graded to block more light at the center of the image than at the edge, thereby compensating for the falloff in coronal brightness. With such filters coronal structures can be detected at distances of many solar radii from the photosphere.

In a photograph taken through a radially graded filter the delicate beauty of the solar corona is revealed in intricate detail. Long, graceful streamers extend outward for several million kilometers. Closer to the solar disk are bright loops of coronal material. In some pictures straight, narrow rays can also be seen. At times of maximum solar activity such coronal features surround the entire solar disk. At times of minimum activity, in contrast, coronal features are confined to the lower solar latitudes. At such times the corona shows large

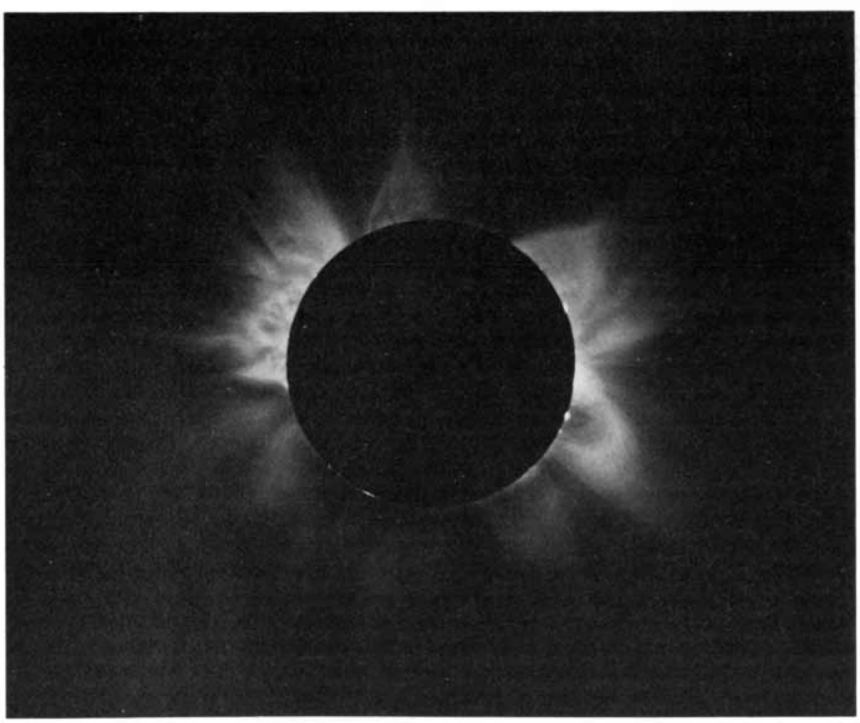


TRANSIENT LOOP IN THE SOLAR CORONA was recorded in April, 1980, by the coronagraph aboard the Solar Maximum Mission (SMM) satellite. A coronagraph incorporates a disk that blocks the brilliant light from the sun's photosphere, or visible surface, so that the million-times-dimmer corona can be observed. The coronagraph carried by the SMM spacecraft was developed by the High Altitude Observatory of the National Center for Atmospheric Research in Boulder, Colo. The light of the corona is focused on a television camera tube, and the image is transmitted back to the earth in digital form. In this four-part sequence, displayed in arbitrary colors, the corona before the transient event appears at the upper left. The other three pictures are made by a digital subtraction technique in which the light value at each point represents the change in light intensity from the level recorded in the first image. At the upper right the coronal loop has attained nearly the diameter of the sun. At the lower left, 25 minutes later, the loop, expanding at the rate of about 300 kilometers per second, has reached the edge of the frame, and behind it a solar prominence has begun to move upward through the corona. At the lower right, 20 minutes later, only one leg of the loop is visible. It is not known whether the eruptions of the coronal loop and the solar prominence are causally related or whether both of them are triggered by some underlying instability in the magnetic field of the sun.

© 1983 SCIENTIFIC AMERICAN, INC



IN A CONVENTIONAL PHOTOGRAPH OF A SOLAR ECLIPSE the inner corona is overexposed and all delicate detail is washed out. The photograph was made in India on February 16, 1980, by a group from Williams College led by Jay M. Pasachoff. The missing detail is revealed in the photograph below. The conventional photograph nonetheless shows that the corona is symmetrical, a feature associated with maximum activity in the 11-year sunspot cycle.



DELICATE FEATURES OF THE CORONA during the eclipse of February 16, 1980, were captured by interposing a radially graded filter between the telescope lens and the photographic plate. By sharply reducing the brilliant light from the inner corona the filter makes it possible to record coronal features from the photosphere out to a distance of several solar radii. The bright spots just above the disk of the moon are prominences, consisting of cool, dense material supported in the corona by magnetic fields. Many well-defined coronal streamers are visible. The photograph was made in India by a group from the High Altitude Observatory.

"holes" at both poles, where the corona is much dimmer than it is near the equator. Within the polar holes faint polar "plumes" are often visible. A bundle of these narrow structures diverges from each pole, fading to invisibility within one solar radius or so. The dominant features—streamers, loops and holes—are so prevalent that some investigators believe the corona consists only of such features and there is no such thing as an underlying "quiet corona."

The appearance of the corona varies significantly from one eclipse to another. Even ordinary eclipse photographs are sufficiently different for an observer usually to be able to identify a specific eclipse from its photograph. Coronal images made through radially graded filters vary dramatically. In addition to the general changes associated with the sunspot cycle, each eclipse has its own particular configuration of coronal streamers, loops and other features. A set of photographs made at different times during a single eclipse provides tantalizing evidence for a dynamic corona, whose features change on a time scale much shorter than the interval between eclipses.

In order to study coronal dynamics one would like to observe the corona even when there is no eclipse. Given the brightness of the solar photosphere, this would seem a hopeless task. At visible wavelengths the dense photosphere is a million times brighter than the tenuous corona. Since the corona is much hotter than the photosphere, however, it produces high-energy forms of electromagnetic radiation that are virtually absent in the spectrum of the photosphere. At two million degrees the coronal gas is a strong source of X rays. Except for the X rays released sporadically by solar flares and other local "hot spots" the cooler photosphere is a negligible source of such rays. Accordingly the corona can be observed by pointing an X-ray telescope directly at the sun.

Such observations must be made from space, because the earth's atmosphere is opaque to X rays. The first X-ray telescopes were flown on brief rocket flights in the early 1970's and made crude images of the corona that showed bright areas suggesting regions of high density or temperature. By comparing eclipse photographs made from the ground with X-ray images made from space at about the same time the correspondence between X-ray-bright regions and optically visible coronal features was confirmed.

X-ray imaging technology has advanced rapidly over the past decade. In the mid-1970's astronauts aboard the Skylab orbiting observatory obtained many X-ray images of the corona. A motion picture made from such images shows clearly how coronal features

© 1983 SCIENTIFIC AMERICAN, INC

Wir sprechen die Sprache derer, die unsere Zukunft prägen.

The German-language edition of SCIENTIFIC AMERICAN, SPEKTRUM DER WISSENSCHAFT, has a fully paid circulation that has surged in only two years from 60,000 to 92,000* today.

German industry has evolved as a formidable competitor to leading American, Japanese and other European companies because its technology is of the highest caliber. The well-educated, affluent professionals that read SPEKTRUM DER WISSENSCHAFT are making the decisions and shaping the opinions of German businesses through their technical expertise.

In business, industry, government and public life, SPEKTRUM's readers are emerging as the most influential members of their nation's professional-managerial population.

We speak the language of more than 3,000,000 world shakers in 140 countries each month. We reach the people who make the future happen in German, French, Spanish, Italian, Japanese, Chinese and English.

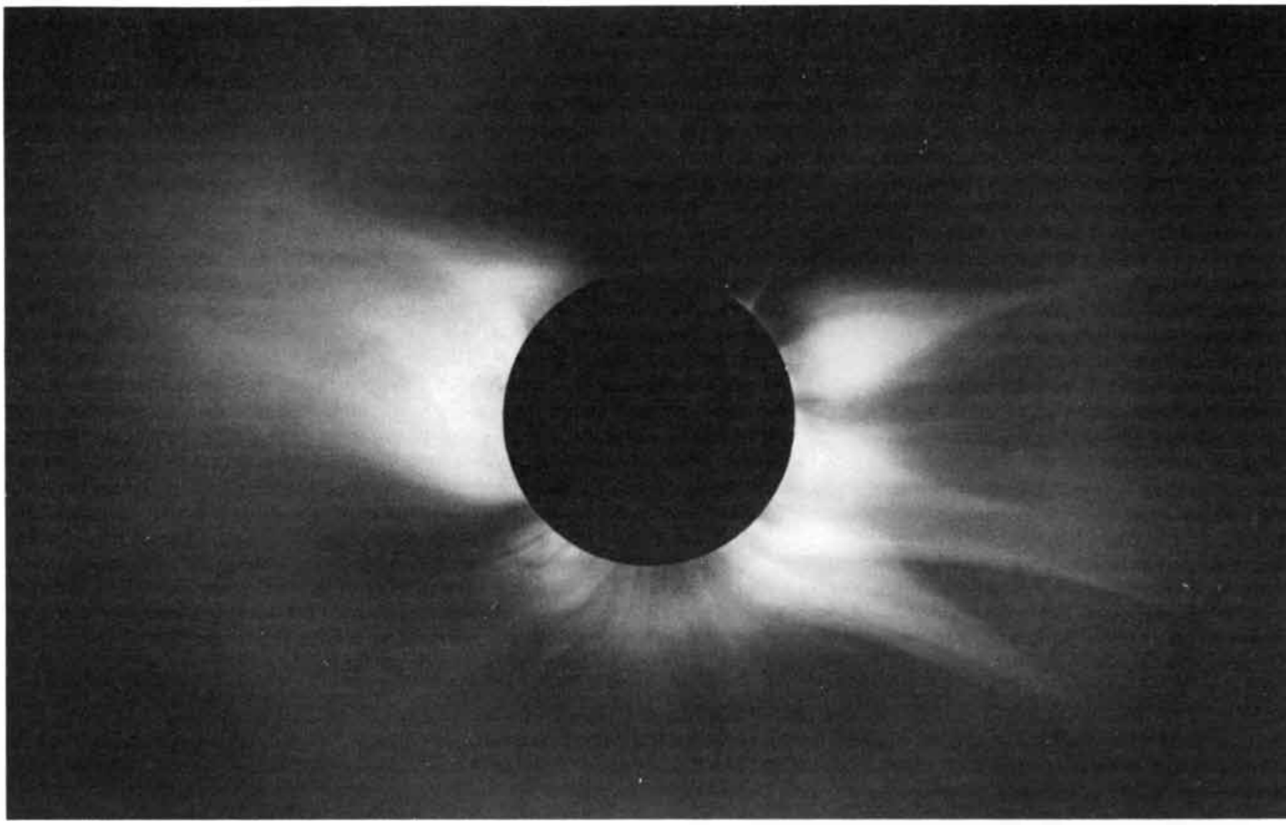
For more information contact:

SPEKTRUM DER WISSENSCHAFT
Anzeigendirektion Deutschland
Mönchhofstr. 15
D-6900 Heidelberg
West Germany
Telephone (011-49) 6221-46037/38

*IVW audited

SCIENTIFIC
AMERICAN
We reach the people who make the future happen.

© 1983 SCIENTIFIC AMERICAN, INC



HIGHLY ASYMMETRIC CORONA was photographed through a radially graded filter during the eclipse of June 30, 1973, when the 11-year sunspot cycle was near a minimum. At such times coronal streamers are concentrated near the solar equator, and the solar poles exhibit "holes": regions of depressed activity. The photograph was made in Kenya by workers from the High Altitude Observatory.

evolve and move with the sun's rotation. Skylab X-ray images were instrumental in establishing that coronal holes exist not only at the poles but also on a smaller scale throughout the corona. Improved X-ray-imaging and spectroscopic instruments on recent spacecraft have supplied data of still higher resolution on the X-ray corona.

Additional information about the corona can be gleaned from the ultraviolet radiation emitted by highly ionized heavy elements in the corona. At these wavelengths too the output of the photosphere is negligible. Ultraviolet data are valuable because particular spectral lines are formed only within narrow temperature ranges, making it possible to probe different layers of the corona, the chromosphere and the transition region between them.

At the other end of the spectrum ground-based radio telescopes can identify and track signals associated with characteristic oscillations of the ionized coronal gas. The frequency of such oscillations is a direct measure of the density of the corona, so that different coronal regions can be probed simply by changing the frequency to which the radio telescope is tuned. By linking several telescopes into an interferometer the evolution of individual coronal features can be followed. At Culgoora in Australia the Commonwealth Scientific and Industrial Research Organization (CSIRO) has set up an instrument called a radioheliograph, consisting of 96 interconnected radio telescopes dedicated to solar studies.

The most direct approach to investigating the corona in the absence of an eclipse is to contrive an artificial eclipse. In principle a thumb held at arm's length could block the light of the photosphere. The reason this does not work is that the earth's atmosphere scatters enough sunlight to overwhelm the coronal light. Above the atmosphere, however, and even on high mountains it is possible to observe the corona by artificially blocking the photosphere. An instrument designed for this purpose is the coronagraph. It basically consists of a small refracting telescope equipped with a disk that occults the sun as the moon does in an eclipse. It includes, however, a number of other occulting devices designed to minimize scattered light and diffraction effects that distort the coronal image. Once the coronal light is isolated it can either be recorded directly or passed through polarizers or filters to isolate special features.

The earliest coronagraphs were developed some 50 years ago. Since then they have been installed at mountaintop observatories, flown on balloons and launched into space. A coronagraph on Skylab discovered coronal transients: rapidly evolving features that appear to eject coronal material into space. There are currently coronagraphs aboard several unmanned spacecraft. The major ground-based coronagraphs are at the Sacramento Peak Observatory in New Mexico, on the Pic du Midi in France and on Mauna Loa in the Hawaiian Islands. In 1979 an orbiting coronagraph operated by the Naval Research Laboratory recorded an event never observed before: a comet colliding with the sun. The newest coronagraphs record their images by means of digital electronic detectors rather than on photographic film. Data from the electronic detectors can be processed by computer to yield enhanced images with high resolution and contrast.

Some of the best coronal images come from a coronagraph on the Solar Maximum Mission satellite. The SMM was launched in February, 1980, near the peak of solar activity in the current sunspot cycle. The spacecraft carries an array of instruments designed for coordinated study of the sun at wavelengths from visible light to gamma rays. For a

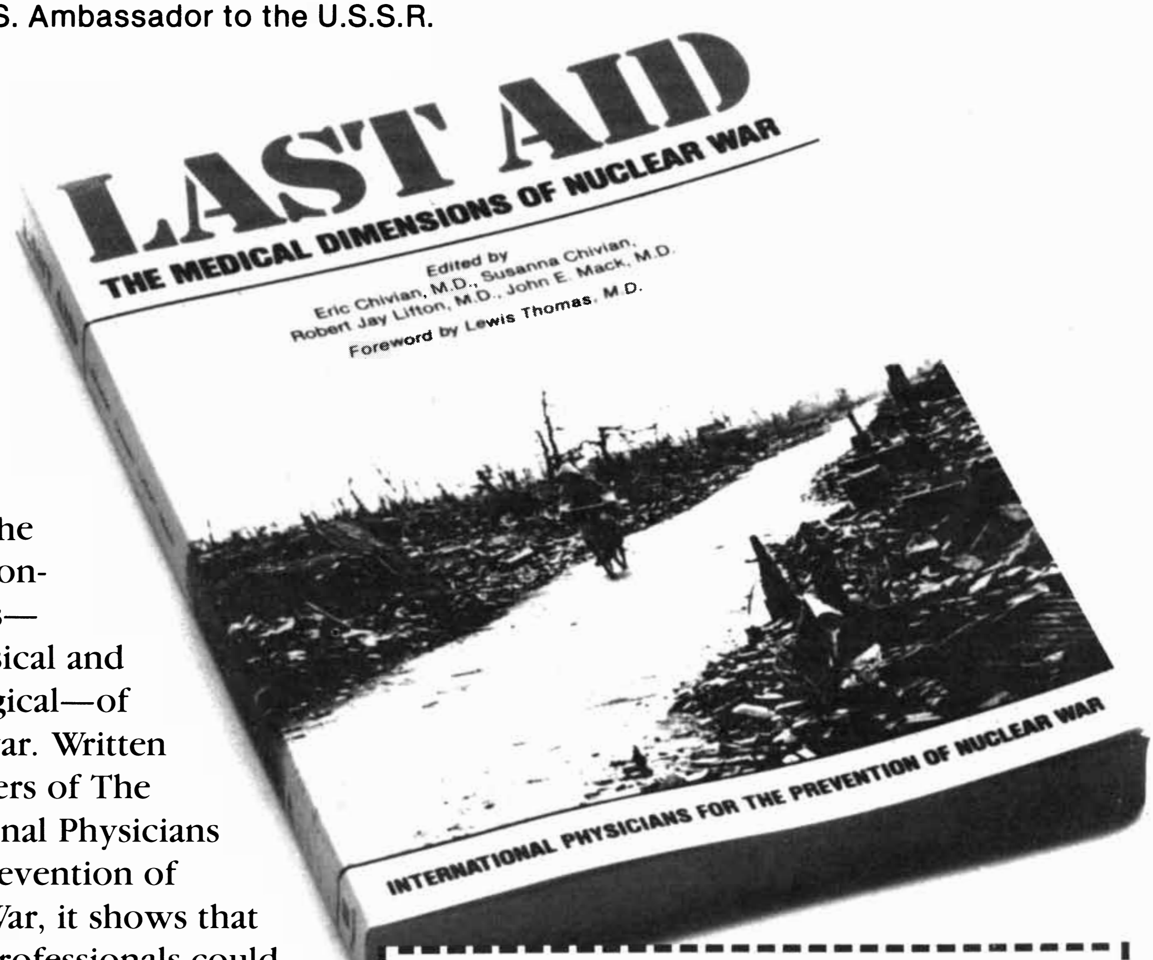
© 1983 SCIENTIFIC AMERICAN, INC

"To ignore this passionate presentation is to turn our backs on the most urgent issue of our time."

Walter Cronkite

"The facts revealed and documented in this book are fundamental for any constructive thinking about the problems of nuclear weaponry, and deserve the widest possible attention."

George F. Kennan, Ph.D.
Former U.S. Ambassador to the U.S.S.R.



Last Aid presents the medical consequences—both physical and psychological—of nuclear war. Written by members of The International Physicians for the Prevention of Nuclear War, it shows that medical professionals could not begin to cope with the overwhelming problems that would ensue in the aftermath of a nuclear holocaust. A reaction to heads of government who speak rationally of a conflict that can have no rational basis, ***Last Aid* is must reading for every intelligent, thinking human being.**

Yes, please send me a copy of *Last Aid.*

_____ I enclose a check for $9.95 (paper) or $19.95 (hardbound), plus $1.50 to cover postage and handling.
(California residents add appropriate sales tax.)

_____ Please charge my ☐ MasterCard ☐ VISA

Account # ______________________ Expires ____________

Signature ______________________________________
(All credit card orders must be signed.)

Name ______________________________________

Address ______________________________________

City ______________ State __________ Zip __________

W. H. Freeman and Company
660 Market Street
San Francisco, CA 94104

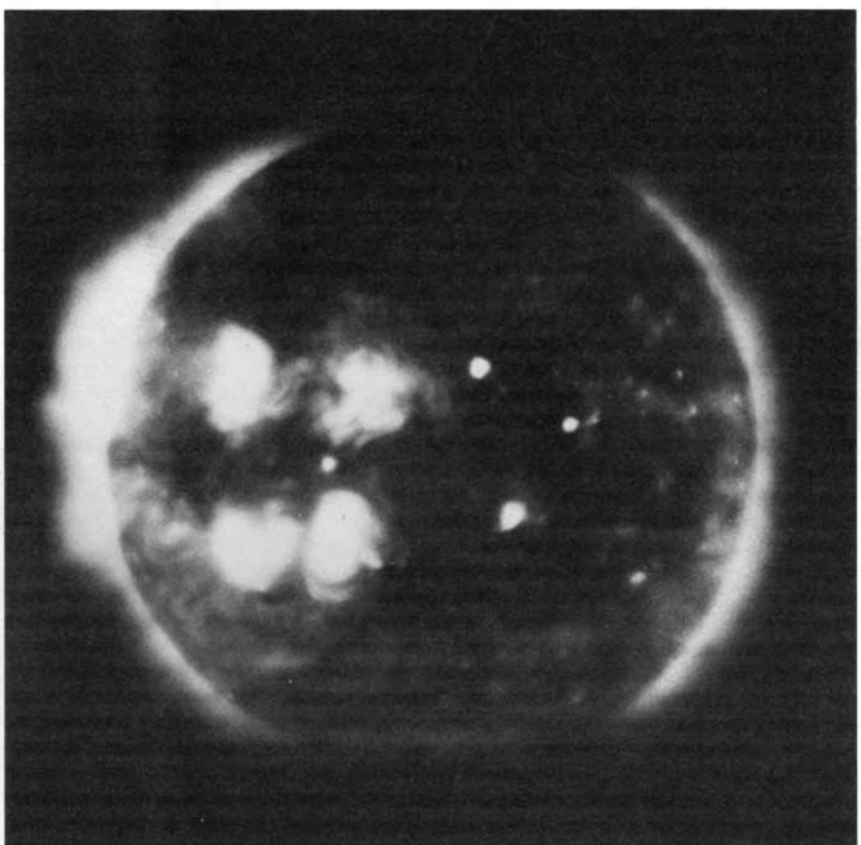
ISBN hardbound 0-7167-1434-5
ISBN paper 0-7167-1435-3
SA

© 1983 SCIENTIFIC AMERICAN, INC

little more than six months the SMM yielded a wealth of data, but then a series of failures in the system that points the spacecraft at the sun left it spinning out of control, cutting short a mission that was projected to last for two years. The SMM, however, is the first spacecraft designed to be repaired in space. It is equipped with a grappling hook and its instrument packages are replaceable. The Space Shuttle mission scheduled for April, 1984, will include a rendezvous with the SMM during which an astronaut will spacewalk to the ailing satellite and stop it from spinning. Once the SMM is stabilized the shuttle's mechanical arm will bring it into the shuttle's cargo bay, where it will be repaired and then redeployed in its orbit. The revived SMM should resume its observations during the sunspot minimum, providing valuable data to compare with those gathered during the maximum.

The coronagraph on the SMM was developed by a group under the direction of Lewis L. House at the High Altitude Observatory of the National Center for Atmospheric Research. In it coronal light is passed through interchangeable filters and polarizers and focused on a television camera tube. Data from the camera are transmitted to the ground in digital form and stored on magnetic tape. Each image from the instrument includes well over a million bits of information and shows one quadrant of the corona. On the ground the images are processed by computer and displayed on a color television monitor. The computer enables the observer to zoom in on selected areas of the image and to enhance particular features.

Before its temporary demise in the fall of 1981 the SMM coronagraph sent back some 30,000 coronal images. A motion picture prepared by Rainer M. E. Illing of the High Altitude Observatory from a sequence of images taken in April, 1980, reveals the corona as it has not been seen before: a restless, dynamic structure of ever changing aspect. Although major coronal features remain recognizable for several days or longer, they clearly evolve from day to day. Smaller features come and go on shorter time scales. Occasional transient events disrupt the corona in a matter of hours. A typical transient event captured by the SMM coronagraph shows a coronal loop suddenly expanding and arching outward at a rate of hundreds of kilometers per second.



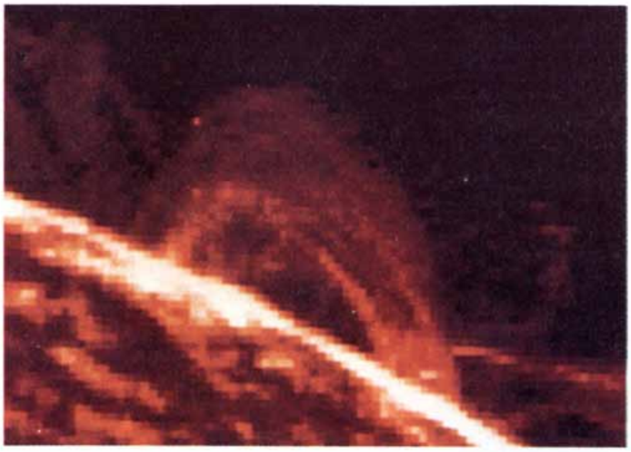
X-RAY IMAGE OF THE SUN records radiation that originates predominantly in the corona, where the temperature is about two million degrees Kelvin (degrees Celsius above absolute zero). The sun's photosphere, at 6,000 degrees K., is much too cool to emit X rays. The X-ray image was made by astronauts aboard Skylab on September 18, 1973, about two and a half months after the solar eclipse shown in the illustration on page 108. Most bright areas in the X-ray image are regions of enhanced coronal density or temperature, corresponding to coronal streamers or loops. The dark areas are coronal holes. In addition to the polar holes a prominent coronal hole can be seen near the equator. Images such as this one demonstrated that coronal holes are not limited to the polar regions of the sun and that they are associated with high-speed streams in the wind of particles from the sun. The image was provided by American Science & Engineering, Inc., which designed the X-ray telescope aboard Skylab.

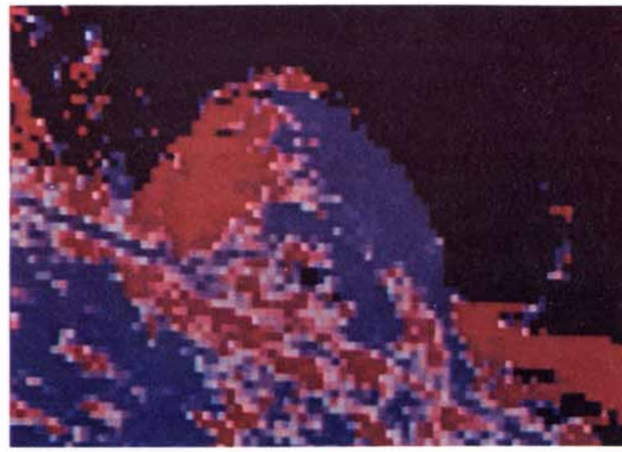
What is the physical nature of coronal features? How can a hot, tenuous gas organize itself into such distinct patterns? Why do some coronal features persist for days and others evolve suddenly and rapidly? Neither the spherically symmetrical force of gravity nor the forces associated with the rotation of the sun can mold the corona into its intricate forms. A clue to the origin of coronal structures is given by the forms themselves. Individual coronal loops and polar plumes suggest the patterns formed when iron filings are sprinkled near a magnet. Just as the filings align themselves with the lines of magnetic force, so the structure of the sun's magnetic field imposes itself on the coronal gas. Magnetic force is responsible for the detail, variety and evolution of coronal structure.

Like the magnetic field of the earth and many other celestial objects, the sun's magnetic field is thought to originate in the interaction of convective motion and rotation of the electrically conducting material in the object's interior. For reasons that are still poorly understood the sun's field is highly variable. Roughly every 11 years the sun's north and south poles reverse. The magnetic poles are strongest near sunspot minimum and give the solar magnetic field the simple overall structure of a magnetic dipole such as a bar magnet or the earth. The dipole structure is evident in the polar plumes seen in the corona near sunspot minimum. At sunspot maximum the solar magnetic field is chaotic, showing no distinct poles but rather many local regions of intense magnetism, often associated with the sunspots. The coronal loops, which at this time are particularly numerous, outline regions of strong local magnetic fields.

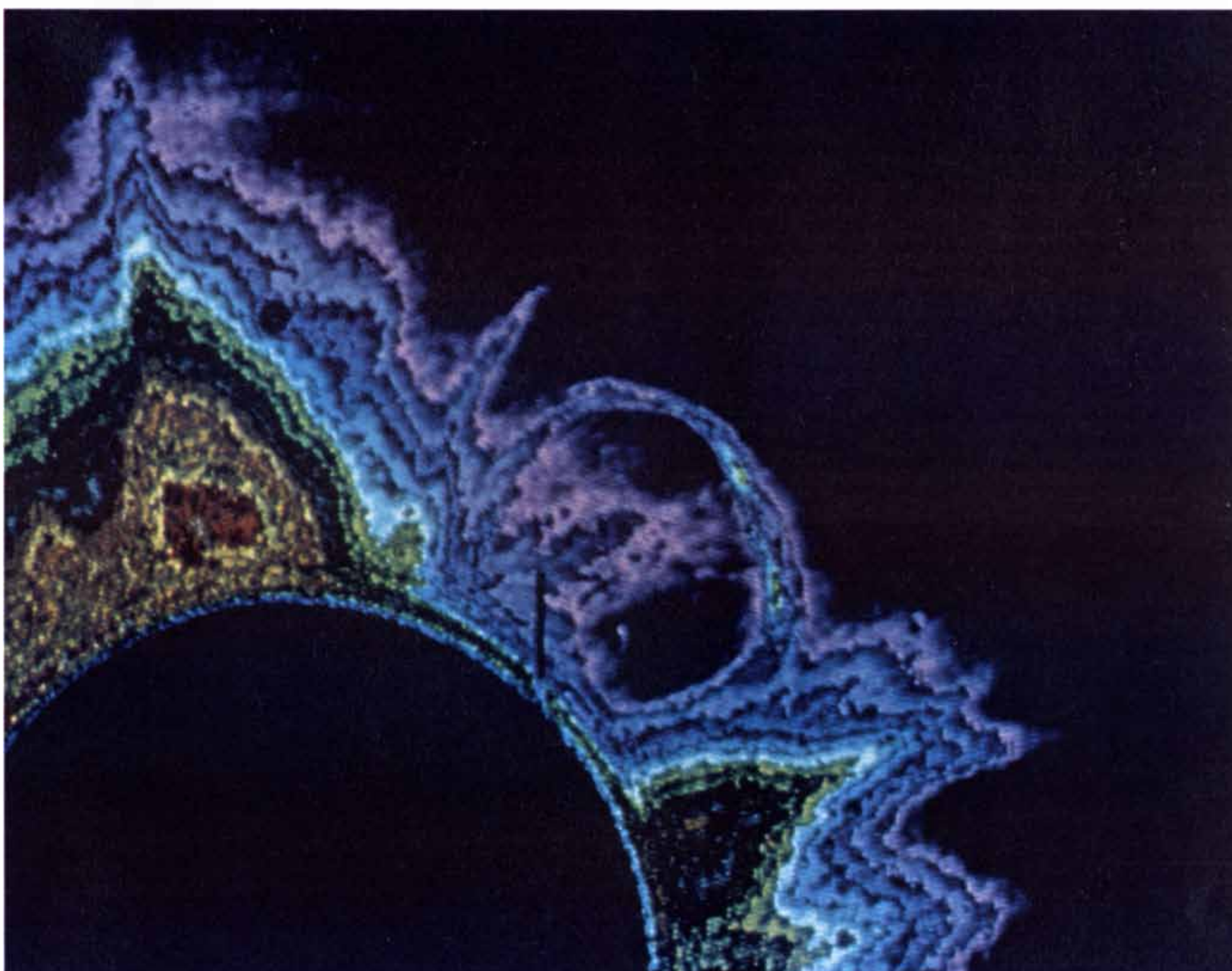
The interaction of the coronal gas and the magnetic field is governed by the basic laws of electromagnetism. We are most familiar with magnetism through the behavior of ferrous metals such as iron. Fundamentally, however, a magnetic field is generated by moving electric charges. In magnetic materials the charges are electrons moving in orbit around atomic nuclei, collectively generating the magnetism of the entire piece of material. In an ionized gas such as the corona the constituent particles of the gas are charged and hence interact with magnetic fields. These electromagnetic interactions make ionized gases so

© 1983 SCIENTIFIC AMERICAN, INC





LOOP NEAR THE PHOTOSPHERE, supported by a strong local magnetic field, has a temperature of about 100,000 degrees K. The images record the ultraviolet radiation at a wavelength of 1,548.19 angstrom units emitted by triply ionized carbon: atoms of carbon stripped of three electrons of their normal complement of six. The image at the left is formed from all the radiation in a narrow band centered at that wavelength. The image at the right is a "velocitygram" of the same loop made by splitting the radiation into two component wavelengths, one slightly shorter than the central wavelength and one slightly longer. The shorter component, represented in blue, is from matter approaching the observing instrument (the ultraviolet spectrophotometer on the Solar Maximum Mission satellite); the longer component, in red, is from matter receding. A possible interpretation of the image is that the right half of the loop is the closer of the two and that material is flowing down toward both feet. Such velocitygrams reveal the dynamic character of the sun's outer layers.



CORONAL STREAMERS and a rising prominence loop were recorded by the High Altitude Observatory coronagraph aboard the Solar Maximum Mission satellite on April 14, 1980. The contrast of the image was enhanced by "effective contouring," one of several computer techniques for bringing out details. The different colors in the image correspond to different values of the density of the corona. This image is actually a high-contrast version of the coronal event appearing in the four-part sequence reproduced on page 105.

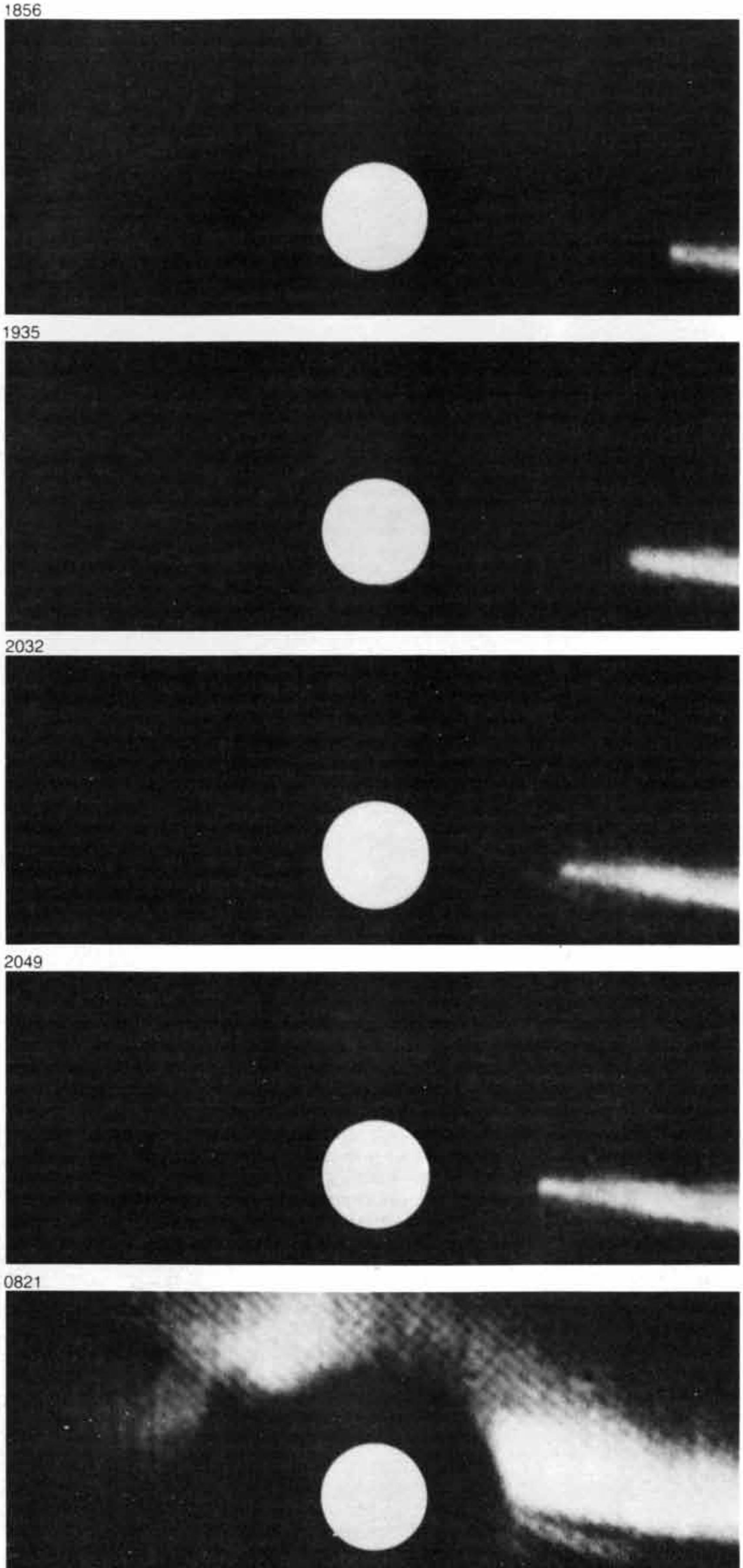
© 1983 SCIENTIFIC AMERICAN, INC

different from ordinary gases that an ionized gas has acquired a special name: plasma. The plasma state is often called the fourth state of matter. On our relatively cool planet earth plasmas are rare. Elsewhere in the universe, however, most matter is believed to exist in plasma form.

A charged particle in a magnetic field is subjected to a force that depends on its charge, its velocity and the strength and direction of the field. The direction of the magnetic force is at right angles to both the direction of motion of the particle and the direction of the field. The force is greatest when the direction of motion and the direction of the field are themselves at right angles, and it diminishes to zero when the direction of motion and the direction of the field are parallel. Since the force is always at right angles to the direction of the particle's motion, it can change only that direction and not the particle's speed. And since there is no force when the motion is parallel to the field, only those components of particle motion that are at right angles to the field are affected.

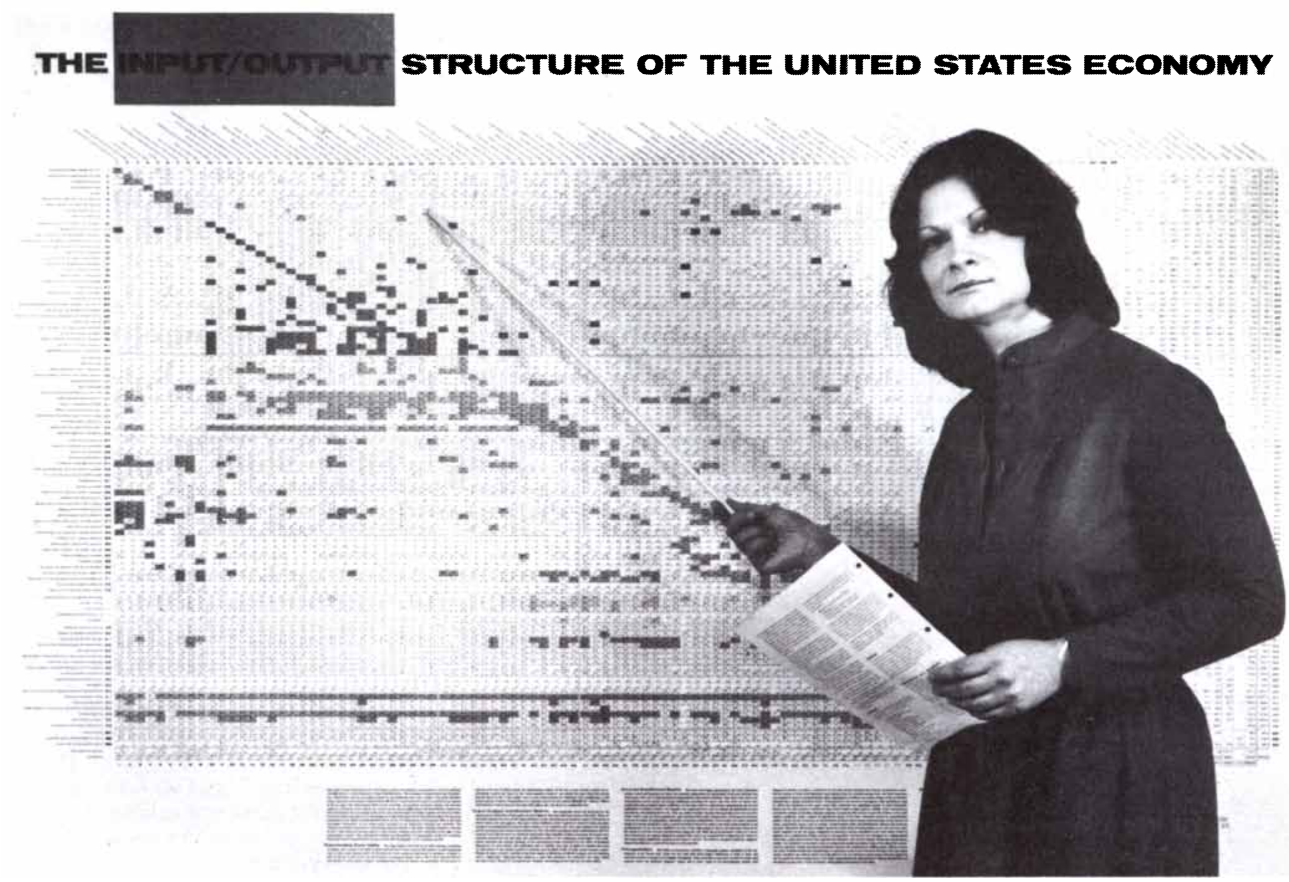
The net result is that charged particles move in helical paths around lines of force in the magnetic field. A particle is free to move in the direction of the field, but if it is pushed at right angles to the field, it moves in a circle rather than in a straight line. Charged particles in a magnetic field are said to be "frozen" to the field, much like beads on a wire. The particles are free to move only along the direction of the field. Nonuniform fields and collisions between particles make the freezing less than perfect, but in a thin plasma such as the corona the assumption that the plasma is frozen to the field is an excellent approximation.

The many charged particles in the corona constitute a gas that exerts pressure just as the earth's atmosphere does. Imagine the coronal gas trying to push its way under pressure into a region where there is a magnetic field. Since the plasma cannot move readily across the lines of force in the field, the field effectively counteracts the gas pressure. A more quantitative analysis shows that a magnetic field indeed exerts pressure on an ionized gas, with the magnetic pres-



COMET COLLIDING WITH SUN, the first recorded event of its kind, was pictured by an orbiting coronagraph operated by the Naval Research Laboratory. The sequence (*top*) begins at 1856 Universal Time on August 30, 1979. The sun itself is completely obscured by the disk of the coronagraph; the white disk represents the sun's size and location. The comet, which is one of the family known as sungrazers, was never detected by earth-based observers. The brightness after the "splashdown" persisted for some 24 hours. Since 1979 two more such collisions have been observed.

© 1983 SCIENTIFIC AMERICAN, INC



WHAT MAKES THE U.S. ECONOMY TICK?

The editors of SCIENTIFIC AMERICAN have prepared a wall chart displaying for the 1980's the Input/Output Structure of the U.S. Economy based on the latest interindustry study from the U.S. Department of Commerce.

The SCIENTIFIC AMERICAN Input/Output wall chart does for economics what the table of elements does for chemistry. It answers at a glance questions about the linkage between the microeconomics of the firm and the macroeconomics of the system; about the web of technological interdependencies that tie industry to industry; about the industry-by-industry direct and indirect consequences of swings in public and private spending; about the impact of change in technology, and about any other topic you can think of. You are rewarded by surprise as well as by confirmation of your hunches. For teaching and practical and theoretical studies, here is a powerful, graphic tool.

In the familiar format of the SCIENTIFIC AMERICAN Input/Output wall charts for the 1960's and 1970's, the wall chart for the 1980's measures 65″ × 52″ and is printed in eight colors. Each of the nearly 10,000 cells in the 97-sector interindustry matrix shows (1) the interindustry commodity flow, (2) the direct input/output coefficient and (3) the "inverse" coefficient. Where the direct input/output coefficient exceeds .01, the cell is tinted in the color code of the industrial bloc from which the input comes. This device, combined with triangulation of the matrix, brings the structure of interindustry transactions into graphic visibility.

A supplementary table displays, industry by industry, the capital stock employed; the employment of managerial, technical-professional, white-collar and blue-collar personnel; the energy consumption by major categories of fuel, and environmental stress measured by tons of pollutants.

The editors of SCIENTIFIC AMERICAN are happy to acknowledge the collaboration, in the preparation of this wall chart, of Wassily Leontief, originator of input/output analysis—for which contribution to the intellectual apparatus of economics he received the 1973 Nobel prize—and director of the Institute for Economic Analysis at New York University.

Packaged with the chart is an index showing the BEA and SIC code industries aggregated in each of the 97 sectors.

SCIENTIFIC AMERICAN

415 Madison Avenue, Dept. I/O, New York, N.Y. 10017

YES, send me the new SCIENTIFIC AMERICAN 1980's INPUT/OUTPUT wall chart—printed in eight colors on heavy paper measuring 65″ × 52″.

I understand my chart(s) will be shipped prepaid, uncreased*, flat-rolled and in a sturdy mailing tube. (Allow 4-6 weeks for delivery)

I enclose a check for $_____ for _____ chart(s) at $35 each, saving me $5 each in handling charges.

Bill me for $_____ for _____ chart(s) at $40 each.

Name ____________________

Company ____________________

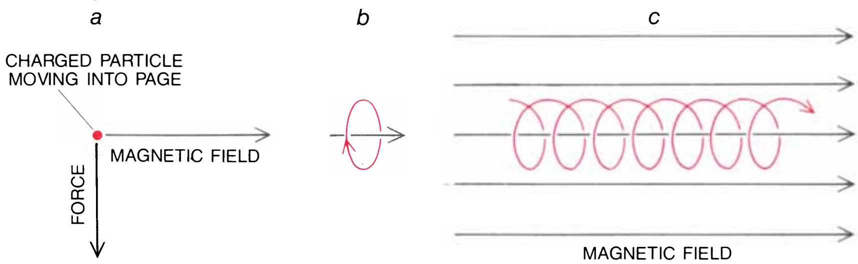
Street Address ____________________

City, State, Zip Code ____________________

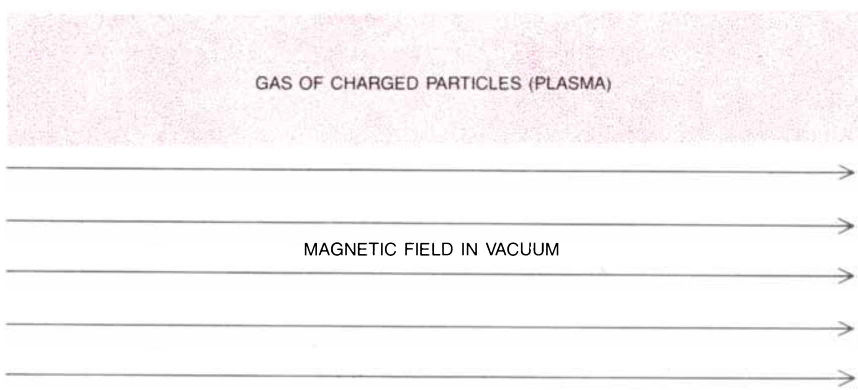
*Charts shipped international require one fold. 2/83

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

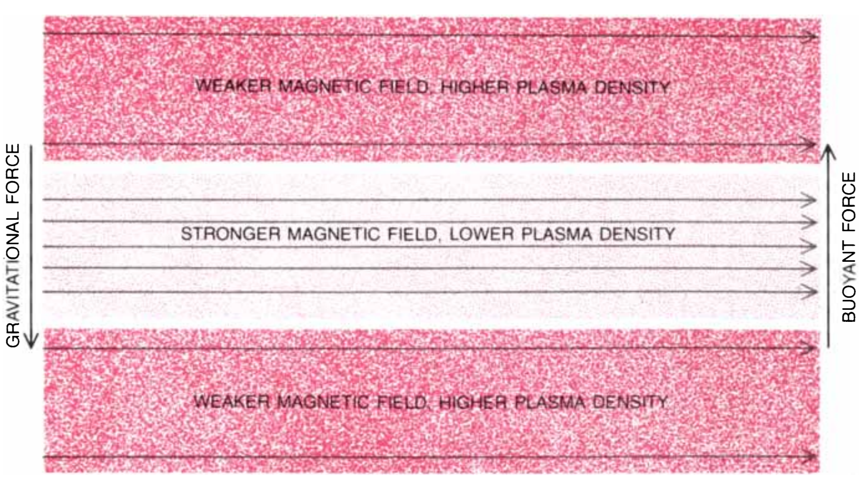
© 1983 SCIENTIFIC AMERICAN, INC



CHARGED PARTICLES IN THE CORONA exhibit behavior characteristic of all charged particles under the influence of a magnetic field. The force on such a particle is at right angles to both the direction of the field and the direction of the particle's motion. For a positively charged particle traveling directly into this page (*a*) in the presence of a magnetic field pointing to the right the force on the particle would be directed downward. As the particle moves, the force continuously deflects it from a straight path so that it assumes a circular orbit (*b*). If the particle also has a component of velocity in the direction of the field, the particle's trajectory becomes a helix (*c*). The particle can move in either direction parallel to the field, but in response to a force acting at a right angle to the field, the particle moves to a helix of larger diameter. A particle of negative charge would circle in a direction opposite to the one depicted here.



MAGNETIC PRESSURE arises because plasma, a gas of charged particles, cannot move readily at right angles to a magnetic field. In the simple case shown here a vacuum containing a magnetic field is immediately adjacent to a plasma under pressure. The plasma is prevented from entering the vacuum by the pressure of the magnetic field. Although the plasma pressure and the magnetic pressure are assumed to be in balance, the configuration is not necessarily stable. A small disturbance in either the plasma or the field can trigger a drastic rearrangement.



MAGNETIC BUOYANCY is exhibited by a region of higher magnetic field strength that is sandwiched between regions of lower strength when all three regions are in a gravitational field. Because the total pressure in each region is the sum of magnetic and plasma pressures, the region of higher magnetic field strength must have a lower plasma pressure if it is to be in pressure equilibrium with its surroundings. If the regions are at the same temperature, lower pressure implies lower density. Since the less dense region is less tightly held by gravity, it tends to float upward. Such buoyancy may account for events in which loops are ejected into space.

sure being proportional to the square of the strength of the field. The behavior of the coronal gas is vastly more complicated than that of the earth's atmosphere, since both gas pressure and magnetic pressure determine the dynamics of the gas.

The coronal plasma does not merely respond passively to the solar magnetic field. Since it is highly ionized and tenuous, the corona is an excellent electrical conductor and can sustain large electric currents. The currents in turn give rise to magnetic fields that modify the original fields. This two-way interaction between the corona and the sun's magnetic field gives coronal structure a richness and variety that is simply unavailable to a nonconducting gas.

The richness of coronal phenomena greatly complicates attempts to explain coronal behavior theoretically. A full understanding of the corona calls for the self-consistent solution of a complicated set of equations in which the magnetic field shapes coronal structure even as coronal currents are modifying the field. In addition the effects of gravity, rotation and ordinary gas pressure cannot be neglected. Needless to say, a full theoretical understanding of all coronal features seen in a particular eclipse or coronagraph image has yet to be achieved.

An important first step has been taken, however, in a mathematical model developed by Gerald W. Pneuman and Roger A. Kopp of the High Altitude Observatory; the model clearly shows the development of coronal streamers at the solar equator. Refinements of equatorial-streamer models by Pneuman, Tyan Yeh of the National Oceanic and Atmospheric Administration and others have led to a detailed understanding of streamer structure. Current research in this area includes attempts to model the association of a coronal streamer and an adjacent coronal hole. A model I have been developing deals with the formation of streamers at high solar latitudes.

The discoveries of the past decade have stimulated much theoretical effort to understand the forces behind coronal dynamics, particularly very rapid phenomena such as coronal transient events. Recent work by B. C. Low of the High Altitude Observatory and me suggests that magnetic buoyancy may be responsible for the uplifting of coronal loops in transient events. Just as a piece of wood is subjected to buoyant forces when it is immersed in water, so regions of stronger magnetic field in the corona are buoyed up by the surrounding gas. Theoretical models show that a coronal loop can suddenly become unstable and rise through the corona either because of the random motion of photospheric gas at the base of the magnetic loop or in response to a violent photospheric event such as a solar flare.

Since the coronal plasma and magnet-

© 1983 SCIENTIFIC AMERICAN, INC

© 1983 SCIENTIFIC AMERICAN, INC

The world's densest computer is now the heart of the world's most 32-bit computer.

Our 32-bit CPU.

From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits—these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.

As a minicabinet, it can handle multiple users.

A rack-mountable version is available, too.

All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

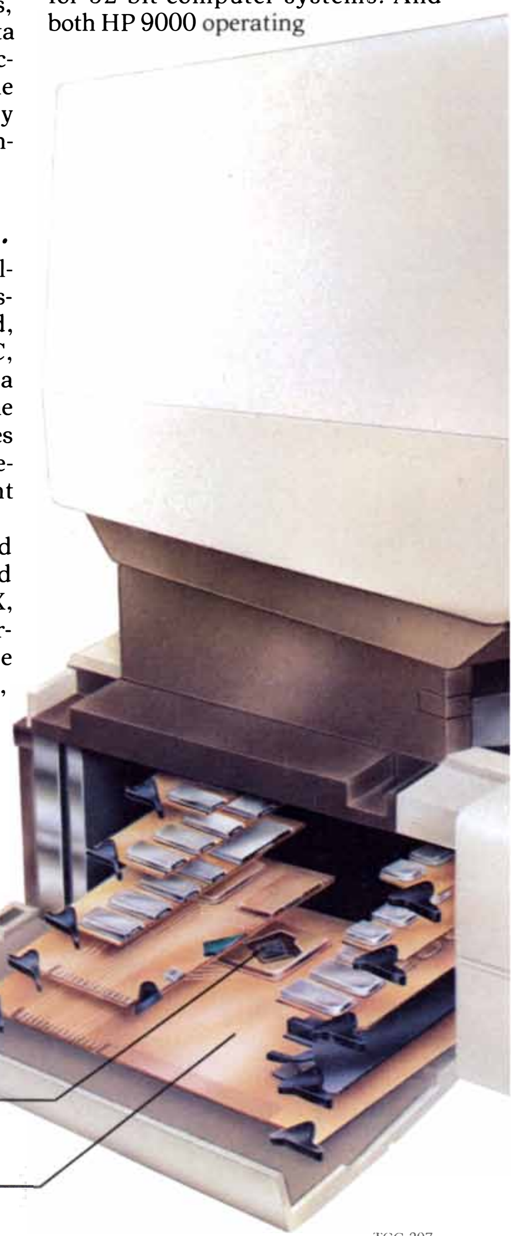
Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX.® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating



The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

© 1983 SCIENTIFIC AMERICAN, INC

chip affordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals. Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 56151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

HEWLETT PACKARD

UNIX is a registered trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

© 1983 SCIENTIFIC AMERICAN, INC

ic field are tightly frozen together, it is impossible to move plasma at right angles to the field without dragging magnetic field lines along. Thus a disturbance in the plasma can create a kink, or bulge, in the field lines. The kink can propagate along the magnetic field much as a kink in a stretched spring moves along the spring. Such a traveling disturbance is a magnetohydrodynamic wave, which carries energy as it displaces both the plasma and the magnetic field from their low-energy equilibrium states. As noted above, magnetohydrodynamic waves may be responsible for transporting energy from the convective layer under the solar photosphere, out through the photosphere and chromosphere and into the corona. Steepening into magnetohydrodynamic shock waves, the waves may dissipate their energy in the corona, generating the two-million-degree coronal temperature.

Why is there interest in coronal dynamics? Does coronal activity have any direct influence on the earth? If one understood the workings of the corona, one would have much insight into the complex behavior of hot, ionized gases in magnetic fields. One might then be able to understand the behavior of more bizarre and distant astrophysical objects, including pulsars, quasars and active galaxies. A better understanding of the corona might also contribute to the harnessing of nuclear fusion as an energy source, since most promising approaches to fusion involve the confinement of hot plasma by magnetic fields.

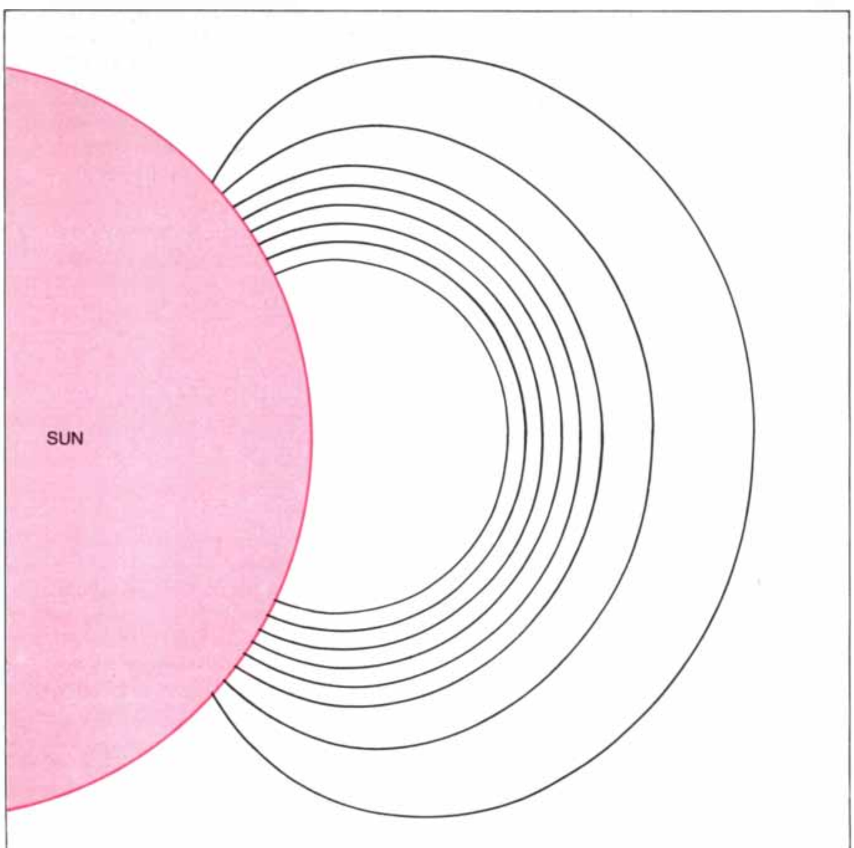
Of greater significance, however, is the fact that the earth is immersed in the corona. It has been recognized since the late 1950's that the high temperature of the corona is responsible for the solar wind: the outpouring of solar gas that rushes past the earth at some 400 kilometers per second. The solar wind is the extension of the corona into interplanetary space.

Since the coronal plasma is tightly gripped by the sun's magnetic field, parts of the corona in which magnetic loops are embedded cannot contribute significantly to the solar wind. With magnetic loops bound to the solar surface at both ends there is no way that coronal gas in the loops can escape into interplanetary space. In those regions where the lines of force in the sun's magnetic field stretch indefinitely outward, however, the solar wind can escape along the field lines. It is now thought that much of the solar wind, particularly its high-speed streams, originates in coronal holes. Coronal transient events may provide another avenue for material to escape from the sun altogether. In many transients an initially closed loop becomes unstable, rising and expanding to the point where its field lines open up, ejecting coronal material into space. Recent observations have also revealed the existence of "coronal bullets": knots of cool, dense material that accelerate rapidly through the corona and may contribute to the outflow from the sun.

The earth's magnetic field shields the earth from the direct impact of the solar wind, but the wind distorts the field. In the direction toward the sun the earth's field is compressed; in the direction away from the sun the field is drawn out into a long tail. The region in which the earth's magnetic field is confined by the solar wind is termed the magnetosphere. Like the corona itself, the magnetospheres of the earth and other planets are laboratories for studying the interactions of plasmas and magnetic fields.

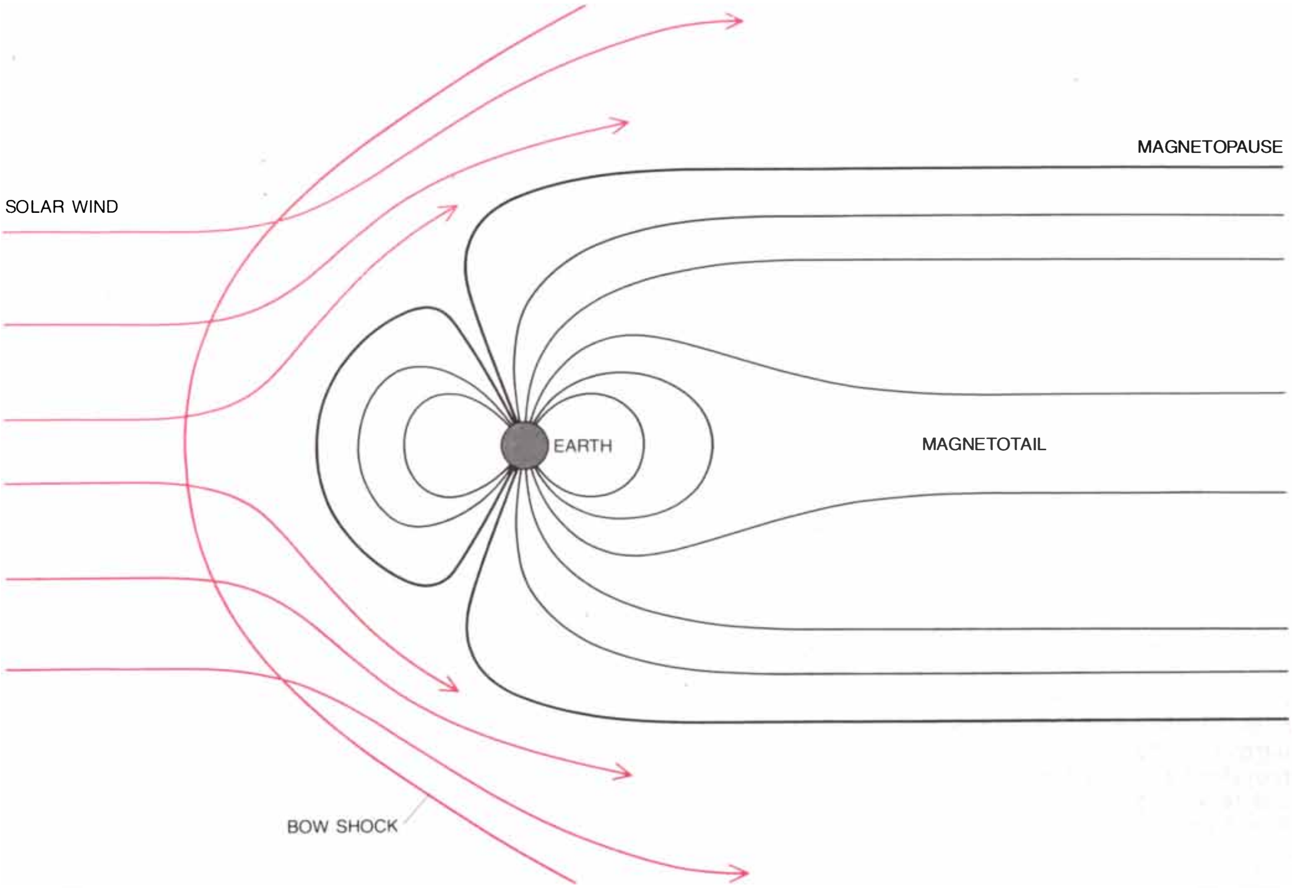
A magnetosphere is an obstacle in the path of the solar wind. Because the wind flows faster than the speed of sound or of magnetohydrodynamic waves, no signals can travel upstream from the magnetosphere to cause the solar wind to adjust to the obstacle. The situation is the same as when an airplane travels faster than sound or a boat moves faster than water waves. In all such cases the fluid adjusts suddenly and dramatically to the obstacle, rather than gradually as it would if its speed were less than that of sound or some other kind of wave. This sudden adjustment is a shock wave, where fluid properties such as pressure, density and velocity change abruptly. The shock wave created by a supersonic airplane is a sonic boom; that of a boat is a bow wave. A huge bow shock stands in space ahead of the earth's magnetosphere, about 60,000 kilometers from the earth. Spacecraft crossing the bow shock can gather valuable information about the structure of shock waves in plasmas.

If the solar wind were unvarying, the magnetosphere and the bow shock would remain at fixed locations determined by a balance between the total pressure of the incoming solar wind and the pressure of the earth's magnetic field. The wind, however, is an extension



DEVELOPMENT OF CORONAL-LOOP TRANSIENT is plotted according to a mathematical model recently devised by the author. The curves represent successive positions of an expanding loop of coronal material embedded in a magnetic field. The top of the loop ascends as the feet of the loop spread apart, perhaps in response to a violent event such as a solar flare or as a result of random motions in the gas at the surface of the sun. The model predicts that the rate of ascent of the loop is progressively faster than the lateral displacement of the feet of the loop. This result indicates an unstable situation where relatively small disturbances at the surface of the sun can initiate rapid changes in the structure of the corona. The model includes the effects of gas pressure, magnetic field and gravity. For mathematical tractability the magnetic interaction is assumed to be confined to thin layers coincident with the plotted curves.

© 1983 SCIENTIFIC AMERICAN, INC



PRESSURE OF THE SOLAR WIND confines the earth's magnetic field to the region called the magnetosphere, which is compressed in the direction facing the sun and drawn out into a long magnetotail in the direction away from the sun. The magnetopause, the boundary between the solar wind and the magnetosphere, acts as a barrier that prevents most of the particles in the solar wind from reaching the earth. Near the earth's poles, however, particles can move along magnetic field lines into the upper atmosphere, giving rise to auroras. Ahead of the magnetosphere is a bow shock that is analogous to the shock wave that generates the sonic boom of a supersonic aircraft.

of the corona, and the corona is itself highly variable in both space and time. As the sun rotates, the spatial irregularity of the corona—here a coronal hole, there a magnetic loop—translates into temporal variations in the solar wind at the distance of the earth. Dynamic events on the sun, such as photospheric flares or coronal transients, also give rise to variations in the solar wind.

In response to changing solar-wind pressure the earth's bow shock and magnetosphere move about, compressing the earth's magnetic field and altering its strength. Changes in the magnetic field induce electric currents in the ionosphere, the conducting layer of the upper atmosphere. This in turn can lead to sudden disturbances of long-distance radio communication. Similar currents can be induced in telephone and power lines, both buried and aboveground, and in the earth itself. Strong disturbances on the sun can therefore result in major disruptions of compass readings, communications and even electric-power transmission. At such times high-energy particles spiral down magnetic field lines at high latitude and give rise to auroras.

Streams of high-velocity solar wind emerge from coronal holes. As the high-speed wind pushes into the slower wind ahead of it, the transition zone between the fast and the slow wind is squeezed ever thinner and the transition becomes more abrupt. By the time the solar wind reaches the orbit of Jupiter, five times farther from the sun than the orbit of the earth, the transitions have steepened into hammerlike shock waves that pound Jupiter's magnetic field. Jupiter's magnetosphere fills a volume so large that even at its great distance from the earth it would appear larger than the sun if we could see it. In response to the pounding of the solar wind it can vary in size by a factor of three in a matter of days. Pioneer and Voyager spacecraft traveling near Jupiter crossed the planet's bow shock and magnetopause repeatedly, indicating that the two structures move rapidly back and forth, now swelling, now contracting.

If the corona is still much in evidence at Jupiter, where does it end? What is the full range of the sun's influence? At some point the decreasing pressure of the solar wind must become weaker than the minuscule pressures of the interstellar gas and the magnetic field of the galaxy. This transition region marks the heliopause: the end of the sun's physical extent. Theoretical calculations a decade ago suggested that the heliopause might be between the orbits of Jupiter and Saturn. With active spacecraft nearing the orbit of Neptune, no sign of the heliopause has been found. It is now believed the solar wind extends well beyond the planetary limits of the solar system.

Beyond the heliopause is a vast bow shock caused by the motion of the entire solar system through the interstellar medium. Driven by variability in the solar corona, the heliosphere, the region within the heliopause, may expand and contract like a balloon. Everything from the thermonuclear furnace at the sun's core out to the heliopause—through the convective zone, the photosphere, the chromosphere, the corona and the solar wind—is a component of the sun. Human beings occupy a privileged position within this giant celestial object, far enough from its core for comfort but close enough to be able to begin to understand the complex dynamics of the extended corona.

© 1983 SCIENTIFIC AMERICAN, INC

Dart-Poison Frogs

In Colombia, Indian hunters poison blowgun darts with highly toxic alkaloids secreted by small frogs. Both the alkaloids and the evolutionary biology of the frogs raise interesting questions

by Charles W. Myers and John W. Daly

Poisonous organisms are so numerous that chemists and pharmacologists are still many years away from analyzing all the noxious molecules found in various microorganisms, plants and animals. Basic research in this area is of interest to the broader fields of ecology and evolutionary biology; it may also lead to new drugs or other useful substances. Some particularly intriguing subjects for investigation are provided by the richness of life in tropical regions. An example in the New World Tropics is the plant genus *Strychnos,* a source of the poison strychnine and of the curare alkaloids. Curare is used by Indian hunters for poisoning arrows and blowgun darts and by anesthesiologists as a muscle relaxant. Another example is provided by a family of frogs, certain species of which secrete one of the strongest animal poisons known. This substance too is used as a dart poison and has found a place in biomedical research. Other species in the same family secrete many additional toxins, and the frogs themselves are also remarkable in several aspects of their biology.

The poisonous frogs and their nontoxic relatives belong to the family Dendrobatidae, which is confined to South America and southern Central America. They occupy a wide range of habitats. Some species are found along streams; others live away from water, on or near the ground in lowland or montane rain forest. A few forest species even spend most of their lives up in the trees. At the other extreme are a few species that live in open dry country, where they find sufficient humidity on shaded ground under low vegetation. In spite of this ecological diversity the dendrobatid frogs share a life style that sets them apart from virtually all other frogs. They are active only in the daytime, and they lay their eggs in moist places on land. They tend their eggs until hatching, when a "nurse" frog carries the tadpoles, literally glued to its back, to a suitable aquatic environment.

The dendrobatid family includes more than 100 species, currently divided among four genera. They are small frogs, varying in head-and-body length from little more than a centimeter to about five centimeters. Whereas among most frogs the male is smaller than the female, in many populations of dendrobatids the male is about as large as the female. Large males are associated with a high degree of territoriality and aggressiveness. The males of such species invest much time and energy in advertising their territory and attracting mates by prolonged calling, and in actively challenging and grappling with other males. Many territorial males are noisy little creatures; their insectlike chirps, peeps or trills contribute noticeably to the daytime sounds of some tropical forests. The sight of two miniature frogs wrestling on a leaf, chirping vigorously at each other, can be amusing to a human observer, but sexual fitness and successful breeding are evolutionarily serious matters. In some species of dendrobatids the female is also aggressive and defends a territory.

The relatively large size of male dendrobatids is also associated with unusual kinds of mating behavior. Among most frogs courtship concludes with the male mounting the larger female and clasping his mate either behind her forelegs or in front of her hind legs. Male dendrobatids clasp the female not around her trunk but around her head, with the backs of the fingers pressed up under the female's chin. This forward position better ensures the fertilization of the eggs being extruded by the female; if a male of similar size were farther to the rear, its sperm could fall to the ground. In some species of dendrobatids clasping is absent from the courtship ritual, and the eggs are fertilized in the course of rather complicated maneuvering by both the female and the male.

Depending on the species, dendrobatid eggs are laid in leaf litter, in crevices under rocks, on shaded leaves aboveground or in such tropical forest growths as bromeliads and arums. The size of the egg clutch is small compared with the hundreds or even thousands of eggs laid by many water-breeding frogs; it can be as few as one or two eggs among the smallest dendrobatids and is rarely more than 30 or 40. In many dendrobatid populations a small clutch size is partially offset by the fact that breeding is continuous throughout the year. Some individual frogs can breed every month. Statistics on this point are provided by observations of a pair of frogs of the species *Dendrobates tricolor* kept at the American Museum of Natural History for more than two years. These frogs mated from two to four times a month. Clutches were produced every 10th day on the average and consisted of from 12 to 30 eggs. Even with this prodigious effort the female laid only about 600 eggs each year.

Once they are deposited the eggs may be "guarded" by either parent or may be left alone except for periodic brief visits. On such a visit the frog may moisten the eggs, apparently with water from its bladder. The nurse frog occasionally wriggles its hindquarters into the egg mass, an action that ultimately aids the hatching process and helps the tadpoles to squirm up onto their parent's back. There they adhere to a patch of mucus secreted by glands in the frog's skin. Adhesion is facilitated by the flat or slightly concave belly that is characteristic of dendrobatid tadpoles. In some species the attachment is no more than superficial; in others the mucus is gluelike and the tadpoles are firmly fixed to the adult's back. Again depending on the species, the tadpoles may remain attached to the nurse frog for an interval ranging from a few hours to more than a week. In this period the tadpoles may grow to some extent by absorbing yolk reserves.

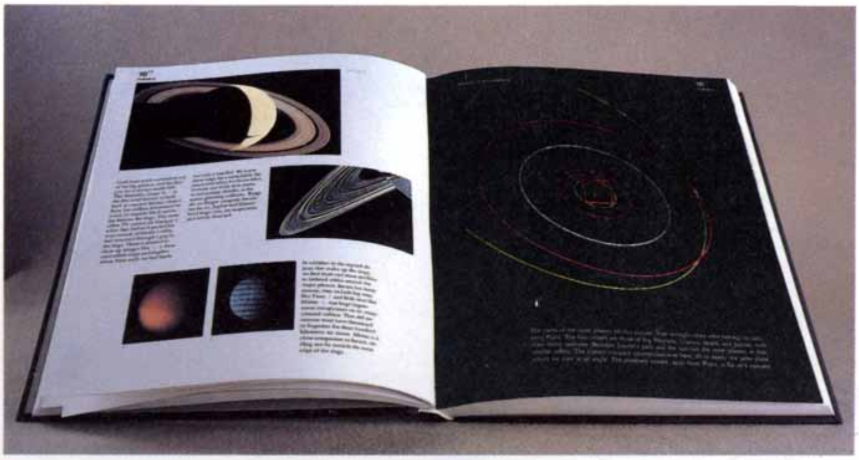
The nurse frog eventually travels to a suitable body of water, where a period of soaking loosens the mucous bond and the tadpoles swim free. Many dendrobatids take their tadpoles to small streams. Other species release them in the small

© 1983 SCIENTIFIC AMERICAN, INC

POISONOUS FROGS of the genus *Phyllobates* are shown 1.7 times life size in these paintings by David M. Dennis. The first two frogs at the top, from the left, are *P. lugubris* and *P. vittatus,* two Central American species that secrete relatively small amounts of toxins from glands in their skin. The striped frog at the upper right is *P. aurotaenia,* one of three Colombian species used for poisoning blowgun darts. In the middle row are two specimens of *P. bicolor,* the second of the three Colombian species. They show extremes in size and color variation; the legs, of a hue different from the body, vary from pale green to black. At the bottom is the third and most poisonous Colombian species, *P. terribilis.* The immature *P. terribilis* (*left*) is striped like its relatives at the top, but its stripes are being obliterated by the spread of bright pigment. The adult *P. terribilis* (*right*) may be yellow, orange or pale green, but normally it is uniformly colored above and below. *P. terribilis* is a bolder, less secretive frog than the other species and is at least 20 times as toxic. It is potentially dangerous even to handle.

© 1983 SCIENTIFIC AMERICAN, INC

SCIENTIFIC AMERICAN LIBRARY



MAN'S CORNER OF THE UNIVERSE lies open here at the pages in *Powers of Ten* that display the universe at the 10^{13} meters dimension (10 billion kilometers), about the diameter of the solar system.

POWERS OF TEN

A wondrous journey across the universe from its largest dimension to its smallest

Here is an atlas of the universe and a handbook to its exploration by the disciplines of science. In vivid color illustration—from the informed imagination and from the instruments of science—*Powers of Ten* shows how the universe looks from the largest to the smallest dimensions now in the grasp of human understanding.

This spectacular journey, with its sweeping changes in perspective, proceeds in ordered steps, each a multiple of ten of the last step or of the next.

There are 42 of these equal logarithmic steps from the outermost regions to the innermost recesses of the known universe. More than half were taken in the lifetime of scientists now alive.

The visual magic of this first-time-in-history journey is the bold invention of Charles and Ray Eames, who have here turned their mastery of multi-media technology to book design. Your tour guides are Philip and Phylis Morrison. They supply instructive and poetic pages of commentary, answering questions and showing how the sciences get those answers at each of the 42 steps.

The SCIENTIFIC AMERICAN LIBRARY publishes *Powers of Ten* as its premier volume because it sets out the agenda of the books that are to come. Each succeeding book further illuminates the ideas and images displayed, step by step, on the pages of *Powers of Ten*.

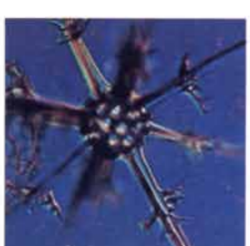
The SCIENTIFIC AMERICAN Library opens an entirely new approach to understanding and sharing in the work of science. Each specially commissioned volume

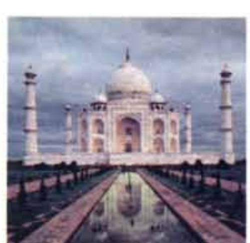
. . . comprehends a large topic that engages more than one discipline and, typically, recurs in many articles in SCIENTIFIC AMERICAN;

. . . is written by a scientist who has made his own pivotal contribution to the work on the topic;

. . . rewards its reader not with "information" but, thanks to the sympathetic collaboration of the author and editor in the preparation of text and illustration, with command of the concepts central to understanding of the work and appreciation of its meaning.

Readers of this magazine are invited to accept membership in the SCIENTIFIC AMERICAN Library. Members are entitled to receive for inspection, without any charge, each volume as it is published. Membership, cancellable at any time, will gather in time a reference library of the member's own choosing, with a reach and accessibility unprecedented in the literature of science.

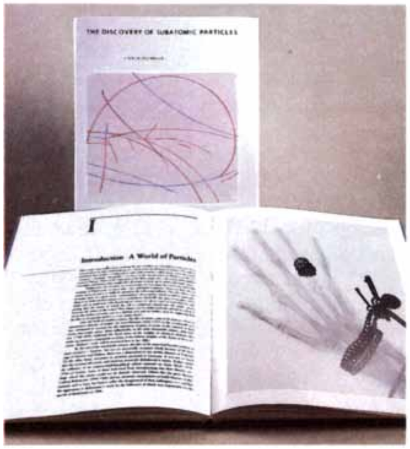




A SAMPLING OF THE DIMENSIONS of the known universe explored on the journey through the Powers of Ten is shown here. At center, the Taj Mahal adorns the 10 meter world in reach of our unaided senses. Dimensions grow smaller to the left: first, crystalline skeletons of single-celled diatoms at 10^{-4} meter (.1 millimeter) and at far left, tungsten atoms, 10^{-10} meter in diameter, brought into visibility by a field-emission microscope. At right, looking across the Indian Ocean toward the South Pole through the window of Gemini II, the curvature of the Earth, 10^{7} meters in diameter. At far right, a quasar, 10^{25} meters away (one billion light-years), caught by a radio telescope and imaged by computer graphics.

© 1983 SCIENTIFIC AMERICAN, INC

SUBATOMIC PARTICLES
by Steven Weinberg



DISSECTION OF ATOM began with accidental discovery of X-rays by Roentgen in 1895.

As everyone knows, the once irreducible atom has been shown to be composed of still more fundamental particles: the electron, proton and neutron. But how do we know this? In *The Discovery of Subatomic Particles* the reader comes to know the answer just as Steven Weinberg and other physicists did, thereby sharing in the intellectual enterprise that has reshaped physics in the twentieth century.

We do not accept scientific truth on faith or authority. We know truths of this kind are grounded on experiment and observation—operations we might have performed ourselves, but in any case can perform inside our heads. With Weinberg the reader of this book reenacts the historic succession of experiments that disassembled the atom. From each experiment to the next, the reader will gain a confident understanding of the laws of physics that govern the events encountered.

HUMAN DIVERSITY
by Richard C. Lewontin

The relevance of science to human values finds compelling demonstrations in the proper understanding of the genetics of our species. In *Human Diversity* Richard Lewontin shows that each human being differs from all others because of the interaction of genetic differences, environmental differences, and chance events that occur during development. There is no unique "normal" genetic constitution.

The genetic markers that supposedly divide the species into races are only a trivial fraction of the total spectrum of diversity in which each person finds singularity. The genetic endowment held in common across racial, national and cultural boundaries is a consequence of the biological history of our species. The segregation of people into social castes can never again invoke genetics as its rationalization.

FOSSILS
by George Gaylord Simpson

One of the architects of the modern evolutionary synthesis here establishes the nature—and the true splendor—of the solid evidence upon which much of the theory and the facts of evolution rest. Simpson declares: "The primary record of the history of life is written in the successive strata of rocks as in the pages of a book. Fossils may be called the writing on those pages. Traces of organisms living at successive geological times, they represent once living things and should be seen as such. They must be put in their sequence in time. They were influenced by and bear witness to geographic and geological changes on the earth. They are basic materials for the study of organic evolution. Their study combines historical geology and historical biology into one great synthesis."

THE SOLAR SYSTEM
by Roman Smoluchowski

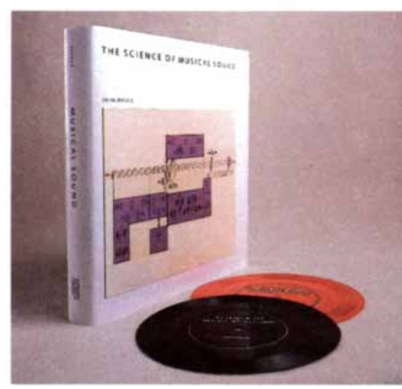
The fine-grained picture of the Earth and its immediate cosmic neighborhood now in the possession of mankind is here comprehensively assembled. The author has brought together the latest graphic images from the satellites and from earthbound telescopes as well. (We know the terrain of cloud-wrapped Venus from radar echoes returned to the 600-foot radio telescope at Arecibo in Puerto Rico.) He shows us our solar system in time as well as in space. In the wealth of images of the Sun, of the planets, of their satellites, of the swarming asteroids and of the cometary messengers from the primordial ice clouds on the outermost circumference of the system, we comprehend the origin, the history and the fated future of our Sun and its planets.

MUSICAL SOUND
by John R. Pierce

In this volume, the author will share with you his joy in the discovery that understanding from physics and mathematics enlarges and enhances one's appreciation of music. Thereupon you will find yourself in the company of a long succession of natural philosophers who have pursued just that strategy.

John R. Pierce is the latest of these natural philosophers. To his work in music he brings his study of sound, its transmission and reproduction, at Bell Telephone Laboratories plus the insights of such contemporary composers as Pierre Boulez, John Cage and Edgard Varèse.

In the live recordings that come with this volume Pierce demonstrates something of what the psychology of acoustics has learned about the perception, the illusion and the effect of sound. He imbues the reader with his confidence that rational enquiry into this most intensely subjective of all aesthetic experiences will open up new realms of sound to that experience.



SOUND SHEETS, for 33-r.p.m. turntable, demonstrate acoustic illusions discussed in book.

Charter Membership Reservation Form

To SCIENTIFIC AMERICAN LIBRARY
Holmes, Pennsylvania 19043

Please enroll me as a Charter Member of the SCIENTIFIC AMERICAN LIBRARY. Send me the Premier Volume — *Powers of Ten* — by return mail. Thereafter, I am entitled to receive for inspection a new volume every two months, as it is published, for fifteen day's *free examination.*

I need send no payment now. For each volume I keep I will be billed $19.95 plus postage and handling. This price is guaranteed for at least one year, regardless of any price increases asked of new members.

I understand that I may cancel my Charter Membership at any time for any reason.

NAME ____________________
(Please print clearly)

ADDRESS ____________________

CITY ____________ STATE ________ ZIP ________

Canadian residents: Please pay in U.S. dollars. FD

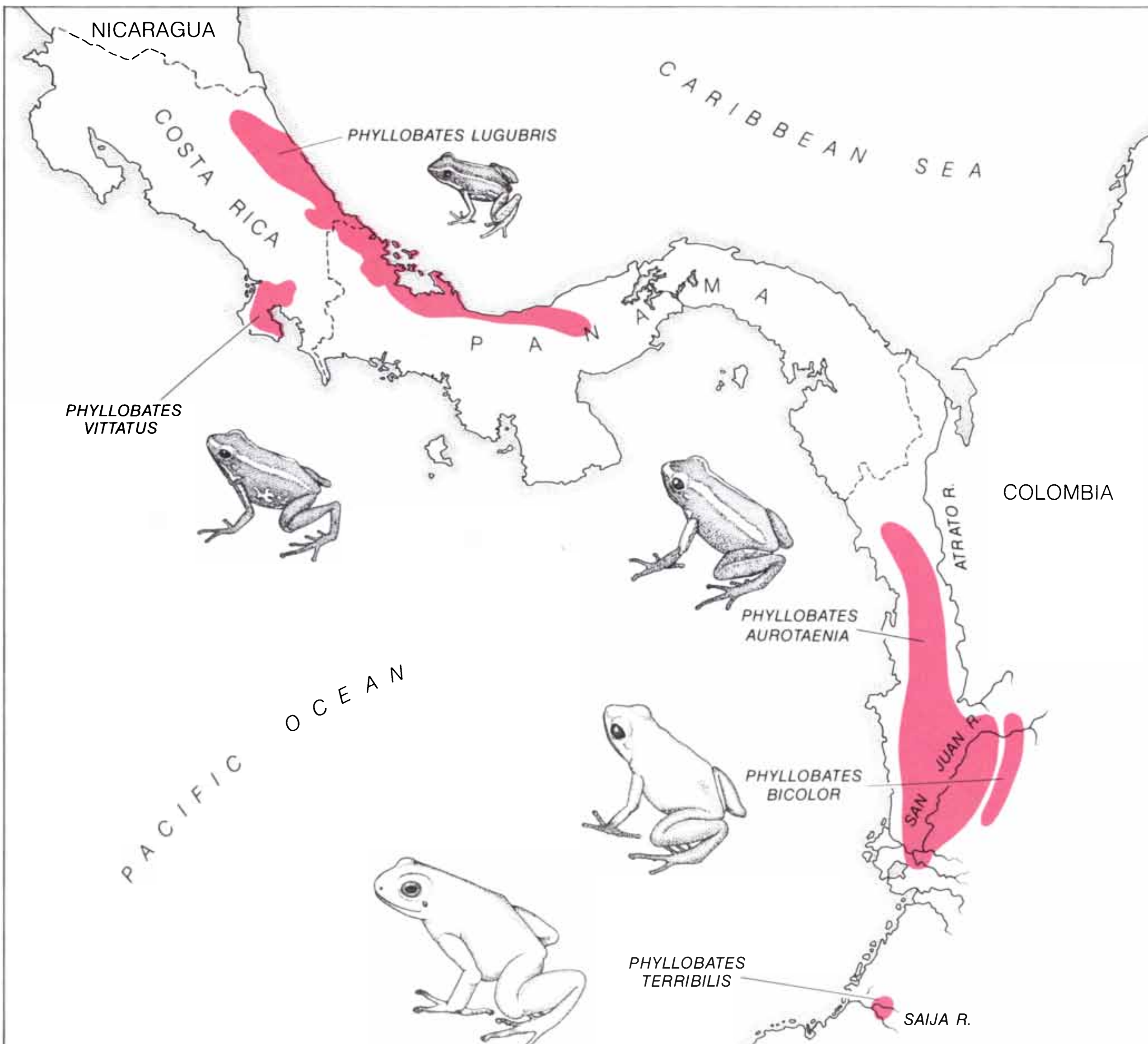
© 1983 SCIENTIFIC AMERICAN, INC

pools of rainwater that collect in bromeliads or between the leaf and the stem of certain other tropical plants. The nurse frogs of one species, *Dendrobates auratus,* often release their tadpoles in water retained by hollows in tree trunks. The tadpoles of *D. auratus* tend to resort to cannibalism in this nutrient-poor environment, and only one of them may survive to maturity.

Rainwater in small bromeliads and in the axils of leaves also seems to hold few nutrients, a circumstance that helps to explain a surprising observation made recently by Peter Weygoldt at the University of Freiburg in West Germany. Working with laboratory specimens of *Dendrobates pumilio,* a species that releases its tadpoles in such places, Weygoldt found that the females regularly visit their free-living tadpoles to deposit unfertilized eggs in the water. The tadpoles then feed on the eggs!

Biologists sometimes attempt to determine the relative "cost" of parental care by male or female animals, because such care may impose limitations on reproduction or may expose the parent to predation. These are factors to be weighed against the benefit of ensuring success for one or a few sets of eggs or for sperm from an individual parent. Among dendrobatid frogs the reproductive cost may be higher in species (such as *D. pumilio*) where the nurse frogs are usually females than it is in species where the nurse frogs are usually males. Of the *D. tricolor* pair at the American Museum the male was able to care for several clutches of eggs at a time. Furthermore, it vocalized frequently and presumably would have mated with other females if they had been available. It fed even while it was carrying on its back a full load of tadpoles, which were released into the water within hours of hatching.

It has been supposed that the sex of the nurse frog (male in most species and females in some) is a species-specific trait. In a few species, however, both sexes have been found carrying tadpoles, a fact that is hard to explain. How can such a novel behavioral trait be so variable? Experiments with the *D. tricolor* pair suggest that the answer may involve the outcome of sexual competition for the nurse role. The female of the pair was both able and willing to



DISTRIBUTION OF THE FIVE SPECIES of the dendrobatid genus *Phyllobates* is patchy. The three species found in western Colombia are the only frogs of any kind known to be used by Indian hunters for poisoning blowgun darts. The two species found in Costa Rica and Panama secrete small amounts of the same toxins as their Colombian relatives, but they do not inhabit the areas where blowguns are used.

© 1983 SCIENTIFIC AMERICAN, INC

be the responsible parent but was kept from the role by the territorial aggressiveness of the male. The female stayed with her clutch of eggs for the first hour or so; thereafter she was apparently allowed no further visitation rights. If, however, the male was removed after breeding, the female cared for the eggs and eventually carried the tadpoles as well. If the male was not removed until a few days later, the female showed no reluctance to cannibalize either her eggs or her tadpoles as though they belonged to another female. Such cannibalism may be a mechanism by which population density is regulated in some species of dendrobatids.

Much remains to be learned not only about the natural history of dendrobatids but also about their evolutionary relations and classification. The approximately 130 species are currently placed in four genera: *Atopophrynus, Colostethus, Dendrobates* and *Phyllobates. Atopophrynus* has only one species. *Colostethus* has more than 70, mainly brown in color. With a few exceptions these species are nontoxic. Most of the approximately 50 species of *Dendrobates* and the five species of *Phyllobates* are characterized by bright "warning" coloration. The brilliance of these frogs advertises the presence of poisonous or at least distasteful skin secretions that are effective in deterring many potential predators. Toxic dendrobatids are not, however, completely immune from predation; for example, they are preyed on successfully by large spiders and certain snakes. The frogs' secretions are released from microscopic glands in the skin at times of stress, and a predator that has seized a frog in its mouth will experience sensations of burning, numbing or foul taste that tend to make it drop the prey.

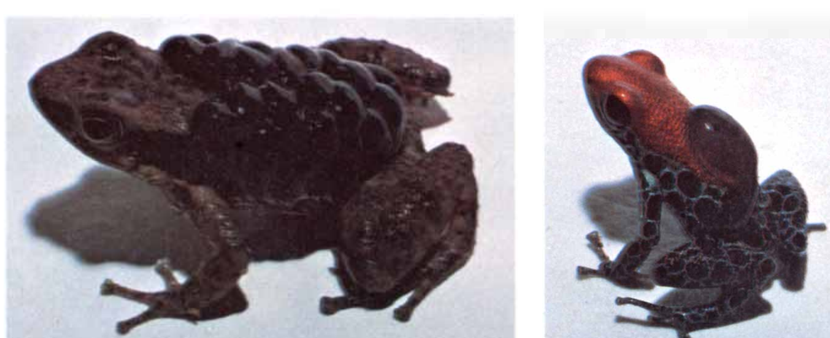
The most toxic dendrobatids are three species of *Phyllobates* from west of the Andes in the Pacific drainage of Colombia. These frogs secrete a poison that is much stronger than curare. Just as curare is a widely used poison for arrows and blowgun darts east of the Andes, so these dendrobatid toxins serve as dart poisons in western Colombia. Possibly the earliest report of their use is one written by Captain Charles Stuart Cochrane, who explored in Colombia in 1823–24 while he was on leave from the British navy.

CARE OF THE YOUNG by the male or female parent includes attention to the eggs. Here a male *Dendrobates silverstonei* watches over a clutch deposited on a leaf on the forest floor.

Cochrane, crossing the western Andes on foot, had his attention drawn to frogs "called *rana de veneno* [by the Spanish], about three inches long, yellow on the back, with very large black eyes." "Those who use [their] poison catch the frogs in the woods, and confine them in a hollow cane, where they regularly feed them until they want the poison, when they take one of the unfortunate reptiles and pass a pointed piece of wood down his throat, and out at one of his legs. This torture makes the poor frog perspire very much, especially on the back, which becomes covered with white froth: this is the most powerful poison that he yields, and in this they dip or roll the points of their arrows, which will preserve their destructive power for a year. Afterwards, below this white substance, appears a yellow oil, which is carefully scraped off, and retains its deadly influence for four or six months, according to the goodness (as they say) of the frog. By this means, from one frog sufficient poison is obtained for about fifty arrows."

Cochrane's *Travels in Colombia* seems not to have been widely read. A few later observers independently verified this curious use of frogs, however, and unsuccessful attempts were made to analyze samples of the poison scraped from Indian darts. The Swedish anthropologist S. Henry Wassén carried out field investigations in Colombia in 1934 and in 1955. Wassén determined that frog poison was used by two related Chocó Indian groups along the San Juan River. The frogs he obtained from the Indians, erroneously identified by zoologists as *Dendrobates tinctorius,* were actually *Phyllobates aurotaenia* and *P. bicolor.* Wassén also got some frog-poison darts from an isolated group of Chocó far to the south, on the Saija River. The frogs at this locality were collected by our colleague Borys Malkin and us in the early 1970's. Our specimens represented an undescribed species, which proved to contain at least 20 times more poison than its relatives along the San Juan. Because of the frog's awesome toxicity, we gave it a name with obvious connotations: *Phyllobates terribilis.*

The northern Chocó Indians poison their darts with the secretions of *P. bicolor* and *P. aurotaenia* much as described by Cochrane a century and a half ago. The frogs are impaled on a stick and are sometimes held near a fire before the darts are poisoned. The southern Chocó Indians, however, poison their darts simply by wiping them across the



"NURSE" FROGS carry their tadpoles to a suitable body of water after hatching. A female of the species *Colostethus inguinalis* from Panama (*left*) is shown with a load of 27 tadpoles. A male of the small Peruvian species *Dendrobates reticulatus* (*right*) carries only one. The young may remain attached to the parent in this manner for a few hours or for more than a week.

© 1983 SCIENTIFIC AMERICAN, INC

DATSUN 200-SX

IT'S GOT PERFORMANCE DOWN TO A SCIENCE.

The sleek Datsun 200-SX is a scientific paradox. Although it burns gas "faster" than conventionally powered cars, the 200-SX gets even better gas-mileage ratings this year than last. And last year's ratings were already remarkably high for a car that won two dozen races in just nine months.

How? The 2.2-liter Nissan NAPS-Z engine. A microprocessor continuously adjusts the fuel/air mixture injected into the engine's crossflow hemihead combustion chambers and selects which of each cylinder's twin sparkplugs to fire, a unique concept in engine design.

Combustion in this "fast-burn" system is completed within 90° of crankshaft rotation. Ordinary engines don't complete combustion until 170°, wasting both power and fuel. The Nissan NAPS-Z saves both by burning gas "faster."

Most important, you can have this scientific advance without compromising looks or luxury features. This unique power plant delivers its high performance wrapped in a sculptured body that turns science into art. See and test drive the 200-SX at your Datsun dealer now.

38 EPA EST HWY | 27 EPA EST MPG

Use EPA estimated mpg for comparison, with standard 5-speed. Actual mpg may differ depending on speed, trip length and weather. Highway mpg will probably be less.

BELT YOURSELF

DATSUN
WE ARE DRIVEN

NISSAN

Product of NISSAN

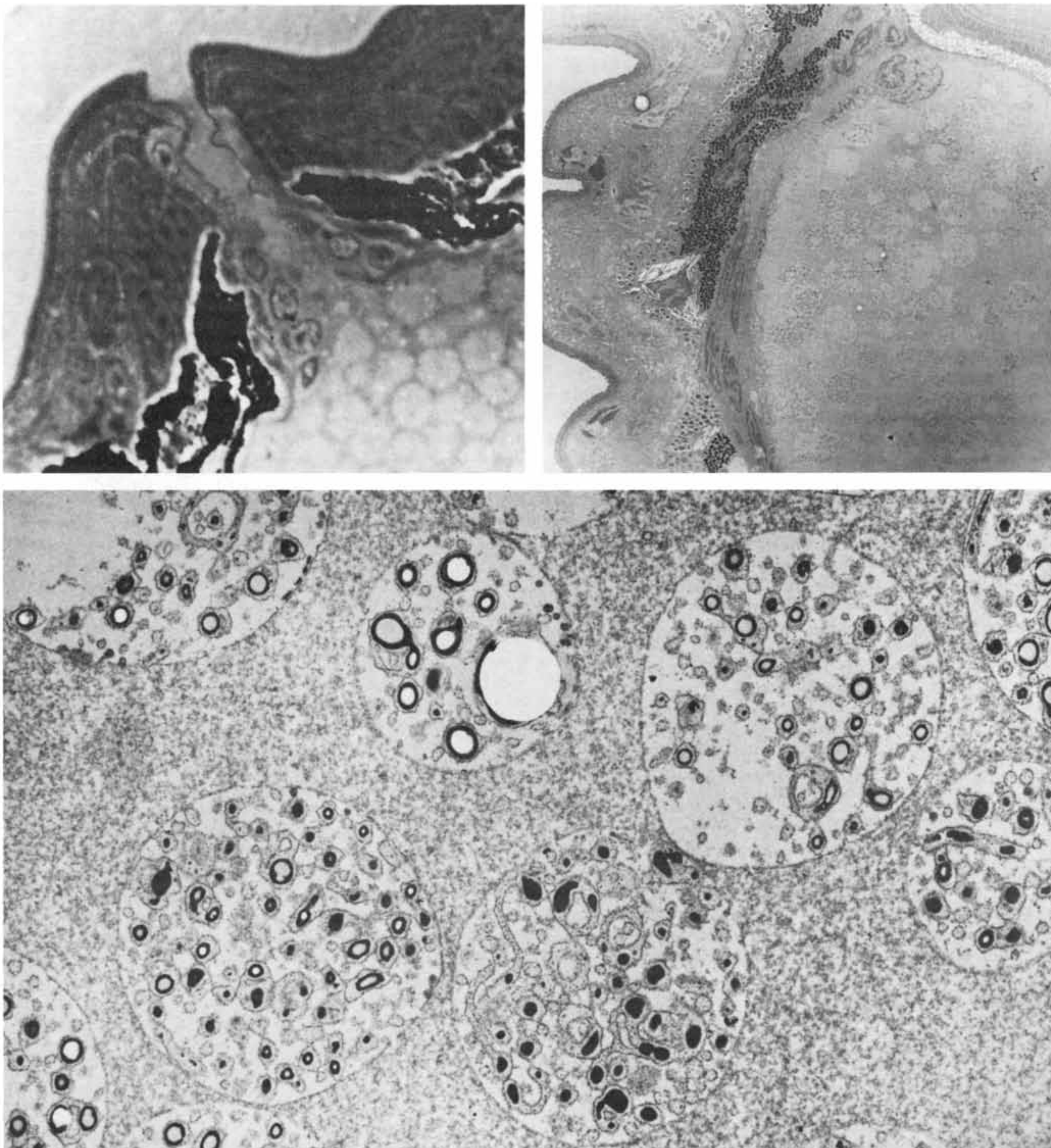
© 1983 NISSAN MOTOR CORPORATION IN USA

© 1983 SCIENTIFIC AMERICAN, INC

back of a living specimen of *P. terribilis.*

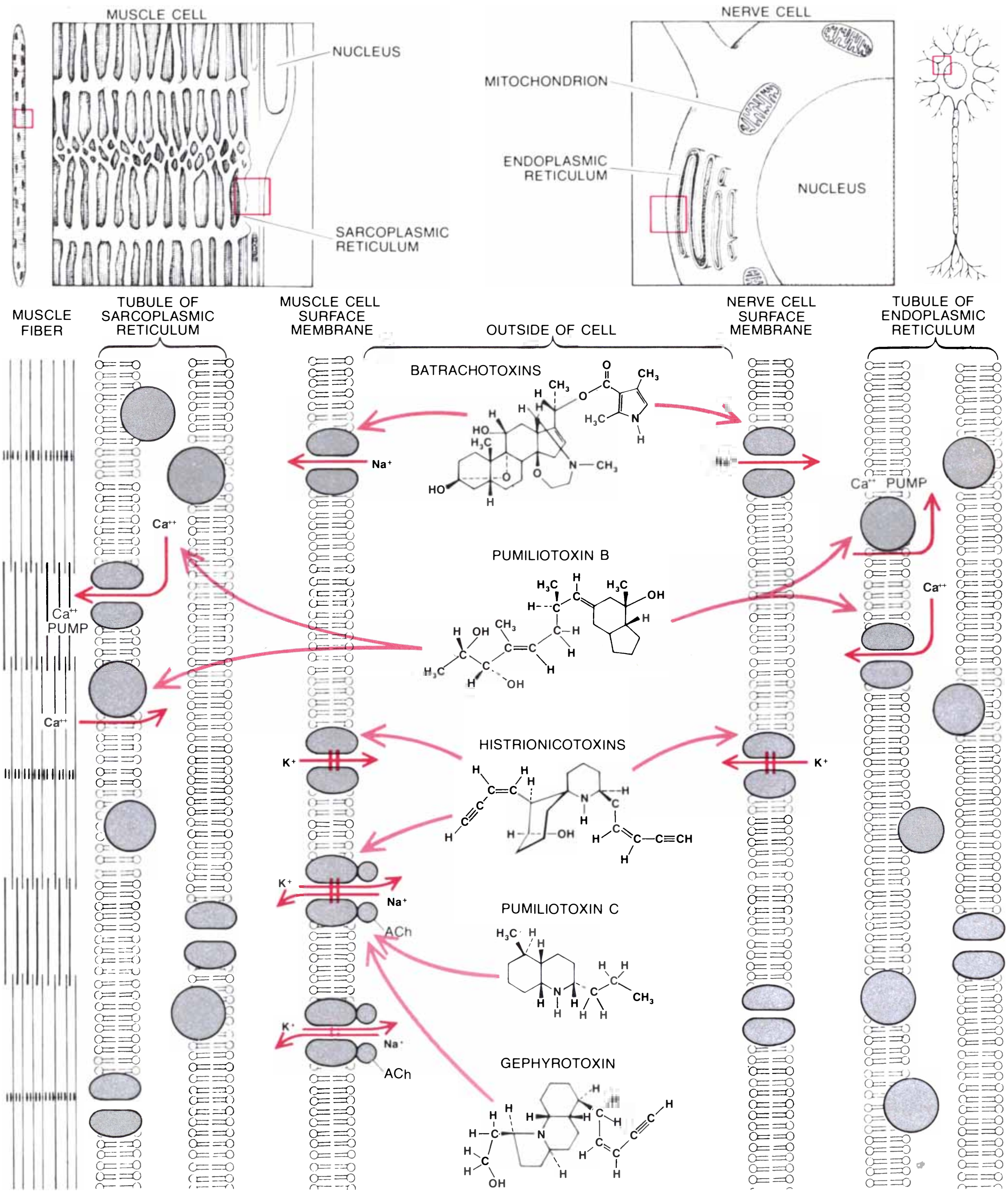
Today reliance on the blowgun is declining among the Chocó, mainly because of increased accessibility of firearms. Contrary to some literary accounts, dendrobatid frogs are not used for poisoning arrows, and their use for poisoning blowgun darts does not seem to have been widespread. This is understandable in view of the fact that the three most toxic dendrobatids are limited to relatively small areas of western Colombia. Two other species of *Phyllobates* are found in Central America, but they have much less toxin than their Colombian relatives and their range is outside the area where dart poisons are known to have been employed.

The skin organs that secrete the dendrobatid toxins are "granular" glands whose minute openings are scattered over the surface of the skin among the openings of the frog's mucous glands. Granular glands seem to be a primitive



"GRANULAR" GLANDS OF THE SKIN store and probably manufacture the toxic secretions of dendrobatid frogs. The photomicrograph at the top left shows a vertical section through the upper part of such a gland, magnified 900 times. The gland is filled with large, circular granules; the duct of the gland penetrates a layer of pigment cells and passes through the epidermis to open onto the surface of the skin. The electron micrograph at the top right shows a section through a gland, magnified 1,500 times. Vesicles in the granules may contain enzymes involved in the synthesis of the toxins. The electron micrograph at the bottom shows granules and the vesicles within them magnified 13,000 times. The micrograph at the top left shows a gland of *Dendrobates tricolor,* the other micrographs a gland of *D. auratus.*

© 1983 SCIENTIFIC AMERICAN, INC



MOLECULAR SITES OF ACTION for four kinds of alkaloid compounds from dendrobatid frogs are displayed schematically for muscle at the lower left and for nerve at the lower right. The morphological features of a muscle cell are shown at the upper left, those of a nerve cell at the upper right. In the middle are the structural formulas for the alkaloids. Batrachotoxin prevents the closing of sodium-ion channels in the surface membrane of both muscle and nerve cells. The influx of sodium ions electrically depolarizes the membrane and halts muscle and nerve function. Pumiliotoxin *B* appears to act in two ways. First, it facilitates the release of calcium ions from storage sites inside the muscle cell, thereby potentiating muscle contraction. Second, it inhibits the return of calcium ions to the storage sites, thereby prolonging the contraction. Pumiliotoxin *B* also affects the translocation of calcium in nerve cells. Histrionicotoxin blocks both the outward movement of potassium ions through potassium-ion channels of the surface membrane of muscle and nerve cells and the two-way exchange of sodium and potassium ions through complexes of ion channels and acetylcholine (ACh) receptors in the "end plate" between a nerve fiber and a muscle cell. The blockage of potassium-ion channels promotes the contraction of the muscle cells and prolongs the release of neurotransmitters by the nerve cells. The blockage of the ion-channel/acetylcholine-receptor complexes prevents the acetylcholine released from nerves from triggering muscle contraction. Both pumiliotoxin *C* and gephyrotoxin also block the movement of ions through the ion-channel/acetylcholine-receptor complexes, preventing acetylcholine from triggering muscle contraction.

© 1983 SCIENTIFIC AMERICAN, INC

characteristic common to all frogs. They were evidently a convenient evolutionary preadaptation in some groups of frogs for the synthesis, storage and release of various biologically active secretions.

The active principles of the dendrobatid skin secretions are alkaloids: nitrogenous ring compounds. Although such compounds are found more often in plants than they are in animals, dendrobatid frogs have proved to be a rich source of unique alkaloids. We have collected dendrobatids throughout their range in the tropical Americas for later analysis of their defensive secretions. This work has resulted in the discovery of a dozen new species of poison frogs and more than 200 new alkaloids, the latter representing at least five distinct classes of compounds. Chemical investigation of the frog alkaloids was initiated at the National Institute of Arthritis, Metabolism, and Digestive Diseases by Bernhard Witkop in the early 1960's. Over the past two decades major contributions to this research have been made by Isabella L. Karle, an X-ray crystallographer at the Naval Research Laboratory, and by Takashi Tokuyama, a chemist at Osaka University. The basis for the activity of each class of dendrobatid alkaloids is of considerable interest. The pharmacological investigation of these compounds has been pioneered by Edson X. Albuquerque of the University of Maryland School of Medicine.

The large majority of dendrobatid alkaloids include within their molecular structure a ring consisting of one nitrogen atom and five carbon atoms: a piperidine ring. Simple piperidine compounds are found among all subgroups of the genera *Dendrobates* and *Phyllobates.* From an evolutionary perspective this is evidence that the frogs in these genera form a separate lineage within the family Dendrobatidae. In the five species of the genus *Phyllobates,* however, the biosynthesis of piperidine alkaloids has been largely suppressed in favor of the synthesis of a new class of extraordinarily toxic alkaloids. They are the batrachotoxins (from the Greek *batrachos,* frog), which are the main basis of the poison on the Chocó blowgun darts. These complex alkaloids are structurally related to steroids but have many features hitherto unknown among natural compounds.

The batrachotoxins are among the most potent naturally occurring nonprotein toxins. They selectively increase the permeability of the outer membrane of nerve and muscle cells to sodium ions. This effect results in an irreversible electrical depolarization of the cells, causing in the heart arrhythmias, fibrillation and failure. The specific site at which batrachotoxin acts is associated with the channels that regulate the flow

Which of these languages would you like to speak?

Mark the one you want to speak in 2 or 3 months' time

Afrikaans	☐	French	☐	Japanese	☐
American English	☐	German	☐	Norwegian	☐
Arabic	☐	Greek (Modern)	☐	Polish	☐
Chinese	☐	Hebrew (Modern)	☐	Portuguese	☐
Czech	☐	Hindi	☐	Russian	☐
Danish	☐	Indonesian	☐	Spanish	☐
Dutch	☐	Irish	☐	Swedish	☐
Finnish	☐	Italian	☐	Welsh	☐

A Linguaphone Course makes you feel at home in almost any country. You'll never miss the meaning of conversations or be at a loss for words.

- It must work—over 4 million Linguaphone students in 88 countries speak a second language FLUENTLY.
- Proven learning success. You LISTEN to real conversations on cassettes...UNDERSTAND what you hear by following illustrated textbooks. HOLD CONVERSATIONS with the speakers.
- **You start speaking the very first lesson.** You learn at your own convenience.
- It's like having a private tutor.
- You get a complete, professional language program at little cost.
- You gain a good, working vocabulary.
- In just 2 to 3 months you can speak another language with complete confidence.
- You develop an authentic accent. Only native-born speakers are used.

Linguaphone The Language Masters

MONEY BACK GUARANTEE—28 DAY FREE TRIAL

World Language Courses, Inc. Dept. 227 313 Nolana Ave. McAllen, Texas 78501

FREE INFORMATION: Please mail me FREE information about learning the languages I have checked. FREE brochure.

Name (please print) ____________

Street ____________

City ____________ State ______ Zip ______

SCIENTIFIC AMERICAN

is now available to the blind and physically handicapped on cassette tapes.

All inquiries should be made directly to RECORDED PERIODICALS, Division of Volunteer Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

Free!

Edmund Scientific Catalog

4,000 Unique Items

Astronomy, Microscopy, Biofeedback, Weather, Alternate Energy, Binoculars, Optics, Magnets, Magnifiers, Tools, Unique Lighting, Lab Equipment, and much more. Over 4,000 unique and fascinating products. Send for our **FREE**, colorful Edmund Scientific Catalog...Today!

Rush me your free catalog!

Name ____________

Department ____________

Institution ____________

Address ____________

City ____________

State ______ Zip ______

Clip And Mail Coupon Today To:

Edmund Scientific, Dept. 8332 2110 Edscorp Bldg., Barrington, N.J. 08007

© 1983 SCIENTIFIC AMERICAN, INC

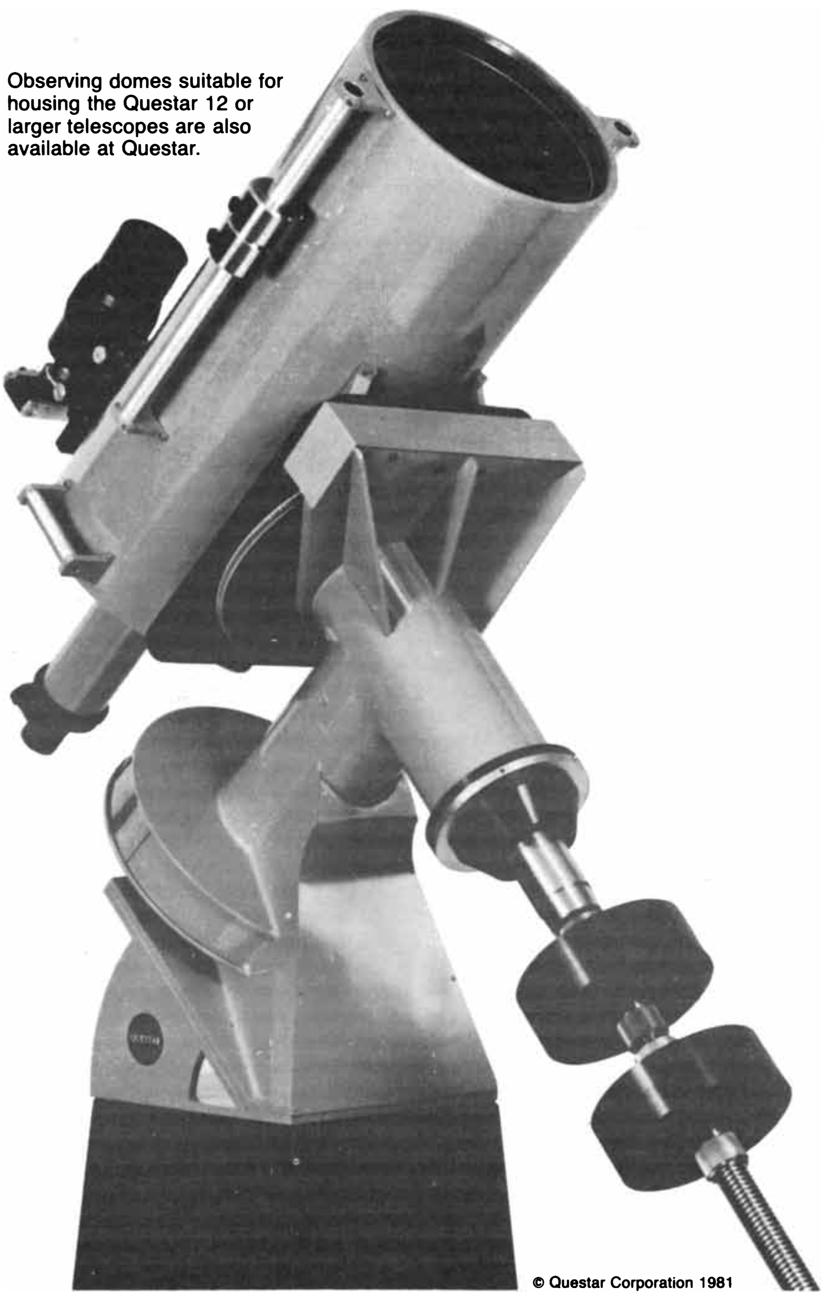
QUESTAR® 12 on the QUESTAR® MOUNT

The Questar 12, latest addition to the Questar family of fine optical instruments, now has its own Questar-designed mount. A German equatorial type, it is notable for its 360° continuous tracking in R.A. with precision tracking to better than 4 arc seconds. The Questar Mount is designed with over-sized components so that it can accommodate any Questar up to 18 inches. The standard mount shown is straightforward in design but can be modified so as to be compatible with more sophisticated tracking devices or other special equipment.

The Questar 12 is a superb instrument for the serious astronomer, for the university astronomy department or the engineer seeking sophisticated tracking and surveillance equipment for which Questar Corporation has a noted reputation.

Questar Corporation Box C, Dept. 20, New Hope, Pa. 18938 (215) 862-5277

Observing domes suitable for housing the Questar 12 or larger telescopes are also available at Questar.



© Questar Corporation 1981

of sodium ions through the cell membrane. Such sodium channels play a key part in the conduction of electrical impulses by nerve and muscle cells. The binding of batrachotoxin prevents the normal closing of the channels. As a result there is a massive influx of sodium ions and the cell is depolarized. Nerve cells can no longer transmit impulses and muscle cells remain in an activated, contracted state. The use of batrachotoxin in research has yielded important new knowledge of the function of sodium channels.

The other dendrobatid alkaloids are all much simpler in structure than batrachotoxin and are much less toxic, but they are no less interesting. Many of these piperidine alkaloids act on an ion channel in the "end plate" at the junction between a nerve fiber and a muscle cell. This channel is associated with a receptor for the neurotransmitter substance acetylcholine. The interaction of the neurotransmitter and its receptor opens the channel in the cell membrane through which both sodium and potassium ions can flow. If the resulting electrical impulse is large enough, it can trigger the opening of separate sodium channels in the membrane. With muscle cells this process causes the release of calcium ions inside the cell, which leads to contraction.

Histrionicotoxins, unusual spiropiperidine alkaloids first isolated from the frog *Dendrobates histrionicus,* interact with sites on the complex formed by the end-plate channel and the acetylcholine receptor, thereby blocking the passage of ions. Such blockage can prevent the transmission of signals from nerve to muscle, an end result similar to the one brought about by curare. Histrionicotoxins also block the passage of ions through separate potassium channels. When these channels are open, potassium ions flow out of both nerve and muscle cells, allowing them to return to their "resting" state after nerve transmission or muscle contraction. The blockage of such channels can lengthen the transmission of nerve messages and prolong muscle contraction.

Pumiliotoxin *B,* a unique indolizidine alkaloid first isolated from the frog *D. pumilio,* appears to affect the transport of calcium ions. In muscle cells an electrical impulse causes the release of calcium ions from internal storage sites. The ions then interact with intracellular proteins of muscle and cause contraction. Rather than blocking this process, pumiliotoxin *B* seems to facilitate it and also seems to delay the return of the calcium ions to their storage sites. As a result the force of contraction is augmented and contraction in both heart and skeletal muscle is prolonged. Such tonic actions on muscle may lead to clinical applications for this class of alkaloids.

Dendrobatid frogs with toxic or nox-

© 1983 SCIENTIFIC AMERICAN, INC

ious secretions are usually brightly colored. As is often the case with noxious animals, these species are mimicked by a few other species of frogs. An example is *Eleutherodactylus gaigeae,* which is found together with *Phyllobates lugubris* and *P. aurotaenia* and in the intervening territory, where an ancestral form of *Phyllobates* probably once lived. Unlike the dendrobatids it resembles, the nontoxic *E. gaigeae* is active at night. By day, however, it is sometimes found in the leaf litter where the *Phyllobates* species seek shelter when they are pursued. The different activity cycles may therefore be inconsequential as far as potential daytime predators are concerned.

The genus *Phyllobates* shows less variation than the genus *Dendrobates,* the species of which exhibit a remarkable variety of colors and patterns. Many *Dendrobates* species have primitive striped patterns, but others have evolved colorations that are uniformly bright or variously spotted, banded or mottled. Such variation may even occur among populations of the same species. It would be of much interest to know how such variability originates and whether it bears on the origin of new species. One species in particular, the Central American *D. pumilio,* seems most likely to yield answers to these basic questions.

D. pumilio is a small frog, usually found at elevations below 500 meters in the rain forest on the Caribbean side of Nicaragua, Costa Rica and western Panama. Variability is relatively low in Nicaragua and Costa Rica, where the frogs are usually red or reddish orange with black or bright blue hind legs. In Panama, however, variability is extremely high. It is associated with the geologic history of western Panama. Starting about 12,000 years ago the rising level of the sea created an archipelago of offshore islands and on the adjacent mainland a mixture of lowland swamp forest and rain forest. The archipelago had probably been created and reconnected to the mainland several times before. The populations of *D. pumilio* in the region show all the colors of the spectrum from red to blue; one population is even patterned solely in black and white. The different populations look like different species, and in fact we believe the complex includes at least two species. Nevertheless, most of the variation is found in what seems to be the single species *D. pumilio.*

The variation is not due simply to the independent evolution of populations isolated on islands; different parts of a single island may be inhabited by very different populations. There are also instances of intense variation within single populations. One example we have studied over a period of years is found at the northwestern end of Bastimentos Is-

Discover a small world with a big heart on the Polaris Baltic cruise.

Our Baltic/Russia cruise, stopping in Copenhagen, Stockholm, Helsinki and Leningrad will show you spectacular scenery, majestic castles and warm friendly people. And an equally engaging staff of lecturers will brief you on the region's history, culture, wildlife and heart.

The **M.S. Polaris*** offers only 1st class, outside cabins. Hearty Scandinavian food. And a chance to learn. Cruises leave from May through August. For more information call **800-223-5688** (in New York **212-751-2300**), see your travel agent, or write Dept. SA-2.

*Registered in Sweden

TO LEARN IS TO LIVE.

SALÉN LINDBLAD CRUISING
133 East 55th Street, New York, N.Y. 10022

Learn Spanish the Foreign Service way!

You can be fluent in Latin American Spanish. That's right, *fluent!* You can be speaking with all the confidence of a seasoned traveler in a matter of weeks. As well-versed in Spanish as if you worked for the U.S. Foreign Service Corps.

The Foreign Service Institute has carefully developed, tested, and refined its self-instructional audio cassette Spanish course to train the U.S. State Department's overseas personnel quickly, thoroughly, and effectively. Its clear, easy-to-follow approach and arrangement of material makes this a perfect method for successful self-instruction.

Now the FSI's complete program is available to *you.* We will send you the entire audio cassette series and instruction books to lead you step by step to fluency in weeks. You work when and where you can, at your own pace, following the taped lessons. You progress according to the schedule *you* establish.

The course uses native speakers exclusively so that you hear exactly how each word and sentence should sound. It won't be long before you start speaking and thinking in your new language. **This program has to work—our State Department depends on it!**

MONEY-BACK GUARANTEE: Try the FSI Basic or Intermediate course for three weeks. If you're not completely satisfied, return it for a full refund. **There's no risk and no obligation, so order today.**

SPECIAL OFFER: Purchase any language course from Pinnacle and get the **Toshiba KT-S3 Personal Stereo Cassette Player,** complete with headphones, FM Tuner Pack, and shoulder strap (nationally advertised for $139.00), for **only $89.95.**

French also available. Write us for details or call (212) 772-9202.

pinnacle **Dept. 219, 215 E. 79th St. New York, N.Y. 10021 (212) 772-9202**

TO ORDER BY PHONE—(Credit Card Orders Only Please)—Call us Toll Free: 1-800-528-6050 Ext. 1611
In AZ 1-800-352-0458 Ext. 1611

Please send me the following course(s) shipped in attractive library albums.

☐ **Programmatic Spanish—Vol. 1 Basic:** 11 Cassettes 16 hrs. manual and 464-page text. **$100**
☐ **Programmatic Spanish—Vol. 2 Intermediate:** 8 Cassettes 12 hrs. manual and 614-page text. **$90**
☐ **Both Volumes for Only $180.** ☐ **Toshiba KT-S3. $89.95** plus $4.00 shpg.
☐ Enclosed is my check or MO for $______ (In NY add sales tax.)
Charge my: ☐ MasterCard ☐ Visa No. ______ Expiration Date
☐ Amex ☐ Diners Club
Signature ______
Name (please print) ______
Address ______
City ______ State ______ Zip ______

© 1983 SCIENTIFIC AMERICAN, INC

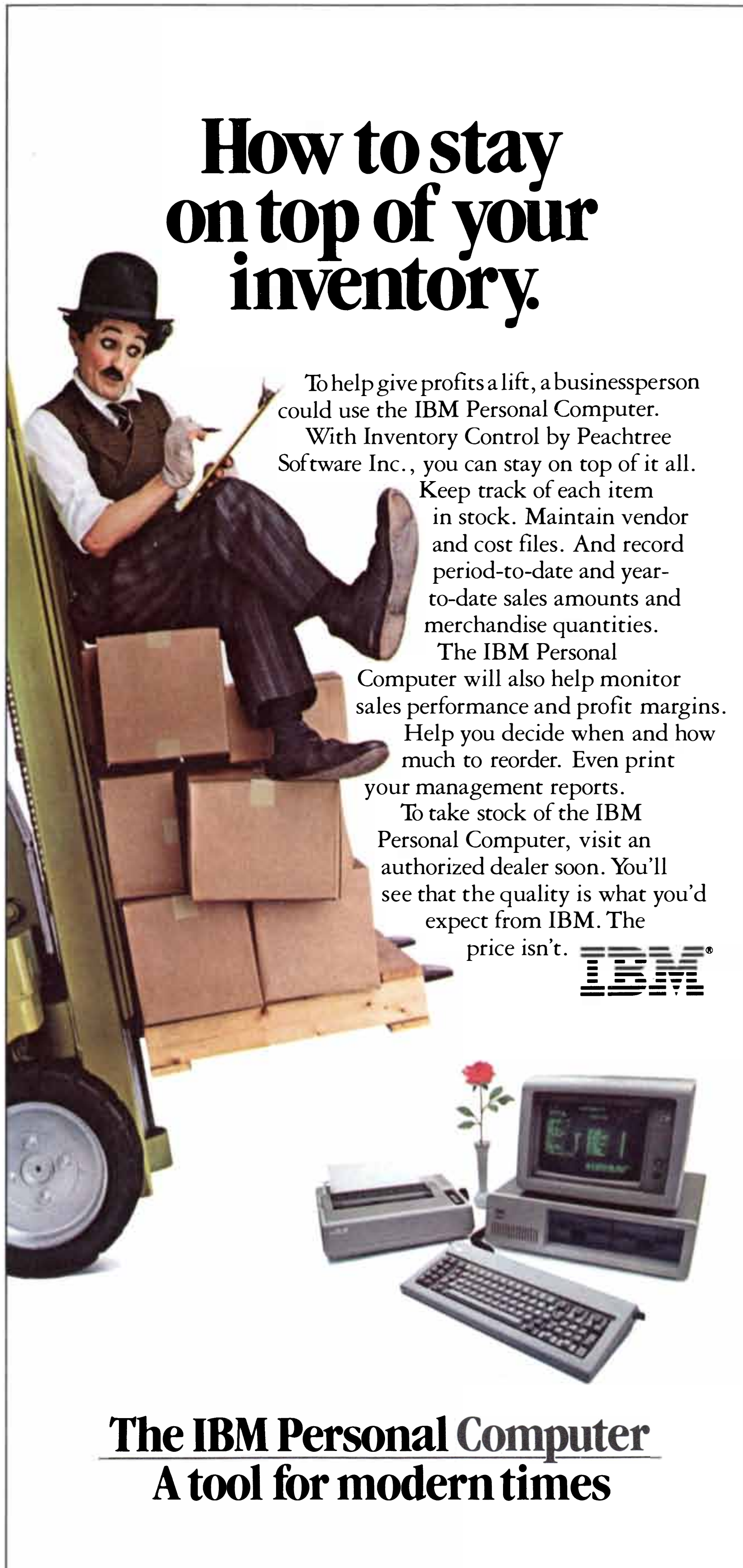
© 1983 SCIENTIFIC AMERICAN, INC

land. There the color of the frogs' backs varies from orange red through bronze to metallic green. In one beachfront forest orange red frogs consistently made up about 96 percent of the population samples we took each year, with only about 4 percent of the samples showing other colors. In a forest-shaded cacao grove away from the beach, however, the orange red frogs accounted for 84 percent or less of the samples, with bronze and green frogs now constituting as much as 32 percent of the yearly totals. These consistent differences between populations of a species in different habitats are an example of "balanced polymorphism." The phenomenon is probably the genetic basis for some of the striking differences among populations of *D. pumilio.*

Although very different populations of *D. pumilio* are usually separated by stretches of forest not inhabited by the frogs, we found areas on the mainland of Panama where nearly pure populations of red frogs were in contact with similarly pure populations of green ones. Frogs of the other color were rare in both populations. A few other rare color variants in the two populations resembled frogs of more distant populations. These findings further indicate that frog populations of different colors can arise from a common gene pool.

There is more to the *D. pumilio* problem than color. The frogs in the different populations also vary in size and behavior. Their usual habitat can be on the ground or in the trees; their response to potential predators ranges from secretive and wary to bold and apparently fearless. None of these traits seems to be correlated with the frogs' defensive skin secretions, the variations in which are equally remarkable. The *D. pumilio* populations of western Panama overall produce more than 80 alkaloids representing all major classes of piperidine toxins. The number of different alkaloids secreted by frogs of individual populations ranges from six to 24, and many of these compounds have not been detected in other dendrobatid frogs. *D. pumilio* is clearly a leading candidate among all vertebrate animals for the title "most variable species," but we are a long way from understanding all the factors that have made it so.

EIGHT FROGS representing different populations of *Dendrobates pumilio* are shown in dorsal and ventral position in the painting on the opposite page. These samples show only part of the extraordinary divergence among frogs of the *D. pumilio* population complex on the Caribbean side of western Panama. All the frogs are shown at the same size, but in life they vary in size as well as in color; they are also remarkably diversified in behavior and in the alkaloids of their defensive secretions.



For more information on where to buy the IBM Personal Computer, call **800-447-4700.** In Alaska or Hawaii, **800-447-0890.**

© 1983 SCIENTIFIC AMERICAN, INC

© 1983 SCIENTIFIC AMERICAN, INC

Fan Vaulting

In the 14th century English masons converted the pointed Gothic arch into an architectural shell of flowing curvature. Their aims were primarily aesthetic, but the vaults are soundly engineered

by Walter C. Leedy, Jr.

One of the central problems in architecture is how to enclose and articulate interior space. The problem becomes more critical as the size of the space increases, and it was particularly acute in a medieval cathedral or abbey. There the upward sweep of the walls strongly emphasized the ceiling, which was consequently expected to have a striking design. Furthermore, the ceiling was generally made of stone and was very heavy. The solution adopted in the most representative Gothic churches was to construct the ceiling as a series of vaults in which planar sheets made up of stone blocks ran between sharply pointed arches. The arches served as ribs bearing at least part of the load.

As medieval masons gained experience with ribbed vaults they found that construction could be made easier and the design more versatile by including additional ribs in the ceiling structure. Eventually the vaulting came to be a complicated lacework of ribs, which could be given many shapes. In England in the 14th century the proliferation of ribs led to the invention of an entirely new structure: the fan vault. In the fan vault the planar stone sheets of the ribbed vault were replaced by a rounded structure called a conoid that resembles a cone cut in half along its axis and erected with the pointed end down. Four conoids were joined to span each modular unit of the vaulting. On the curved surface of each conoid numerous ribs radiate upward and outward toward the peak of the ceiling. It is from this fanlike pattern of ribs that the style takes its name. Between the ribs is intricately carved stone tracery.

In a ribbed vault the arches were expected to bear the weight, but the conoid of a fan vault can act as a shell structure, in which the stresses are distributed fairly evenly throughout the stone fabric. The medieval builder had no way to analyze the stresses in a shell. Nevertheless, modern engineers have shown that the conoid of a fan vault is a very stable structure. Between about 1350 and 1540 English designers exploited this quality to construct more than 100 fan-vaulted ceilings spanning increasingly large spaces. Among the most famous are the ceiling of the chapel of King's College at the University of Cambridge and that of the Chapel of Henry VII in Westminster Abbey.

The elaborate ceiling of a fan-vaulted building provides a powerful visual experience. The rounded conoids flow into one another along the planes of the wall and give an impression of gracefully and continuously molded volumes of space, in contrast to the angularity of the ribbed vault. The intricate carving on the surface of the conoid emphasizes its volume and contributes to the impression of unity. The visual effect was probably foremost in the minds of the masons who built the fan vaults. It should not be forgotten, however, that underlying the aesthetic experience is the solution to a fundamental architectural problem arrived at by designers relying on practical experience in the absence of highly refined theoretical skills.

The development of the fan vault is closely connected to the physical properties of stone and the aesthetic preferences of the English masons and their patrons. Stone can withstand large compressive forces, but it has little tensile strength: it cannot readily withstand being pulled or stretched. The downward pull of gravity can cause substantial tensile stress in a horizontal beam supported at both ends. In a stone ceiling the tensile stress must somehow be converted into a compressive stress. In the pointed arch of the ribbed vault the conversion is accomplished by raising the arch stones above the horizontal plane defined by the points from which the arch springs. The force of gravity is thereby transformed into outward and downward compressive forces that are transmitted through the stones of the arch toward the walls.

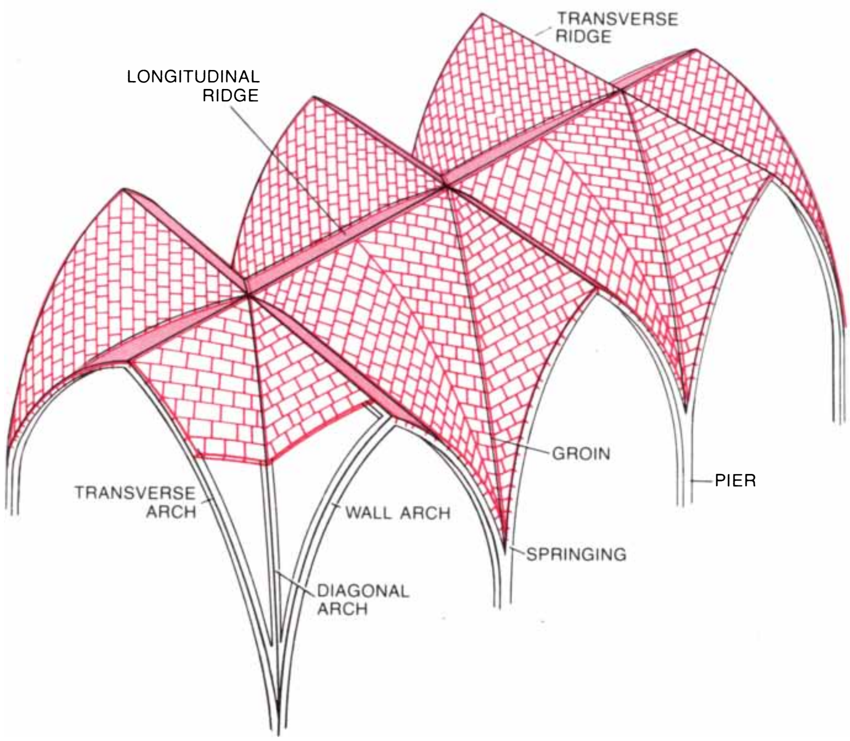
The modular unit of a vaulted ceiling is usually a rectangular compartment called a bay, whose long axis is perpendicular to the major axis of the hall. In a ribbed vault the bay generally includes three pairs of arches. The transverse arches cross the hall. The wall arches run along the wall at the short sides of the rectangle, most often with large clerestory windows in the wall under the arches. The opposite corners of the bay are connected by diagonal arches that intersect in the center of the rectangle. Tall piers placed at the springing, the point at the corner of the bay where the lower segments of three arches meet, help to support the vaulting high above the floor. The thrust from the arches is conveyed to the ground by the piers and by flying buttresses that meet the exterior of the wall near the springing.

In the construction of a rib-vaulted ceiling the arches were erected first, utilizing a complex wood scaffolding called centering as a form on which to lay the stones. When the six arches in each bay had been erected, they served as the framework for the web: the sheets of stone laid between the ribs.

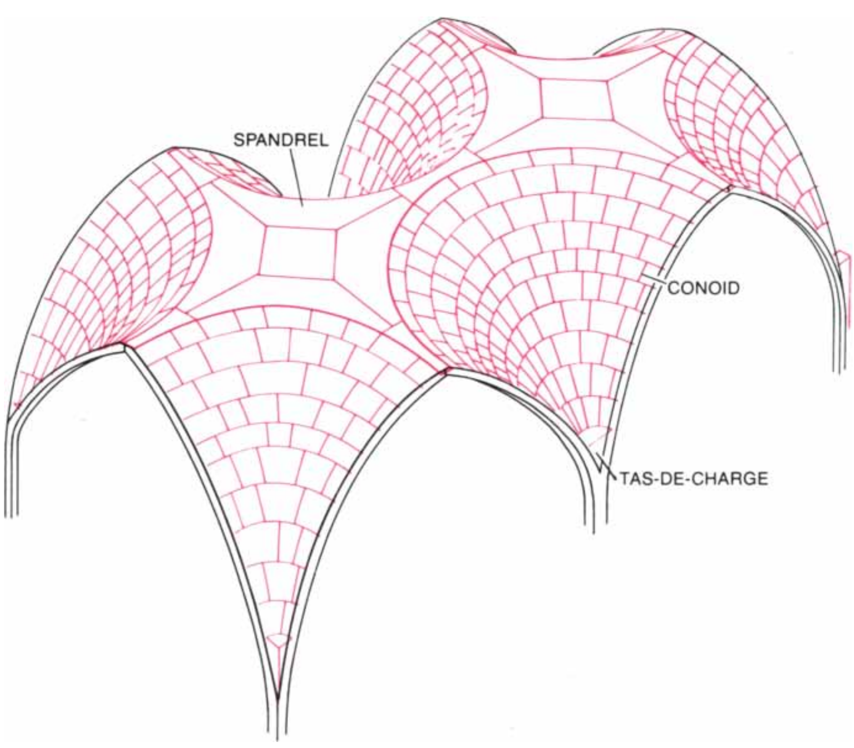
Ribbed vaulting, such as that in the cathedrals of Amiens and Rouen, has a strongly linear appearance, which is the basis of Gothic artistry. The effect was achieved by emphasizing the vaulting ribs as a series of frames with the web filling the space between them. Further-

KING'S COLLEGE CHAPEL at the University of Cambridge has one of the most beautiful examples of fan vaulting. The main vault of the chapel, shown here, was completed in 1515. The ceiling has the longest span of any known fan vaulting: the distance across the hall is 12.7 meters. The rounded intersecting surfaces are the main weight-bearing structures of the fan vault. Called conoids, they act as shells in which stresses are distributed fairly evenly. The carving, derived from the tracery of stained-glass windows, serves to unify the building's interior.

© 1983 SCIENTIFIC AMERICAN, INC



RIBBED VAULT, the classic Gothic ceiling design, was conceived by medieval masons as a frame structure in which pointed arches serve as weight-bearing ribs. Each of the modular units of the vaulting, called a bay, includes six arches arranged as a rectangle with intersecting diagonals. The load of the ceiling is converted into outward and downward thrusts that are conveyed in part through the arches to the piers and walls. The surface of the vault between the arches is filled with stone sheets called webs. In England the stones of the web were laid at an angle of approximately 45 degrees from the longitudinal or transverse ridge to the groin.



FAN VAULT developed from the ribbed vault in the 1350's. The fan-vault conoid is a surface of rotation, created by rotating a curve about a vertical axis. Viewed from the center of a bay, each conoid has a convex horizontal section and a concave vertical section. This form allows the conoid to act as a set of vertical and horizontal arches that convey thrust to the wall.

more, the large transverse arches draw attention to the division of the ceiling into bays. In contrast, although the fan vault is still composed of linear elements, the unity of the interior space is emphasized. The viewer's gaze is drawn along the ceiling by the repetitive and graceful form of the conoids; the distinction between bays generally has little visual importance.

Although the aesthetic principles of the two kinds of vaulting are quite different, the historical transition between them was not abrupt. On the contrary, the fan vault developed gradually out of the construction of ribbed vaults. Ribbed vaulting was common in England, France and other countries of continental Europe, whereas fan vaulting developed only in England; the question naturally arises of why the new form emerged only there.

The answer lies in the aesthetic values and building practices that prevailed in England in the Middle Ages. The comparison with France is instructive. In France the courses, or rows of stone, in the web were laid parallel to the sides of the rectangle covered by the vault, so that the courses run in horizontal rows from the bottom of the vault to the top. Each quadrant of the bay resembles a peaked roof; the ridge at the peak of the quadrant is called a transverse or a longitudinal ridge depending on the orientation of the quadrant. Along the groin, where the quadrants meet, the courses of the web were laid parallel to both ridges, forming right angles to the sides of the bay. For a while English masons followed the French precedent. Soon, however, they began to lay the courses from the ridge of the quadrant to the groin at an angle of about 45 degrees to both ridges. This style of coursing led in part to the conoid's being conceived of as a rounded structure.

English practice came to differ in other ways as well. In the French vault the web courses were laid over the top of the rib; the rib and the web were connected only by mortar. In England the upper edges of the rib were deeply notched and the web stones were set into the notches before the stones were cemented together. This practice, called rebating (or in modern English rabbeting), tended to make the rib and web a single assembly. Here is architectural evidence that the English mason saw the rib as a weight-bearing member.

Even the design of the arches in England differed from that in France. In France height, illumination and openness of structure were the most highly prized architectural qualities. To increase the area of the clerestory windows the wall arches were often stilted by raising the bottom of the arch above the springing on vertical wall shafts. Stilting had several significant conse-

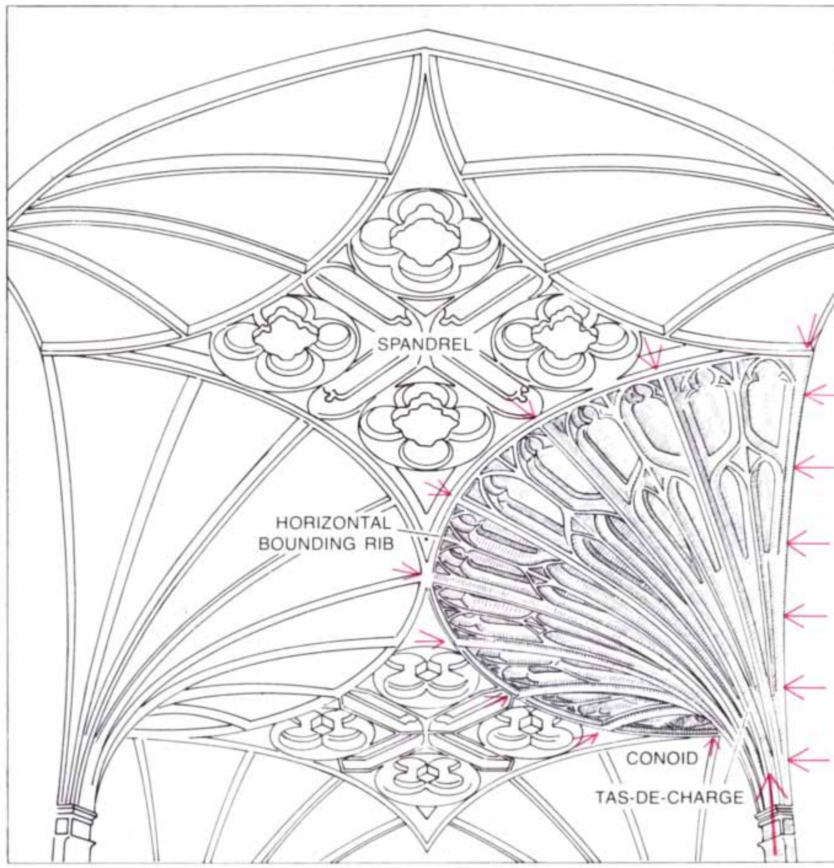
© 1983 SCIENTIFIC AMERICAN, INC

quences. The web near the window had to be warped to accommodate the shape of the arch. In addition the area of the wall between the clerestory windows was much reduced and the outward thrust from the vaulting was concentrated on a narrow strip of wall. As a result the flying buttresses had to be placed near their optimal position or the wall would collapse.

In England there was less concern with height and openness. The walls were generally lower and thicker. The clerestory windows were smaller and hence the wall arches did not have to be stilted. This increased the area of the wall across which the vaulting thrust was distributed and reduced the need for buttressing.

The ribbed vault spread through much of Europe in the 12th century. Medieval masons soon learned that the addition of two further types of ribs could increase the flexibility of the vaulting design. The first type, the tierceron rib, ran from the springing to the ridge of the vault, roughly bisecting the angle between the diagonal rib and either the transverse rib or the wall rib. The lierne rib was a shorter structural member inserted at an angle between the longer ribs.

English masons built numerous tierceron vaults. The tierceron design was well suited to the English style, in part because it made possible great flexibility in adjusting the curvature of the ribs. In contrast to architectural practice in France, where rib curvature was constrained by the striving for height and illumination, in England the tierceron ribs radiating from the springing could be formed into an appealing geometric pattern.



CONOID OF A FAN VAULT must be compressed along all its edges to be in equilibrium. The drawing shows the vaulting in the Church of St. Mary at North Leigh, Oxfordshire, which was finished in about 1440. The heavy spandrel panel between the conoids provides a compressive force along the horizontal bounding rib and serves as the keystone for the entire vault. The tas-de-charge built out from the wall at the corner of the bay supports the bottom of the conoid. The walls provide inward thrust to oppose the thrust transmitted outward through the conoids. Ribs carved on the surface are primarily aesthetic rather than structural elements.

In the high Gothic ribbed vault the volume defined by the webs intersecting along the groin in the corner of the bay was quite irregular and the shape was determined by other design considerations. As more ribs were added to the tierceron vault the volume became more regular and began to appear as a satisfying form worth designing for its own sake. At the same time the English coursing pattern began to yield an intriguing result. As I have mentioned, in England the web courses were laid at an angle from the longitudinal or transverse ridge to the groin. In tierceron vaulting the web courses in the corner of the bay began to approximate the form of regular, horizontal polygons. The polygons were concentric on the springing of the vault and increased in size toward the peak of the ceiling.

By the early part of the 14th century the main preconditions for the fan vault had been created in tierceron vaulting with concentric courses and rebated ribs. Two further technical advances were necessary for the fan vault to come into being. The first was the practice of carving a segment of the rib and the adjoining portion of the web from the same block of stone. The carved stones were then fitted together to form the vaulting much as blocks of ice are fitted to form an igloo. This kind of construction, which is termed jointed masonry, was employed in much fan vaulting.

The use of jointed masonry by English masons does not seem to have been motivated primarily by engineering considerations. On the contrary, it seems to have been largely the result of aesthetic concerns. English builders valued a high degree of carved articulation of the vaulting surface. Such articulation is much easier to achieve with large, closely fitted blocks than with ribs and small panels of stone inserted between them. As the size of the blocks was increased to provide a carving surface, the blocks tended naturally to assume a structural function.

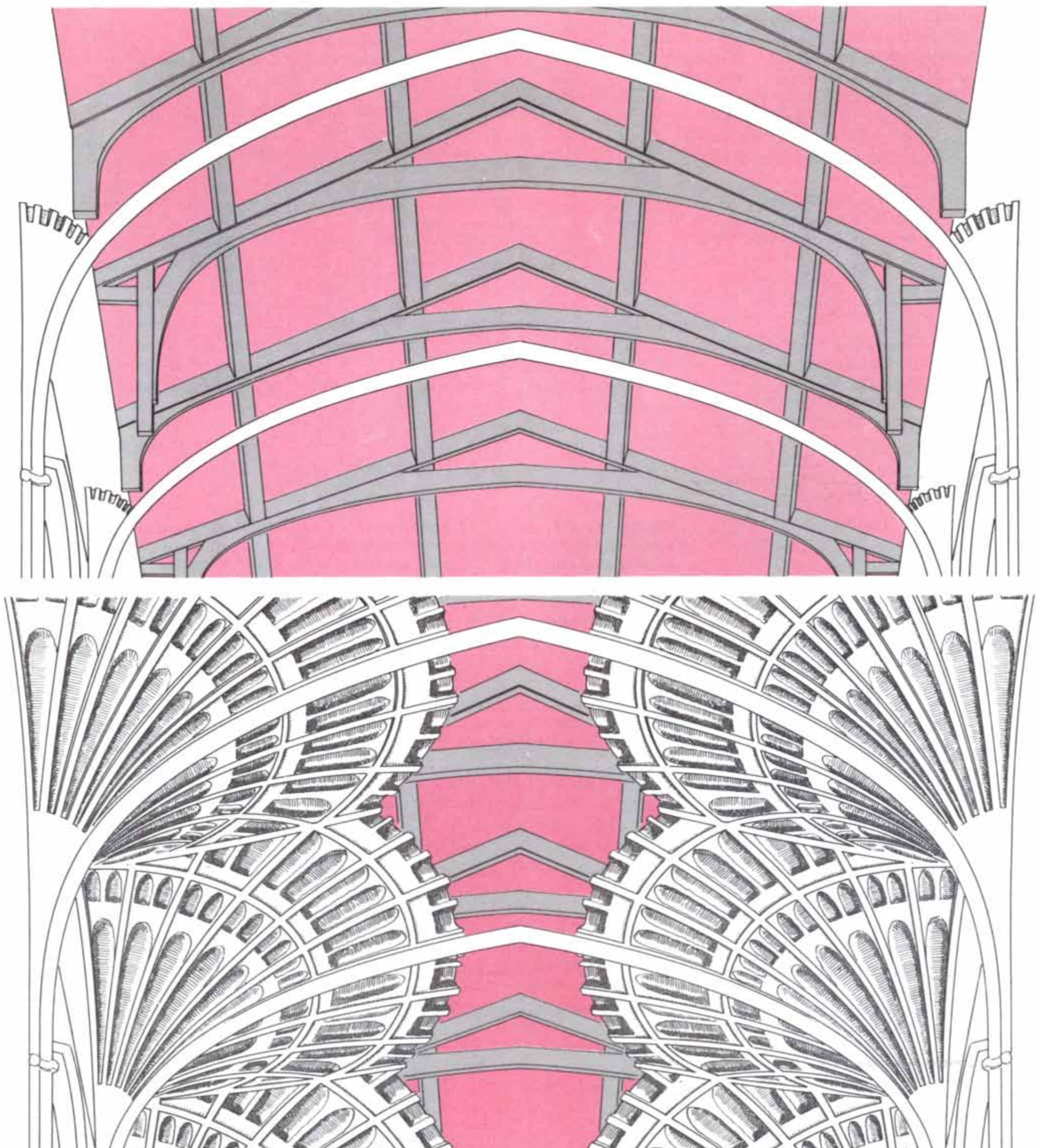
The introduction of jointed masonry led to another innovation: the major axis of the rectangular cross section of all the ribs was made perpendicular to the vaulting surface. In a ribbed vault the major axis was generally perpendicular to the floor. When jointed masonry is used in a fan vault, however, the joint between the portions of a rib on two different blocks can be made perfect only if the ribs are perpendicular to the vaulting surface. The proliferation of ribs and hence of intersections between ribs made this an increasingly important aesthetic consideration.

By the middle of the 14th century the major elements of the fan vault had evolved in England. According to most modern investigators, the defining characteristics of a fan vault are as follows: vaulting conoids with a regular geometric form; regularly spaced ribs, all with the same curvature; a distinct spandrel, or central ceiling panel; ribs perpendicular to the surface of the vaulting, and patterned surface carving. This set of properties seems to have appeared first in small square canopies in the tombs of great nobles. (The origin of the style may never be identified with certainty because of the destruction of many im-

© 1983 SCIENTIFIC AMERICAN, INC

portant sites during the reign of Henry VIII.) Because the tomb vaults are small, they are often referred to as toy, or decorative, fan vaults. The first structural fan vaulting could well be that in the Trinity Chapel of Tewkesbury Abbey in Gloucestershire; the vaulting was completed in about 1380. The first large structural fan vaulting and one of the most significant early examples of the form is the vaulting in the cloister of Gloucester Cathedral, which was begun in the second half of the 14th century.

An essential feature of the fan vault is that the vaulting conoid is a surface of rotation, that is, a form produced by rotating a curved line about a vertical axis. The axis is the corner of the bay and the curved line is the rib. One consequence of this form is that the conoid has a horizontal section that is a circle or a portion

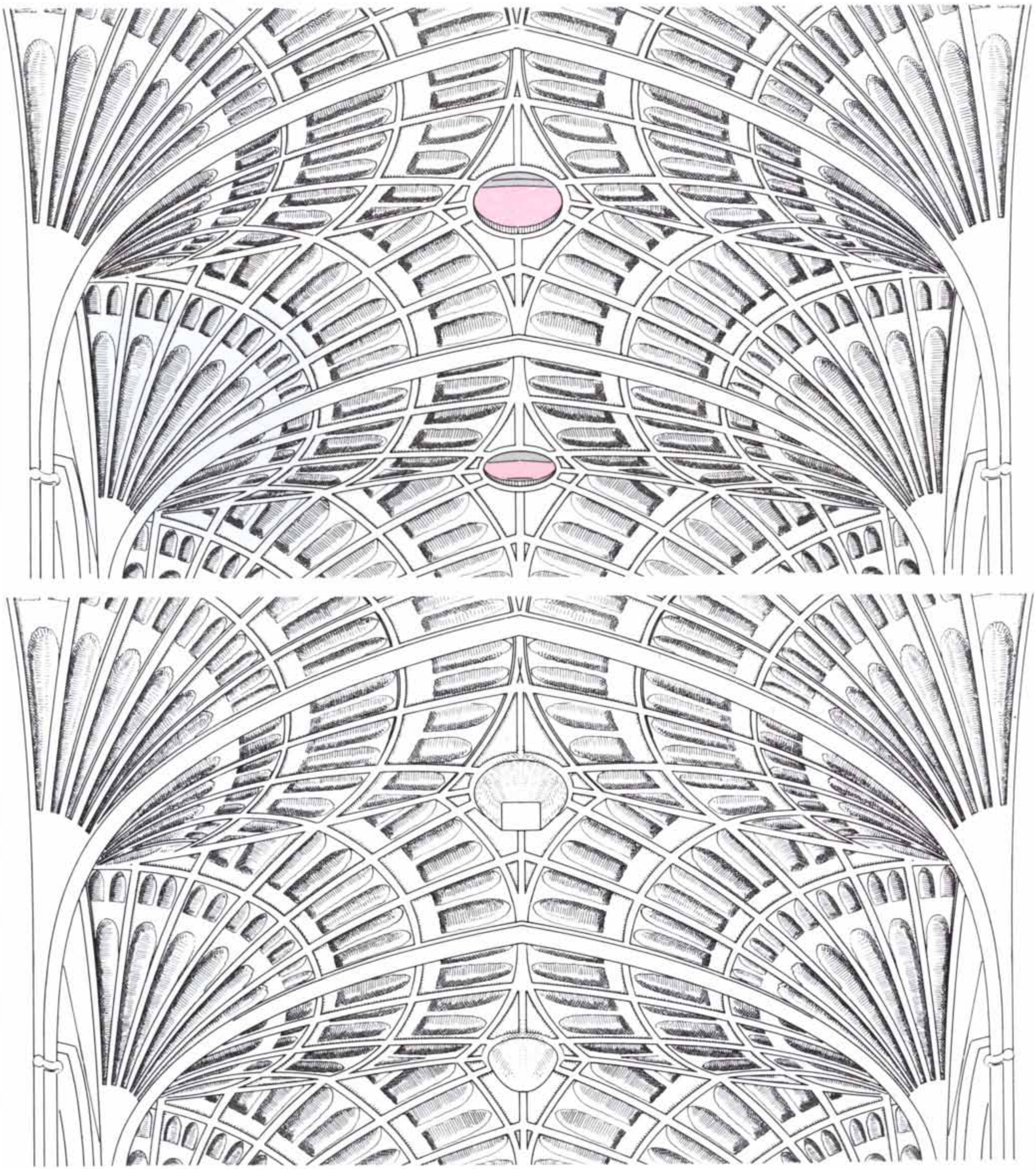


CONSTRUCTION OF VAULTING in King's College Chapel was accomplished in four stages. The stone blocks were carved in workshops next to the building site. It seems likely they were put in place with their decorative surfaces only partially finished. Wood centering, or scaffolding, was utilized as a form for the vaulting during construction. After the walls and roof were up the large transverse arches that divide the ceiling into bays were erected (*upper left*). The conoids were then laid, proceeding inward from the transverse arch-

© 1983 SCIENTIFIC AMERICAN, INC

of a circle at every vertical level. The circle is convex with respect to the center of the bay. The rib, however, is concave with respect to the center of the bay. Thus at every point on the surface of the conoid a convex horizontal curve intersects a concave vertical curve.

Since the conoid extending from the corner of the bay is circular, fan vaulting is best suited to a square bay. (Fan vaults were also built in rectangular modules, but ingenuity was required to adapt the form to the bay.) In a square bay the conoid has a horizontal section that is a quarter of a circle. At the top of the ceiling is an approximately diamond-shaped space formed by the intersection of the four circular conoids. This space is filled by the spandrel. At the curved boundary between the spandrel and the conoid the builder often inserted



es toward the center of the bay (*lower left*). The conoids were assembled one horizontal level at a time. When the horizontal arches of adjoining conoids met in the center of the bay, the large transverse ridge stone that serves as the keystone for both arches was put in position. The stones of the longitudinal ridge were then laid (*upper right*). Finally, the large, heavy boss at the center of each spandrel was dropped in place from above (*lower right*). After the centering was taken down any final carving of the ribs and tracery was done.

© 1983 SCIENTIFIC AMERICAN, INC

Physician, did you miss any of these significant developments in medical science?

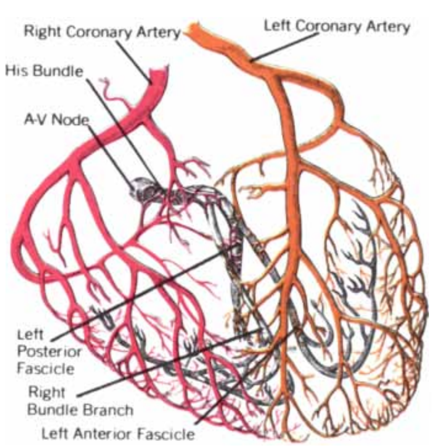
• ***Campylobacter fetus* subsp. *jejuni* is associated with a colitis that can clinically and sigmoidoscopically resemble acute idiopathic ulcerative colitis. Stool cultures are in order for *C. fetus* before beginning nonspecific anti-inflammatory therapy.**

• **Coumarin derivatives cross the placenta. A recent study shows that the consequences for the fetus can be severe. These include embryopathy, stillbirth, and premature delivery.**

• **Nonsteroidal anti-inflammatory drugs may produce a marked reduction in glomerular filtration rate; with termination of the drug, GFR returns to normal.**

• **The first documented incident of indigenous transmission of dengue in the continental United States since 1945 has been reported in Brownsville, Texas.**

IF THESE ITEMS are familiar you must be a prodigiously energetic or prodigiously lucky reader. With 2,000 or more journals published each year, information that significantly affects patient management all too easily slips by. Textbooks are out-of-date before they are published.



Branches of the right and left coronary arteries supply blood to the A-V node and intraventricular conduction system.

SCIENTIFIC AMERICAN *Medicine* is lucidly illustrated with drawings and photographs. Some examples are seen here and on the facing page.

SCIENTIFIC AMERICAN *Medicine* is the busy clinician's answer to this problem.

Because its authors update SCIENTIFIC AMERICAN *Medicine* every month, it is always current. Because the new information appears in a single source, it is there when you need it.

This 2,500-page, innovative union of publishing and electronic technology is the work of leading scholar-practitioners from Harvard and Stanford. The editors are Edward Rubenstein, M.D., F.A.C.P., and Daniel Federman, M.D., F.A.C.P.

Each month as authors update their contributions, revisions are entered on the magnetic tape on which the text and index are stored. The tape drives high-speed phototypesetting equipment so that subscribers receive about eight new chapters and a new index every four weeks; a bulletin highlights new developments.

New material replaces old material in the living text, so that the information is there — up-to-date, at your fingertips.

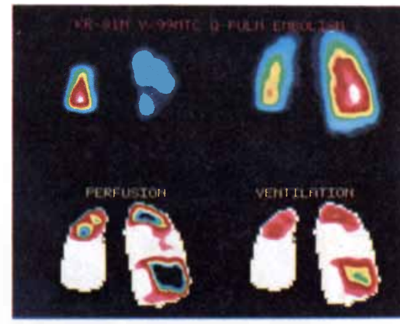
A CME program of eight patient management problems offered over a 12-month period is available at no extra cost. As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions. This program has been reviewed and is acceptable for 32 prescribed hours by the American Academy of Family Physicians. This program has been approved by the American College of Emergency Physicians for 32 hours of ACEP Category 1 credit.

Trial Offer

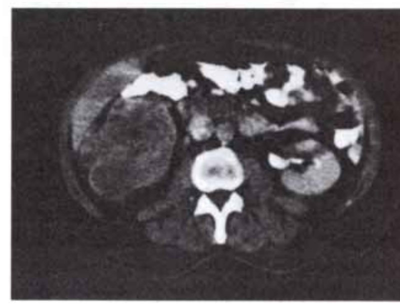
We invite you to try SCIENTIFIC AMERICAN *Medicine* — for two months at no cost. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, we will bill you for $220 for the full 12 months (renewal is currently $170); otherwise return the two volumes.

Please mail the coupon today and let us take the hassle out of keeping up.

© 1983 SCIENTIFIC AMERICAN, INC



Computerized scintigraphy reveals pulmonary thromboembolism.



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

THE DISTINGUISHED AUTHORS AND THE FIFTEEN SECTIONS OF SCIENTIFIC AMERICAN ***MEDICINE***

1. Cardiovascular Medicine

Edgar Haber, M.D., Harvard Medical School and Massachusetts General Hospital

E. William Hancock, M.D., F.A.C.P., Stanford University School of Medicine

Roman W. DeSanctis, M.D., Harvard Medical School and Massachusetts General Hospital

Adolph M. Hutter, Jr., M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital

Eve Elizabeth Slater, M.D., Harvard Medical School and Massachusetts General Hospital

2. Dermatology

Eugene M. Farber, M.D., Stanford University School of Medicine

Elizabeth A. Abel, M.D., Stanford University School of Medicine

3. Endocrinology

Daniel D. Federman, M.D., F.A.C.P., Harvard Medical School

4. Gastroenterology

Gary M. Gray, M.D., Stanford University School of Medicine
Peter B. Gregory, M.D., Stanford University School of Medicine
John Austin Collins, M.D., Stanford University School of Medicine
Douglas Wilmore, M.D., Harvard Medical School and Brigham and Women's Hospital

5. Hematology

Stanley L. Schrier, M.D., Stanford University School of Medicine

6. Immunology

John David, M.D., Harvard Medical School and Harvard School of Public Health

7. Infectious Disease

Thomas C. Merigan, M.D., Stanford University School of Medicine
Morton N. Swartz, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital
Cyrus C. Hopkins, M.D., Harvard Medical School and Massachusetts General Hospital
Adolf W. Karchmer, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital
Robert H. Rubin, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital
Harvey B. Simon, M.D., F.A.C.P., Harvard Medical School and Massachusetts General Hospital
Peter F. Weller, M.D., Harvard Medical School; Beth Israel Hospital, Boston; and Brigham and Women's Hospital

8. Intensive and Emergency Care

Edward Rubenstein, M.D., F.A.C.P., Stanford University School of Medicine

9. Metabolism

George F. Cahill, Jr., M.D., Harvard Medical School, Howard Hughes Medical Institute; and Brigham and Women's Hospital

10. Nephrology

Roy H. Maffly, M.D., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center
William Bennett, M.D., Health Sciences University, Portland, Oregon

11. Neurology

Robert W. P. Cutler, M.D., Stanford University School of Medicine

12. Oncology

Saul A. Rosenberg, M.D., F.A.C.P., Stanford University School of Medicine

13. Psychiatry

Ned H. Cassem, M.D., Harvard Medical School and Massachusetts General Hospital

14. Respiratory Medicine

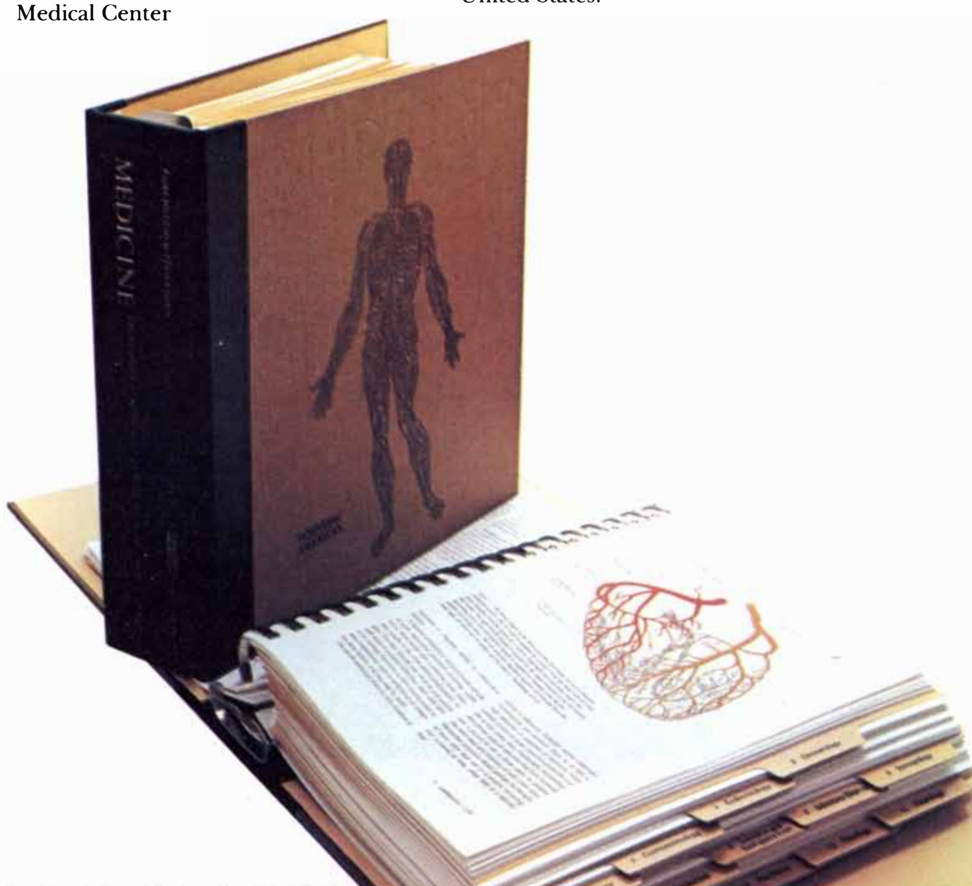
Eugene D. Robin, M.D., F.A.C.P., Stanford University School of Medicine

15. Rheumatology

Stephen M. Krane, M.D., Harvard Medical School and Massachusetts General Hospital
Dwight R. Robinson, M.D., Harvard Medical School and Massachusetts General Hospital
Andrei Calin, M.D., M.A., M.R.C.P., Stanford University School of Medicine and Palo Alto Veterans Administration Medical Center

Order by Phone

You can order SCIENTIFIC AMERICAN *Medicine* by telephone. **Please call this toll-free number: 1-800-345-8112 (in Pennsylvania, call 1-800-662-2444);** you will be billed after your subscription begins. Toll-free calls are acceptable only for orders placed in the continental United States.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, N.Y. 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine*. On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the price of $220 for first year of service is tax deductible, as is current renewal at $170. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund*.

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

Expiration Date______ Account Number______
*Please add sales tax for California, Illinois, Massachusetts, Michigan and New York

Name______

MD Specialty______

Address______

City______ State______ Zip______

Signature______

Please ... ents must be in U.S. dollars. CME is available outside the U.S. and its posses... ed unless airmail delivery is prepaid.

© 1983 SCIENTIFIC AMERICAN, INC

a horizontal bounding rib to divide the two structures visually.

The medieval mason was no formalist. He felt free to alter the elements of the fan vault in order to accommodate the design requirements or his own taste. For example, in the chapel of King's College the vertical ribs appear to be equally spaced, but actually there are small discrepancies resulting from the fact that the decision to fan-vault the ceiling was made after the building had been partially erected. Sherborne Abbey in Dorsetshire has beautiful examples of fan vaulting that were quite influential among 15th-century designers. The conoids in the chancel (the area that includes the altar) have horizontal sections that are polygonal rather than circular. Moreover, the conoids are constructed of separate ribs and small flat panels rather than jointed masonry. Nevertheless, in both King's College and Sherborne Abbey the ceilings must be described as fan vaulting because of their overall appearance, in spite of the deviations from the definition of an idealized fan vault.

The construction of fan vaults can be divided into three periods beginning with the construction of toy fan vaults in tomb canopies. The first period ended with the completion of the cloister of Gloucester Cathedral in 1412. Between 1412 and 1430 no major fan vaults were built. This was a time of labor shortage, high taxes and economic depression; fan vaulting was costly and it is likely that no individual or institution could muster the capital needed to undertake construction. The design of the Sherborne Abbey chancel vaulting in the late 1430's marks the beginning of the second period, which lasted until about 1475. In the third period, from 1475 to about 1540, many of the largest and most important fan-vaulted ceilings were constructed.

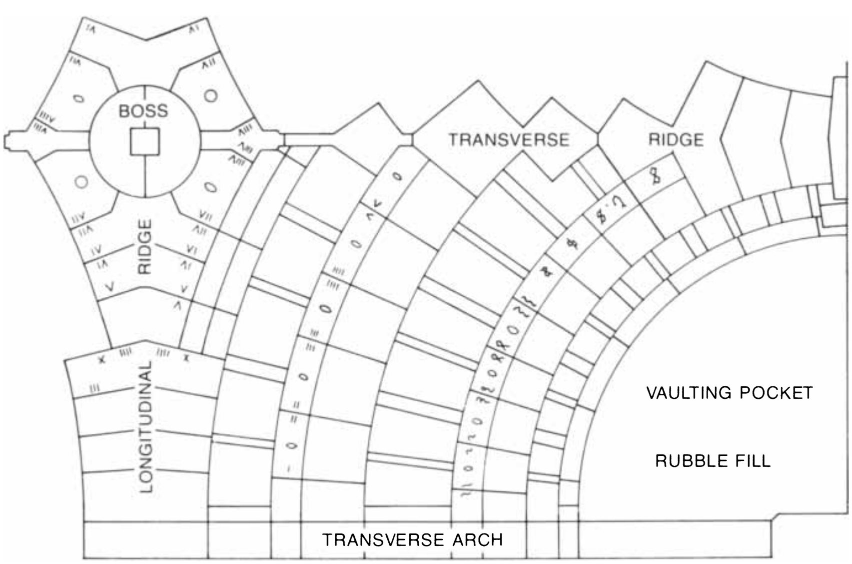
The overwhelming majority of fan vaults were built in ecclesiastical buildings, and a considerable number of these were constructed in chantry chapels. A chantry was a fund established to pay for masses for the soul of the founder; the chantry chapel was where the masses were said. Most of the early chantry chapels were endowed by noblemen and were intended as much to call attention to the greatness of the founder as to ensure the safety of his soul. The fan vault with its striking visual form and intricately carved tracery was well suited to the purpose.

In the second half of the 15th century the taste for fan vaulting spread to the middle class and to the King. St. George's Chapel at Windsor Castle, which was a royal institution, has fan vaults dating from the 1480's. Shortly after 1500 work began on the most magnificent of all fan vaulting: the ceilings in the Chapel of Henry VII at Westminster Abbey, which Henry conceived as a huge chantry chapel for the Tudor dynasty. Hence the church, the nobility, the middle class and the crown all contributed to the diffusion of the new architectural form.

The spread of fan vaulting was accompanied by an increase in the scale of the projects. The vaulting completed about 1380 in Trinity Chapel at Tewkesbury Abbey spans 1.7 meters across the hall. The main vault of the King's College chapel, which was completed in 1515, spans 12.7 meters. This is the longest span of any fan vaulting. One reason it took English builders more than 150 years to increase the span of the vaulting to its maximum is that they had little conception of how a fan vault works.

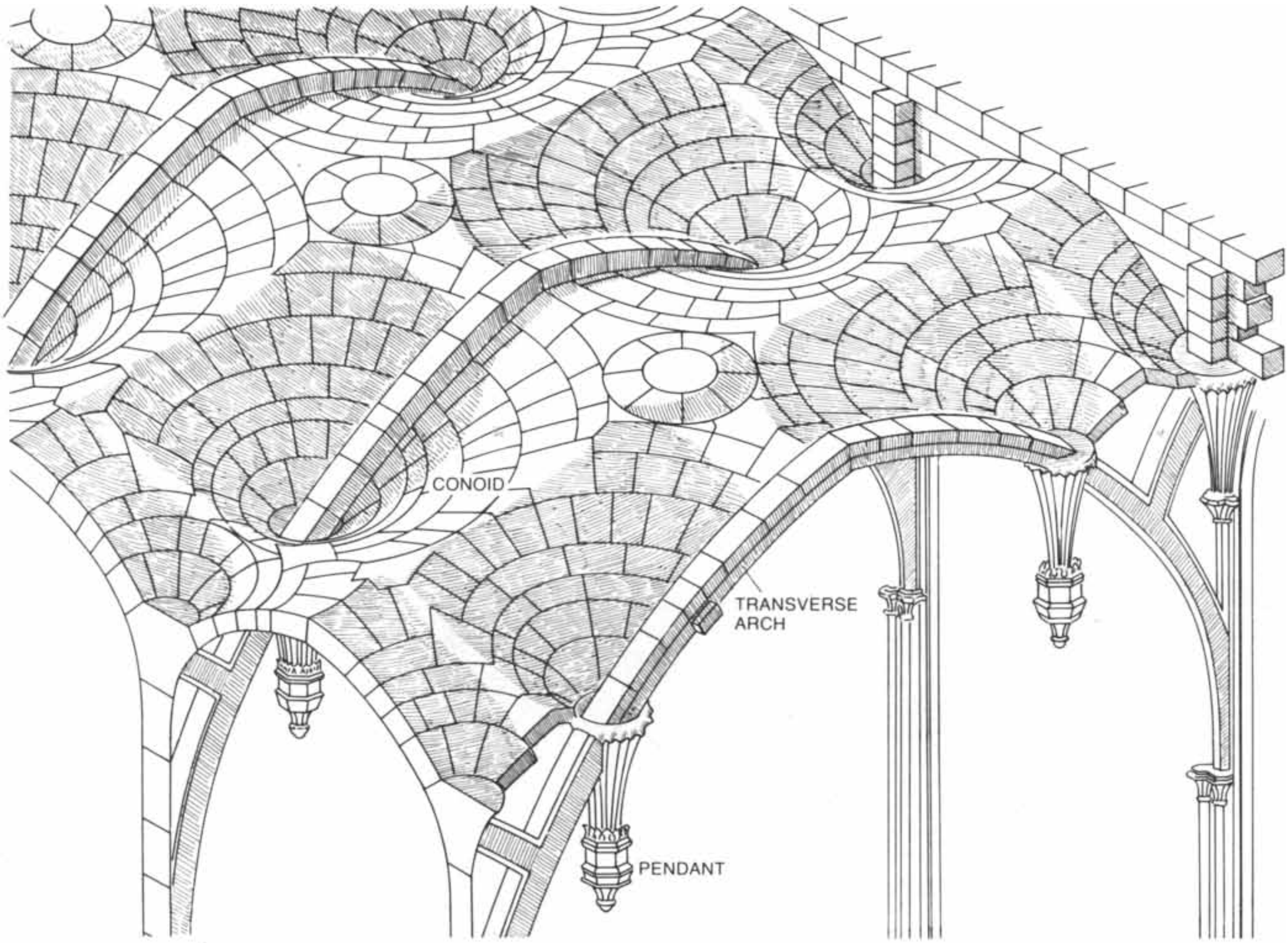
Twentieth-century engineers utilizing advanced mathematical tools have shown that for the conoid of a fan vault to be in equilibrium it must be supported along all its edges. A substantial weight must apply a compressive load at the upper edge along the horizontal bounding rib that separates the conoid from the spandrel. The load is provided by the spandrel itself, which is a very heavy stone plate. The large bosses, or raised decorative stones, that constitute the center of the spandrels in the King's College chapel weigh some 1,400 kilograms each.

The spandrel serves as the keystone of the vertical arches in the conoid. As in the ribbed arch, the thrusts are directed outward and downward. The downward thrusts are conveyed to the bottom of the conoid, where the structure rests



PLACEMENT MARKS on the stones for the vaulting of King's College Chapel enabled the stone setters on the building site to complete the ceiling by merely following the marks. The symbols were carved in the workshops where the vaulting was prefabricated. The large circular mark at the center of the stone indicates the quadrant of the bay in which the stone was placed. The marks at the ends of the stone next to the joints are the Arabic numerals 3 and 4. They indicate the horizontal level in the conoid where the stone was to go, the position in that level and the timing of insertion. The placement marks can still be seen on the upper surface of the vaulting stones, as is indicated by the plan view of one quadrant of a bay in the main vault.

© 1983 SCIENTIFIC AMERICAN, INC



CHAPEL OF HENRY VII in Westminster Abbey has the most magnificent of all fan vaulting. Work on the vaulting was begun in about 1500. The ingenious design of the ceiling combines frame structures and shell structures. Substantial transverse arches support large pendants near the wall of the bay. The conoids are built over the pendants. At the pendants the transverse arches run up through the vaulting and along its upper surface, so that the arches are not visible in the central part of the ceiling. Most of the weight is taken up by the arch where it passes through the conoid at the pendant. The arch functions as a frame member and the conoid as a shell structure.

against the wall. At this point the conoid is supported by the tas-de-charge, a projection built out from the wall in the corner of the bay. As noted above, the conoid is composed of concentric courses radiating upward from the springing, with each course having the form of an arch whose alignment is approximately horizontal. The outward thrust is conveyed through the horizontal arches to the wall, which provides an opposing thrust. Thus the conoid is compressed between the spandrel, the walls and the tas-de-charge.

The stresses in the conoid tend to be distributed fairly evenly rather than concentrated in the ribs. The specific way the stresses are distributed has important implications for the stability of the conoid, which, as I have noted, is a shell structure. An architectural shell is defined as a stressed structure much thinner than it is broad. The thrust acting on each point of the shell can be represented by a vector with components directed outward and downward; the set of all such vectors forms an imaginary surface called the thrust surface. Mathematical analysis shows that for the conoid to be in equilibrium the thrust surface must lie within the physical surfaces of the conoid shell. If the thrust surface emerges from the surfaces of the conoid, the structure is likely to collapse.

The builders of fan vaults had no way of knowing this fundamental fact. The analysis of stresses in a shell was far beyond their theoretical capacities. On the basis of long experience, however, the masons developed several techniques that greatly increased the stability of the vaulting. The narrow volume at the base of the conoid above the tas-de-charge is termed the vaulting pocket. In many fan-vaulted ceilings the pocket was filled to a height of about a meter with a solid rubble made up of stone chips and cement. In the chapel of King's College the rubble reaches precisely the same height in each conoid, indicating that it was a planned element of the structure rather than a casual afterthought.

The presence of the rubble fill has three significant consequences. First, the thrust surface extends into the rubble fill, which conveys the thrust to the wall. In this way the thrust is distributed over a much larger area of the wall than it would be if the shell alone transmitted it to the wall. Second, the rubble acts as a weight cantilevered over the floor. The weight opposes the outward thrust of the conoid reaching the walls in the corner of the bay and thereby further reduces the stress put on the walls.

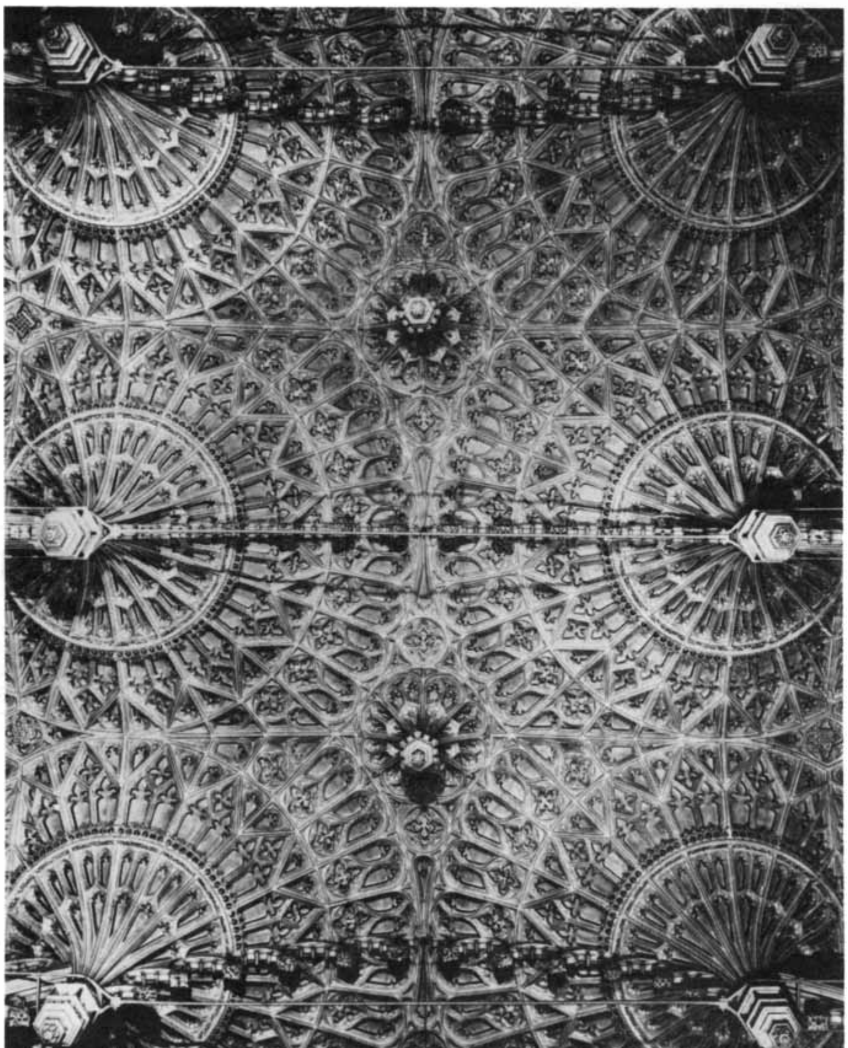
Most significant for the stability of the conoid, the rubble fill reduces the span of the vaulting that functions as a shell structure. The rubble cannot act as a shell because it is a substantial three-dimensional form rather than a thin sheet. Since the vaulting thrust passes into the rubble, the lower part of the conoid does not act as a shell. Jacques Heyman of the University of Cambridge has shown that the part of the vaulting that func-

© 1983 SCIENTIFIC AMERICAN, INC

tions as a shell is the part between the top of the rubble fill and the lower edge of the spandrel. This distance can be quite short in relation to the entire span of the vaulting. Reducing the length of the shell increases the ratio of thickness to length. The shorter and thicker the shell, the greater the probability that the thrust surface lies between the surfaces of the conoid.

Thus without significant theoretical knowledge the mason managed to make the fan vault quite stable. If stability can be ensured, the only other major risk to be considered is compressive failure, that is, the crushing of the building material. Since stone has tremendous compressive strength, however, it is not likely that the vaulting will fail in compression no matter how large the vaults. Therefore large fan vaults can be constructed on the same principles as smaller ones. Nevertheless, it took many decades for builders to understand that the stability of the fan vault is not compromised by its size. By the end of the 15th century this fact had been grasped and quite audacious ceiling designs were being attempted.

The construction of the largest fan vaults entailed some intriguing building techniques. The stone for the vaulting in the King's College chapel was taken by boat up the river Cam or by wagon to the building site. The rough stones from the quarries were in large blocks; a typical rough stone was probably a cube about two-thirds of a meter on a side. Temporary workshops were erected next to the building site and the stones were partially carved there.



CEILING TRACERY in the Chapel of Henry VII is among the most intricate examples of the patterns carved on the surface of fan vaults. The chapel was meant to glorify the Tudor dynasty; the carving is the material analogue of ornate praise for the Tudors. The vaulting is constructed entirely of jointed masonry, in which closely fitted blocks of stone compose the vaulting surface. Jointed masonry affords the best surface on which to carve elaborate designs.

Indeed, in the workshops the chapel vaulting was prefabricated. The rough stones were carved to the correct size and shape and the surfaces where the stones were to join were precisely finished. The rib and the tracery on the face of the block that was to be on the exposed surface of the conoid, however, needed only to be roughly carved. The stone was marked with a symbol indicating the quadrant of the bay it was to go into. Arabic or Roman numerals were carved at the end of the stone near the joint to designate the horizontal level of the conoid the stone was intended for, the exact location within that level and the timing of the insertion of the stone in the construction process.

The stone setter on the building site had only to observe the markings to put the stone in its correct place. The prefabrication of the vaulting stones may have required more skill than the actual construction. From a study of the construction markings visible on the upper surface of the conoids in the King's College chapel I have concluded that the bays in the main vault were built as follows. First the walls were erected and the roof was installed, following the common medieval building practice. Then the large transverse arches between bays were erected. All the stone setting probably required elaborate centering. The next step was the construction of the conoids, beginning at the long sides of the bays along the transverse arches and proceeding inward. One horizontal level at a time was constructed in each conoid. When the arches met in the center of the bay, the large stone that constitutes the keystone for both horizontal arches was placed. When the conoids were complete, the longitudinal ridge stones on each side of the central bosses were put in position. The final step was the insertion of the heavy central bosses.

The bosses were dropped in place from above. Because of their great weight they induced compression in the vaulting, which slightly changed its shape and allowed the centering to be readily taken down. When the centering had been removed, the final carving of the surface of the conoid was probably done. After any cracks that might have appeared were mortared the conoids were sometimes painted. The vaulting in the chapel of King's College was originally to have been painted and gilded.

Although the main vault at King's College has the longest span of any fan vaulting, the most splendid fan vaulting is that in the Chapel of Henry VII. The span of the chapel, 10.6 meters, is only slightly less than that of King's College and the overall design is awesome, combining a simplicity of three-dimensional form with a very complex

© 1983 SCIENTIFIC AMERICAN, INC

and regular surface pattern. Because of the high degree of surface articulation, which was considered fitting for a monarch, the vaulting was constructed entirely of jointed masonry.

The structural solution in the Chapel of Henry VII is ingenious. Large, mostly hidden transverse arches support pendants placed near the wall. Conoids are built upward and outward from the pendants. At the pendants the transverse arch rises through the vaulting conoid and continues along the upper surface of the vaulting, disappearing from view in the central area of the bay. Practically the entire structural load is transmitted to the transverse arch at the pendant. The arch carries the thrust to the walls and buttresses. Hence the ceiling combines shell structures (the conoids) and frame structures (the transverse arches).

Unlike the conoids of many fan vaults, where each horizontal cross section is only a segment of a circle, the central conoids in the Chapel of Henry VII are fully circular in horizontal section. This is symbolically appropriate because the circle, representing the revolution of the heavens and the disk of the sun, was an important symbol in Tudor political iconography. In 1500, when Henry's chapel was designed, the Tudors were a young and self-conscious dynasty; Henry had seized power only 15 years before on the basis of a slim claim to the crown. The magnificently carved ceiling of the chapel with its multiple circular forms was intended as elaborate praise for a monarch anxious to secure a place for himself in the cosmic and historical orders. The ceiling tells us more about the chapel's patron and his society than it does about the personality of the designer, whose identity we can only guess at.

The ceiling of the Chapel of Henry VII represents the culmination of fan vaulting. In the 1540's Henry's son Henry VIII drastically reduced the authority and wealth of the monastic orders in whose buildings much fan vaulting had been included. The construction of fan vaulting came to a virtual halt for almost 100 years, until it was revived by the Church of England at the University of Oxford, which was then a stronghold of Anglicanism.

The construction of fan vaults has never ceased entirely. Some modern cathedrals have vaulting in imitation of the earlier styles. The later examples, however, lack the charm and interest of fan vaults built between 1350 and 1540. This interest comes not only from the beauty of the ceilings themselves, which is considerable, but also from the way the form was worked out by the builders, who were guided solely by aesthetic impulses and empirical results but who were nonetheless dramatically successful engineers.

Now NRI takes you inside the new TRS-80 Model III microcomputer to train you at home as the new breed of computer specialist!

NRI teams up with Radio Shack advanced technology to teach you how to use, program and service state-of-the-art microcomputers...

It's no longer enough to be just a programmer or a technician. With microcomputers moving into the fabric of our lives (over 250,000 of the TRS-80™ alone have been sold), interdisciplinary skills are demanded. And NRI can prepare you with the first course of its kind, covering the complete world of the microcomputer.

Learn At Home in Your Spare Time

With NRI training, the programmer gains practical knowledge of hardware, enabling him to design simpler, more effective programs. And, with advanced programming skills, the technician can test and debug systems quickly and easily.

Only NRI gives you both kinds of training with the convenience of home study. No classroom pressures, no night school, no gasoline wasted. You learn at your convenience, at your own pace. Yet you're always backed by the NRI staff and your instructor, answering questions and giving you guidance.

You Get Your Own Computer to Learn On and Keep

NRI training is hands-on training with practical experiments and demonstrations. You don't just program your computer, you go inside it... watch how circuits interact... interface with other systems ...gain a real insight into its nature.

You also work with an advanced liquid crystal display hand-held multimeter and the NRI Discovery Lab,® performing over 60 separate experiments. Both microcomputer and equipment come as part of your training for you to use and keep.

Computer Assisted Instruction

Your TRS-80 even helps train you. You receive 8 special lesson tapes in BASIC computer language. Using them in your microcomputer, you "talk" to it as you progress. Errors are explained, graphics and animation drive home key points. Within a matter of minutes, you'll be able to write simple programs yourself.

Send for Free Catalog... No Salesman Will Call

Get all the details on this exciting course in NRI's free, 100-page catalog. It shows all equipment, lesson outlines, and facts on other electronics courses such as Electronic Design, Industrial Electronics, TV/Audio/Video Servicing... 11 different career opportunities in all. Send today, no salesman will ever bother you. Keep up with the latest technology as you learn on the latest model of the world's most popular computer. If coupon has been used, write to NRI Schools, 3939 Wisconsin Ave., Washington, D.C. 20016.

(TRS-80 is a trademark of the Radio Shack division of Tandy Corp.)

NRI McGraw-Hill

NRI Schools
McGraw-Hill Continuing Education Center
3939 Wisconsin Avenue
Washington, D.C. 20016
We'll give you tomorrow.

NO SALESMAN WILL CALL.

Name (Please Print) Age

Street

City/State/Zip

Accredited by the Accrediting Commission of the National Home Study Council

Please check for one free catalog only.

- ☐ Computer Electronics including Microcomputers
- ☐ Color TV, Audio, and Video System Servicing
- ☐ Electronics Design Technology
- ☐ Digital Electronics
- ☐ Communications Electronics • FCC Licenses • Mobile CB • Aircraft • Marine
- ☐ Industrial Electronics
- ☐ Basic Electronics
- ☐ Small Engine Servicing
- ☐ Appliance Servicing
- ☐ Automotive Servicing
- ☐ Auto Air Conditioning
- ☐ Air Conditioning, Heating, Refrigeration, & Solar Technology
- ☐ Building Construction

All career courses approved under GI bill.
☐ Check for details

183-023

© 1983 SCIENTIFIC AMERICAN, INC

THE AMATEUR SCIENTIST

Thermal oscillators: systems that seesaw, buzz or howl under the influence of heat

by Jearl Walker

A thermal oscillator is a curiosity that turns up in physics classes because it is an example of a self-sustained, nonlinear oscillating system. What happens is that one part of the system vibrates continuously because of the transfer of heat. Here I shall discuss four such devices, two of them new and two that were demonstrated in the 19th century. Although each one depends on the periodic redistribution of heat, they are all quite different.

Pierre Welander of the University of Washington has devised a novel thermal oscillator based on water. The opening on each side of a U-tube is fitted with a wide container. The tube and about half of each container are filled with water. The tops of the containers are open. The bottom of the tube is heated; the upper sides are cooled. Soon the water level in the apparatus begins to oscillate between the two sides with a period of between several minutes and several hours.

Suppose the system is initially in complete equilibrium with its two sides containing equal amounts of water at the same temperature distribution. Consider a tiny volume of water lying within a thin slice through a cross section at the bottom of the tube. The volume is pressed from both sides because of the weight of the water in the two sides of the tube.

Two factors determine the weight of a column of water: the height of the column and the density of the water. The density is determined by the temperature: cooler water is denser than warmer water. Since initially the two sides of the system are identical in water level and temperature distribution, the central volume of water is pressed equally from both sides. It remains stationary and all the water is in equilibrium.

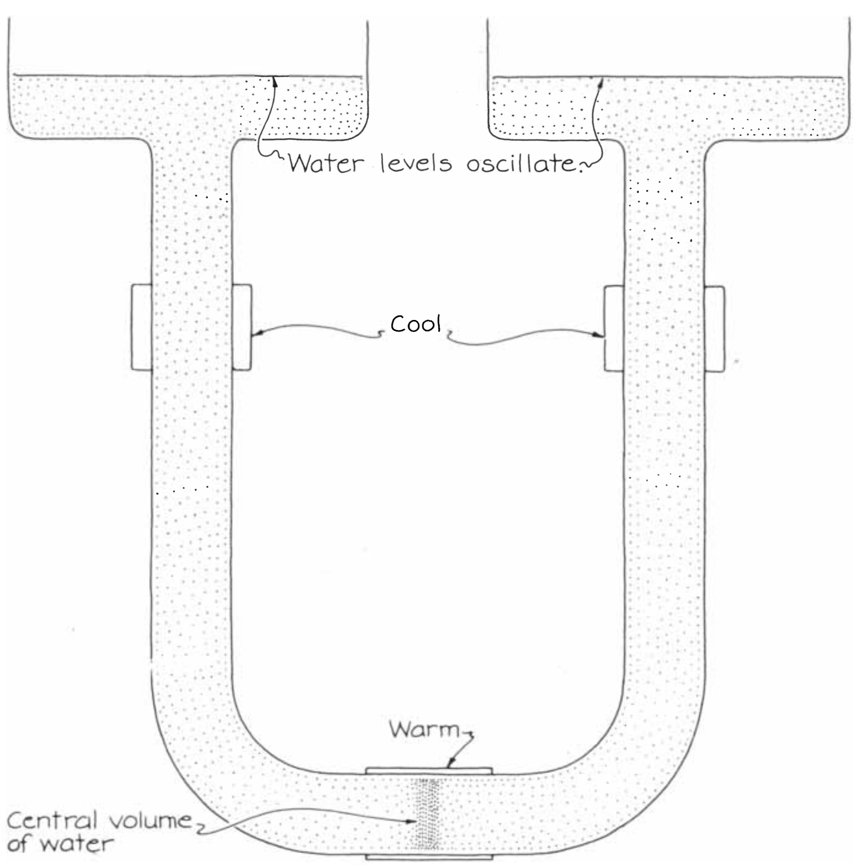
Suppose a disturbance from the environment causes a small amount of water to move from the left side to the right. The flow upsets the symmetry of the system; now the two sides differ in both water level and temperature distribution. The motion lowers the level on the left side and moves cooler water downward. It raises the level on the right side and moves warmer water to the right at the bottom of the U-tube.

If the containers on top of the U-tube are wide, the changes in the water levels are initially small. The broken symmetry of the temperature distribution is more important. Since the cooler water at the bottom left is denser than the warmer water at the bottom right, the central volume of water is no longer under equal pressure from the two sides. The left side now pushes more. The uneven pressures tend to push more water to the right.

Although the initial change in water levels is small, the continued flow increases the difference until it is significant. Eventually the right side applies more pressure on the central volume than the left side because of the difference in the water levels. The flow from left to right is slowed and then stopped. The process is gradual, however, and so the heating and cooling of the tube has a chance to reestablish the temperature symmetry. By the time the flow stops the system is again approximately symmetrical in temperature, but now the level is higher on the right side than on the left. The water begins to flow back to the left side, and the cycle is renewed in the opposite direction. The system remains in oscillation.

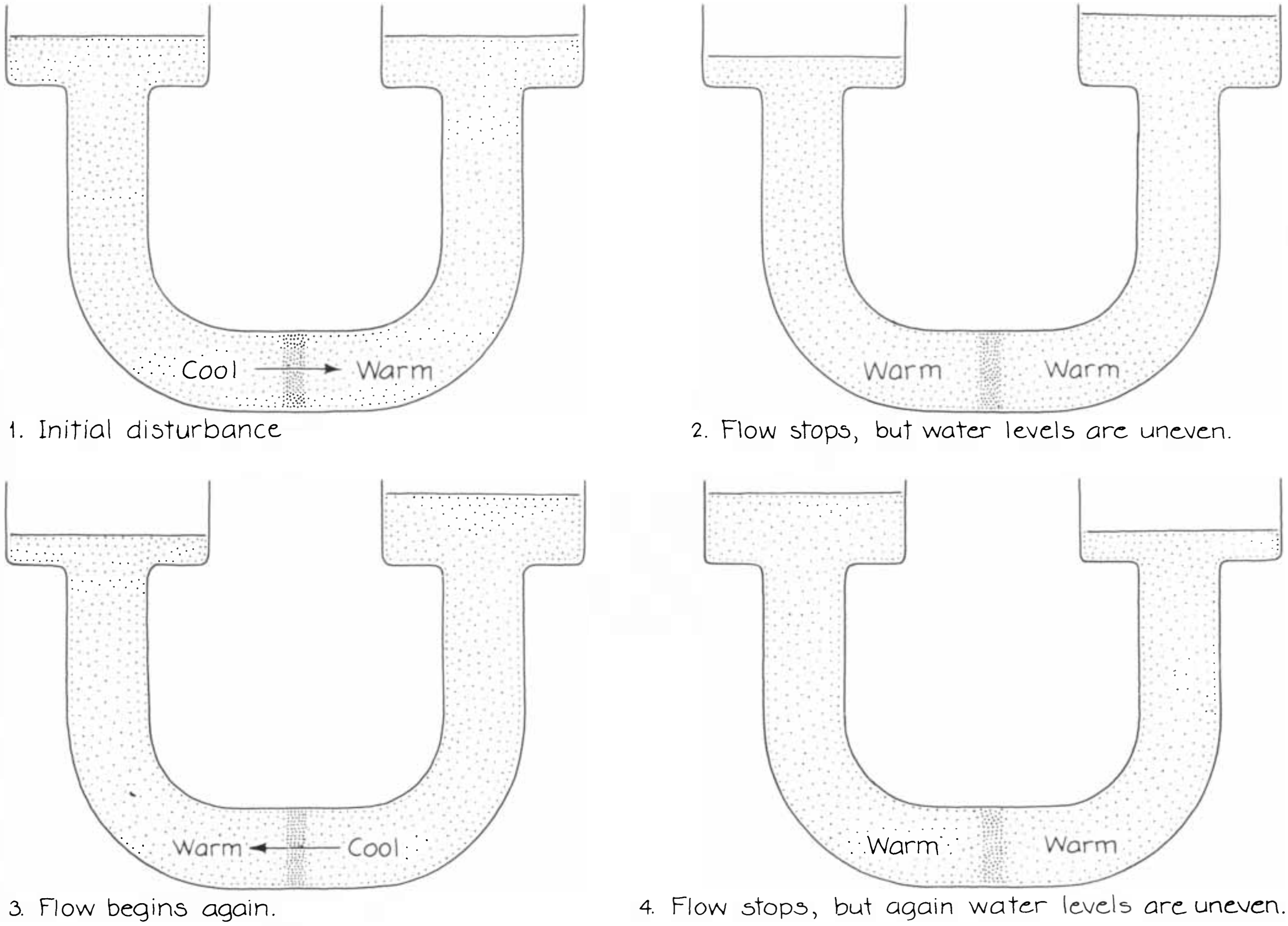
The oscillating flow continues indefinitely because of the heating and cooling of the tube. If the tube is allowed to reach a uniform temperature, the oscillation soon damps down into a stable situation with equal water levels on the two sides. As long as the heating and cooling are continued, however, the oscillation is maintained. Every time the system passes through the state where the two water levels are equal the imbalance in densities drives the water until the levels are unequal, continuing the oscillation.

The dimensions of the system are critical to the onset of oscillation. The top



The physical conditions of a thermal oscillator devised by Pierre Welander

© 1983 SCIENTIFIC AMERICAN, INC

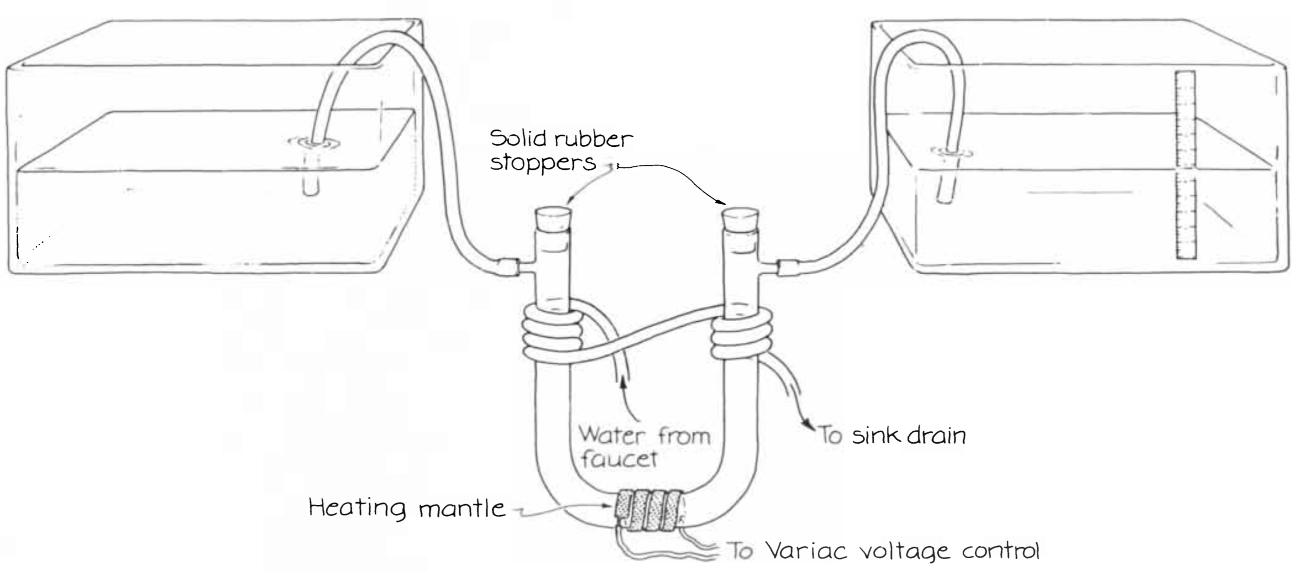


The pattern of flow in a fluid oscillator

surface area of each container must be larger than the cross-sectional area of the tube by an amount that is related to the temperature distribution in the water. The ratio of the area of a container to the area of the tube must exceed 5,000 divided by the maximum temperature difference (in degrees Celsius) that is set up in the water by the heating and cooling.

For example, if the system gives rise to a temperature difference of 50 degrees C., the top surface area of the container should be at least 100 times larger than the cross-sectional area of the tube. A smaller temperature difference demands an even larger ratio of the areas. If the containers are unusually large, of course, the oscillation of the water levels may be difficult to perceive.

The requirement for the relative sizes



A setup for an oscillator similar to Welander's

© 1983 SCIENTIFIC AMERICAN, INC

of the areas arises from the interplay of pressure on the central volume at the bottom of the U-tube. Suppose the initial disturbance from the environment moves a small amount of water to the right. The resulting asymmetry in temperature tends to drive more water to the right. This flow is countered by the difference in water levels on the two sides. If the top surface area of the containers is too small, the flow immediately creates a difference in water levels that is more than enough to stop the flow resulting from the asymmetry in temperature. Equilibrium is quickly reestablished.

If oscillation is to begin, the top surface area of the containers must be relatively large. That way the flow driven by the temperature asymmetry does not immediately yield a large difference in water levels. The flow has time to develop. The period of oscillation is set by the relative areas of the containers and the tube, the length of the tube, the viscosity of the water and the acceleration of gravity.

An apparatus similar to Welander's can be made by fusing glass containers on top of a U-tube. Since the tube should be Pyrex to preclude its shattering from being heated and cooled, fusing containers onto the tube seemed too difficult to me. I was also worried about how to mount such a delicate system so that it would stay upright. I chose a U-tube with side arms near the top. Short lengths of rubber hose connected the arms to plastic containers I took from my kitchen. The containers were identical; the top surface area of each one was about 200 times the cross-sectional area of the U-tube.

I taped a transparent ruler to the inside of one container so that I could monitor the water level. The U-tube and about half of each container were filled with water. Around the bottom of the U-tube I wrapped a heating mantle of the type that can be found among the supplies of a standard chemical laboratory. This mantle contains a high-resistance electrical wire that supplies heat. I plugged it into a Variac voltage control so that I could regulate the amount of heat applied to the tube. (I took great pains to make sure there would be no sudden leak that would let water reach the electrical line.)

For cooling I wrapped a thin-walled rubber hose several times around the top of the U-tube. Tap water could then be run through the hose. Although the hose did help to cool the water in the top of the U-tube, I decided a better system was needed. The conduction of heat through the glass wall of the U-tube and the rubber wall of the hose seemed to be too slow.

I had no way of measuring the temperature distribution of the water in the U-tube. After experimenting with several settings on the Variac I did manage to observe small oscillations in the water level of the containers. The level adjacent to the ruler rose and fell by about a millimeter every five minutes. Although the oscillations were not large or always regularly spaced, they were of the type observed by Welander because they disappeared when the heating and cooling systems were eliminated.

Welander told me about a more sensitive monitor of the water level. A small cork is hung from the edge of the container on a short arm made of wire. The cork floats on the water. A vertical straw is mounted on the cork. When the water level changes, the floating cork rotates about the wire arm. The upper end of the straw magnifies the motion.

Welander has also described a different kind of thermal oscillator in which a thin layer of ice forms periodically in a container of water. His experiment was done with about 50 liters of water in a Plexiglas container having a square bottom 50 centimeters on an edge. The container was thermally insulated. Near the bottom a heating element released about 100 watts of heat. Magnetic stirrers constantly mixed the water to distribute the heat. The apparatus was kept in a compartment where the air temperature was maintained at −20 degrees C. The air above the water was also continuously stirred.

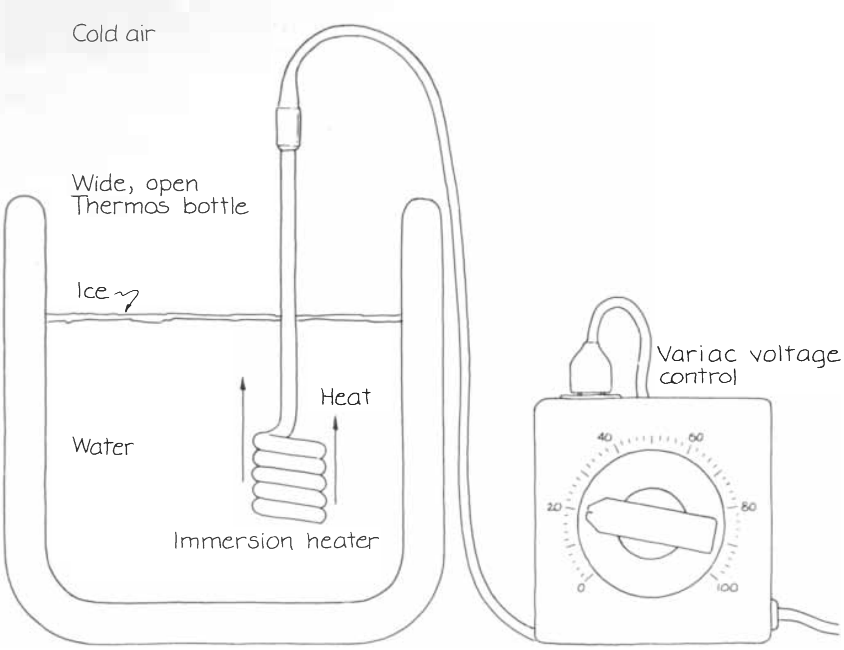
Welander monitored the temperature of the water at half of the depth. He also watched for the formation of ice. The surface periodically froze and melted as the temperature of the water varied by about 10 degrees. The water temperature rose to about 10 degrees C. during the phase of the oscillation when the top surface was frozen. It fell to almost the freezing point during the phase without ice. The period of oscillation was roughly 50 hours.

You can observe similar oscillations of freezing and melting water at home. Welander suggests using a Thermos bottle with a wide top. Fill it with water almost to the top, insert a heating coil below the surface and place the apparatus in a freezer. Ice periodically forms and disappears, although the oscillations are not as uniform as those Welander observed in the larger apparatus.

I followed his procedure with one change. The heating coil was the immersion heater with which I normally heat water for instant coffee. The heater was suspended in place over the center of the Thermos by several strips of reinforced tape. A problem was that the metal casing of the heater extended upward through the surface of the water. Heat from it surely interfered with the periodic formation of ice.

I plugged the immersion heater into a Variac so that I could experiment with the rate at which heat was released in the water. The rubber gasket lining the edge of the freezer door compressed enough so that I could close the door over the cord. If it had not closed completely, I was prepared to pack the edge with cloth to keep out room air. I turned the control of the freezer to its coldest setting. Again I took precautions to keep water from spilling onto the electrical supply.

The oscillations in the formation of ice are due to the poor transfer of heat through ice compared with the transfer through liquid water that is being stirred, either mechanically or by convection. The heat supplied to the system



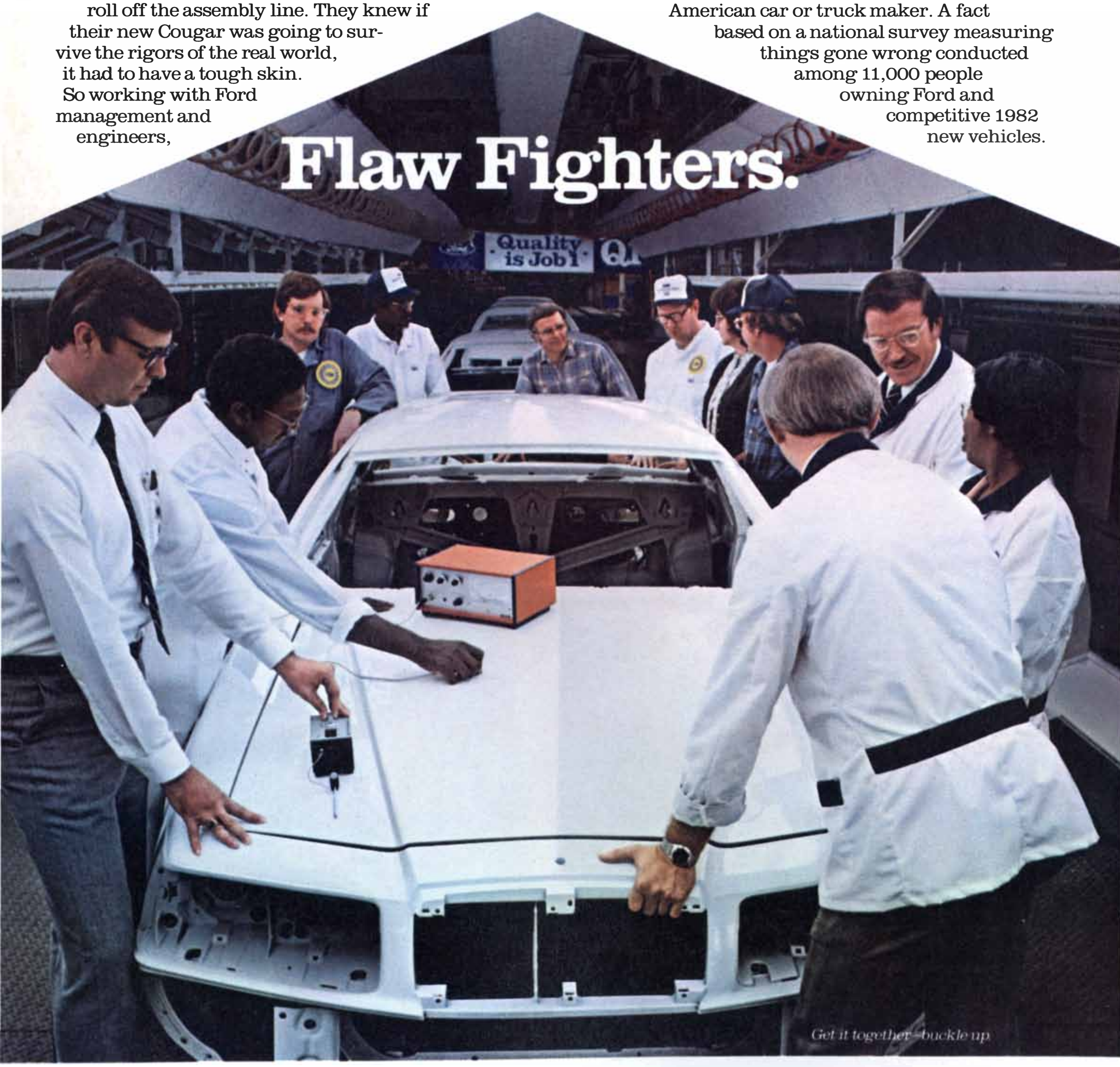
An arrangement in which the oscillation is between ice and liquid water

© 1983 SCIENTIFIC AMERICAN, INC

What is needed in a time when people are keeping their cars longer, is a car people can be proud of longer. That's why, in 65 plants across the country, over 1000 Employee Involvement Groups have been established by Ford Motor Company and the UAW.

Their goal? Build the highest quality cars and trucks in the world. So you'll buy the results of their labor instead of someone else's. For example, members of this group at the Ford Lorain Assembly Plant set out to protect the 1983 Cougar's aerodynamic new body. Months before the first car was due to roll off the assembly line. They knew if their new Cougar was going to survive the rigors of the real world, it had to have a tough skin. So working with Ford management and engineers, Lorain employees used their on-the-job experience to help set up new computer-guided spray equipment. A system specifically designed to apply paint and primer more uniformly to the Cougar's dramatic new shape. Thus enhancing the quality of the car's finish as well as its looks. And that's only one example of 129 employee-inspired quality improvements adopted by Ford management on the new 1983 Cougar. Employees speaking and management listening. Another reason Ford Motor Company achieved the highest quality rating of any major American car or truck maker. A fact based on a national survey measuring things gone wrong conducted among 11,000 people owning Ford and competitive 1982 new vehicles.

Flaw Fighters.



Ford
Mercury
Lincoln
Ford Trucks
Ford Tractors

Quality is Job 1.

Ford

© 1983 SCIENTIFIC AMERICAN, INC

Over 300 manufacturers sell their microcomputers with Microsoft software.

There's a reason.

Powerful simplicity. That's the concept behind every software program we write. Powerful programs that let you spend more time thinking about the problem... and less time thinking about the computer. That concept has earned us the confidence of over 300 microcomputer manufacturers. Starting with the very first microcomputer you could buy. And today, Microsoft® languages, operating systems and applications software are running on well over a million microcomputers. Worldwide.

Made for each other. Manufacturers literally build their computers around Microsoft software. That's no exaggeration. In fact, Microsoft frequently participates in the initial design and development of microcomputers. That's a major reason why Microsoft software runs so well on the majority of the world's 8- and 16-bit systems.

User-oriented. Microsoft is the only software supplier to offer a full range of compatible operating systems, languages, utilities and applications programs. If you're not a computer expert, here's what that means to you: Better programs. Programs that are not only more powerful, but easier to learn and use.

Better tools. We can't buy your trust. We have to earn it. With better products. Tools you can easily use to solve complex problems. We started with the first BASIC for the first microcomputer. Today, we offer a broad range of proven tools for microcomputers. Including SoftCard™ and IBM® RAMCard™ products that materially increase the capabilities of Apple ® and IBM PC computers. Plus, business and management software such as Multiplan™ the powerful electronic worksheet you can learn to use in just a few hours. Better tools. Because they're designed specifically for your computer.

Ask your Microsoft dealer. Most microcomputer manufacturers offer their systems with some Microsoft software, but you'll undoubtedly want more. Programs and languages that solve your specific problems. Programs for your industry, business or home. Programs that do more, yet ask less. Ask your computer software dealer for a demonstration. You'll see why more than 300 microcomputer manufacturers offer their systems with Microsoft software. The reason is powerful simplicity.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION
10700 NORTHUP WAY
BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, SoftCard, RAMCard and the Microsoft logo are trademarks of Microsoft Corporation Apple and the Apple logo are registered trademarks of Apple Computer, Inc. IBM and the IBM logo are registered trademarks of International Business Machines Corporation. The Radio Shack logo is a registered trademark of Radio Shack, a Division of Tandy Corporation. The Burroughs logo is a registered trademark of Burroughs Corporation. The Victor logo is a registered trademark of Victor Business Products. The Wang logo is a registered trademark of Wang Laboratories, Inc.

© 1983 SCIENTIFIC AMERICAN, INC

by the submerged coil must travel upward because the sides of the container are insulated. The surface of the water is being cooled by air at a temperature below the freezing point of water.

The system can develop into either of two possible steady states. In one of them the surface of the water remains liquid because the air is not particularly cold and the supply of heat from the coil always prevents freezing. The heat travels upward by conduction and convection in the water. The transfer to the top surface is sufficiently fast to prevent freezing.

In the other steady state the system permanently freezes at the top because the air is quite cold and the rate at which heat is supplied is not particularly high. The heat from the coil travels upward to the ice layer by conduction and convection. There it is conducted slowly through the ice and carried away by the convection currents in the air. The system is in a steady state because although the conduction through the ice is slow, it is still fast enough to keep the temperature of the underlying water from rising high enough to melt the ice.

Oscillations can begin if the air is quite cold and the rate at which heat is supplied by the coil is high. Suppose the water is not frozen. The cold air removes heat from the surface fast enough to initiate freezing. When the ice layer forms, the removal of heat from the system diminishes because ice does not transfer heat as well as stirred liquid water. Thus the temperature of the water below the ice begins to rise. Eventually the water is warm enough to melt the ice.

When the ice disappears, the rate of heat transfer increases and the temperature of the water drops. Later the surface temperature is again at the freezing point and a fresh layer of ice forms.

A different kind of thermal oscillator was discovered in 1805 by an inspector in a smelting factory in Saxony. When a heated bar of one metal is laid across a block of another metal, the bar may oscillate vigorously and noisily. This oscillator was rediscovered independently about 25 years later by Arthur Trevelyan and has been known since as the Trevelyan rocker.

I have seen two kinds of demonstration of this type of thermal oscillator. In each one the choice of metals is important. In the first demonstration two lead plates are mounted in a vise. The top edges of the plates projecting above the vise are filed sharp so that there is a slight separation between them. Then a brass bar is balanced across the plates. When the bar is heated, it begins to rock back and forth. As it hits the edge of one plate and then the edge of the other there is an almost steady tapping.

In the second demonstration a piece of brass with a triangular cross section has two sharp edges at one angle of the triangle. The two edges rest at right angles on the edge of a block of lead. Extending from the piece of brass on a line with the two sharp edges is a brass rod that has a weight at the other end; near the far end the rod rests freely on a support. When the rod is heated, the sharp edges of the brass piece rock back and forth around the axis of the rod, making a sound that is approximately constant in frequency as they tap on the lead block.

The rod in Trevelyan's device was copper. Brass is more commonly used today. Perhaps certain other metals could serve in place of brass, copper or lead, but the choice must be made with care if the continuous oscillations are to be produced. The oscillations and the associated tone were apparently first explained by Sir John Leslie soon after the rocker was demonstrated by Trevelyan. Recent work has refined the explanation but has not replaced it.

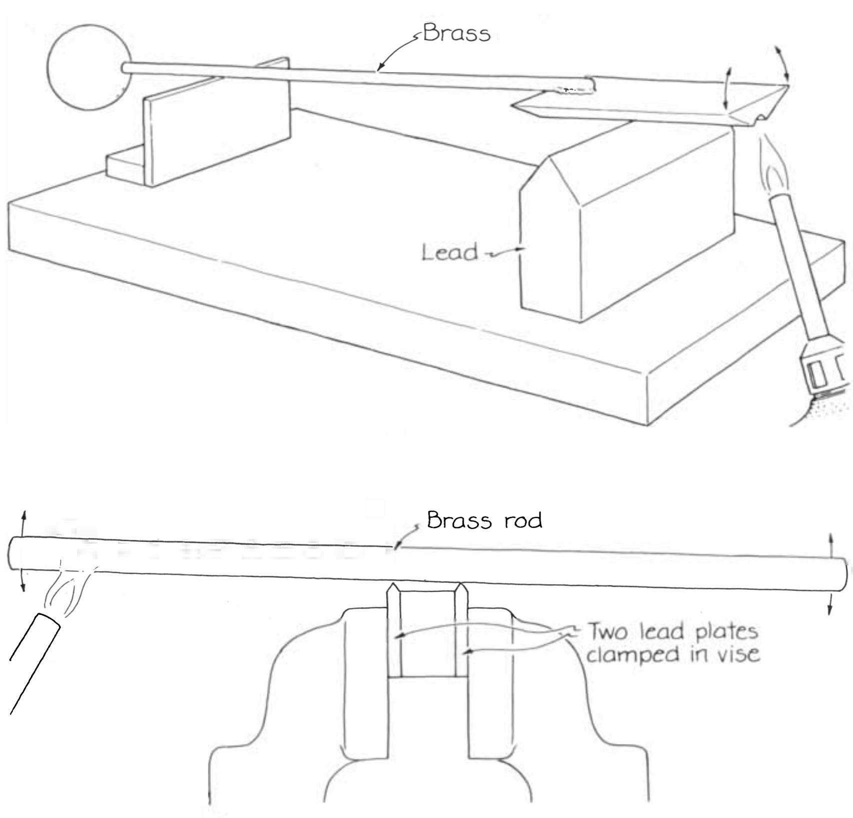
Consider the demonstration in which a brass piece rests on a lead block. The regions of contact are the two sharp edges of the brass piece. As the brass rod and the brass piece are heated by a flame, heat is transferred to the lead through the two edges. The rate of transfer through an edge depends partly on the weight resting on it: more weight provides better contact and a greater transfer of heat. One of the edges initially transfers heat better. Thermal expansion builds up the underlying lead into a tiny mound. The expansion can be fast enough to push the edge upward and off the mound.

Two things happen. Since one edge has been pushed up, the weight of the brass piece is shifted to the other edge. The heat is now transferred to the lead at the second edge. The mound at the first edge cools and shrinks. A fresh mound rises at the second edge. The system oscillates as the cycle of rising and falling mounds repeats.

The rate of tapping depends partly on the dimensions of the rocker. With some rockers the rate is so high that a continuous tone is heard.

Why do some materials not work? One reason is that the block must be of a material that expands when it is heated. Lead expands well, even more than brass and copper do. A second reason is that the vibrations of the rocker can be maintained only if the heat transferred to the block under an edge is not rapidly dispersed. If it is conducted away too fast, the region of the block just under the edge will not have enough time for thermal expansion. Lead conducts just poorly enough to ensure that it is effective for the purpose.

A third reason is that the block must conduct heat fast enough to keep the heat from remaining too long in the region under an edge, otherwise the thermal expansion would persist too long and the mound would subside too slow-



Two versions of the Trevelyan rocker

© 1983 SCIENTIFIC AMERICAN, INC

ly. The edge would no longer be lifted off the block.

You might like to reverse the brass and lead so that a lead piece with two sharp edges rests on a block of brass. The heating procedure cannot be reversed; you must still heat the brass piece. You might also enjoy monitoring the oscillations in either system by flashing a stroboscopic light on the rocker. Although the motion is slight, it can be amplified by attaching a needle or a thin rod to the rocker at right angles to its long axis. The end of the needle moves through a greater distance than the rocker and so makes observation of the motion easier.

Another thermal oscillator was once a standard demonstration in physics classes. A section of wire mesh is pushed about a quarter of the way up inside a vertical pipe. The mesh is bent at the edges so that it remains in place inside the pipe. A burner is set under the pipe so that both the pipe and the mesh get hot. When the burner is removed, the pipe howls for several seconds. The tone is nearly pure because it is the fundamental frequency for acoustic oscillations inside the pipe.

This system was discovered in 1859 by Pieter Leonhard Rijke of the University of Leiden. He thought the sound resulted from thermal contraction of the wire mesh and pipe as they cooled. He believed oscillations accompanying the contractions excite the acoustic wave emitted by the pipe. Rijke was wrong; neither the mesh nor the pipe oscillates.

The correct explanation was supplied by Lord Rayleigh in 1878. The sound from the pipe arises from the forced oscillation of the air inside the pipe. When the burner is in place, it heats the air in the pipe and forces a convection current upward through the mesh and pipe. No sound is emitted at this stage because the air is flowing steadily upward.

When the heat source is removed, the convection current continues for a while because both the pipe and the mesh are hot. The many disturbances of the air in the pipe tend to create an acoustic standing wave in that air. I shall first explain such a standing wave in an unheated pipe and then show how the convection current in the Rijke pipe strengthens it until the pipe howls.

An acoustic standing wave inside a straight pipe can be created by the interference of two sound waves traveling in opposite directions through the pipe. The sound waves must be identical except for their direction of travel. They pass through the pipe, reflect from the open ends and then pass through the pipe again. As they pass through each other their interference forces the molecules of air to oscillate by a small amount parallel to the length of the pipe. Because of the repeated interference of the waves a pattern develops in the oscillation of the molecules. The molecules at some places along the length of the pipe never move. Such a place is known as a displacement node. Molecules at other places oscillate by some maximum amount. Such a place is known as a displacement antinode.

This pattern of repeated interference of two traveling waves is called a standing wave because the positions of the nodes and antinodes in the pipe are stationary. The simplest standing wave in the pipe is termed the fundamental or the first harmonic. Its pattern consists of a displacement node halfway through the pipe and an antinode at each end. Hence the molecules near the open ends vibrate parallel to the length of the pipe by a maximum amount, whereas the molecules near the halfway point vibrate little or not at all.

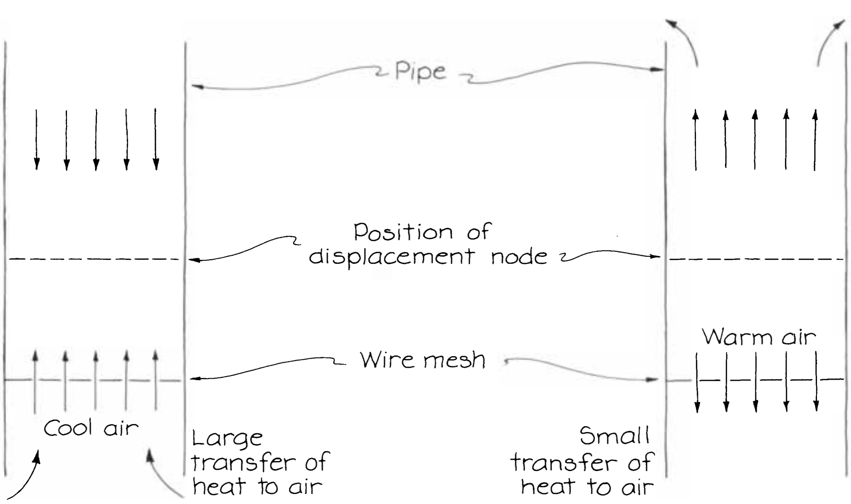
To excite the fundamental in a pipe the traveling waves of sound must have a frequency equal to the speed of sound divided by twice the length of the pipe. Once a standing wave has been initiated it can grow in strength if more energy is given to the sound waves. The pipe is then said to resonate at the frequency of the fundamental standing wave. The sound emerging from the pipe can be quite loud when resonance is achieved.

An acoustic standing wave inside a pipe also sets up periodic variations in the air pressure along the length of the pipe. Consider the node at the midpoint of the pipe when the fundamental is generated. On each side of the node molecules oscillate parallel to the length of the pipe. During one part of the oscillation the molecules on the opposite sides of the node move toward the node, increasing the air pressure there. Half a cycle later in the oscillations the molecules move away from the node, leaving low pressure there. As the molecules oscillate, the pressure at the node varies above and below the atmospheric pressure outside the pipe. The greatest variation in the pressure inside the pipe is at the displacement node.

In some physics textbooks the nodes and antinodes of an acoustic standing wave refer to the pressure variations. Thus a pressure antinode (a large variation in pressure) is found at a displacement node and a pressure node (no variation in pressure) is found at a displacement antinode. In Rayleigh's explanation of the Rijke tube the key is the relative positions of the wire mesh and the displacement node (the pressure antinode) at the midpoint of the pipe. The pipe must be vertical so that the heat forces a convection through it. The mesh must be in the lower half of the pipe, not too close either to the lower end or to the displacement node at the midpoint.

Assume the various disturbances of the air inside the pipe as it cools somehow initiate weak traveling waves that generate the pipe's fundamental standing wave. Air oscillates through the wire mesh. During part of the oscillation cycle the air must move upward through the mesh and toward the displacement node, building up the pressure at the node. During the other part of the cycle the air must move downward through the mesh and away from the node, leaving low pressure at the node.

Rayleigh argued that the oscillation of the air inside the pipe is aided by a transfer of heat from the hot mesh and pipe to the air. The transfer is efficient and beneficial, however, during only part of the oscillation cycle. When the air moves upward through the mesh toward the node, relatively cool air flows into the pipe through the bottom end. Because the temperature difference between the air and the mesh is then large, the mesh transfers a lot of heat to the air. That air is more buoyant because of its increased temperature. The heat supplied by the mesh effectively pushes the mass of air upward toward



How air moves through the Rijke-pipe oscillator

© 1983 SCIENTIFIC AMERICAN, INC

the node, increasing the pressure build-up there and furthering the oscillation.

During the downward motion of the air in the oscillation cycle the heat transferred to the air tends to buoy it upward. This time the push is in the wrong direction to aid the oscillation. Because the push is small, however, it does not seriously hinder the acoustic vibration. The air moving downward is almost as hot as the mesh because it has already been through that region. Hence the transfer of heat from the mesh to the air is small. What transfer does occur can only diminish the oscillation because an addition of heat to the air tends to drive it upward, not downward in synchrony with the oscillation.

The oscillation of the air in the pipe's fundamental standing wave is strengthened by the asymmetry in the transfer of heat during the two parts of the oscillation cycle. The oscillation receives a large push during one part and essentially no push during the other. Rayleigh likened this asymmetry to a force applied periodically to a swinging pendulum. Suppose you push on a pendulum every time it passes through its lowest point. In one part of the swing the push adds energy because you are pushing in the direction of motion. On the next pass of the pendulum through its lowest point, however, you are pushing in the direction opposite to the motion. You remove energy. As long as you apply the same force in each case the pendulum receives no net supply of energy. Its swing eventually decreases because of frictional losses of energy. To build up the swing you must push only when you can add energy to the oscillation.

There is another advantage in transferring heat to the air while the air moves toward the midpoint of the pipe. During the increase in pressure of the air near the midpoint the temperature and pressure of the air are proportional. When the hot mesh increases the temperature, it also enhances the pressure increase at the midpoint. When the air begins to move away from the midpoint and so decreases the pressure there, any addition of heat to the air would again tend to increase the pressure. Hence the fact that the transfer is small during this stage is advantageous to the oscillation.

You can demonstrate the howling of a Rijke pipe with almost any pipe that will not melt when it is heated. Window screen serves well for a mesh. A double layer may work better than a single one. Heat can be supplied by a laboratory burner, a kitchen stove or even a candle.

Once the howling begins try to move the pipe to a horizontal position and then restore it to the vertical. The natural convection through the pipe is destroyed when the pipe is horizontal; the sound ceases immediately. If the pipe is put vertical again before the mesh has cooled, the howling begins anew.

To preserve your copies of

SCIENTIFIC AMERICAN

A choice of handsome and durable library files—or binders—for your copies of SCIENTIFIC AMERICAN. Both styles bound in dark green library fabric stamped in gold leaf.

Files Each file holds 12 issues. Price per file $5.95; three for $17.00; six for $30.00, postpaid.

(Add $2.50 each outside U.S.A.)

Binders Each binder holds 12 issues. Issues open flat. Price per binder $7.50; three for $21.75; six for $42.00, postpaid.

(Add $2.50 each outside U.S.A.)

To: *Jesse Jones Box Corp., P.O. Box 5120, Philadelphia, Pa. 19141*

file binder

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).*

Name ____________________
(please print)

Address ____________________

City ____________________

State ____________ Zip ______

NOTE: Satisfaction guaranteed or money refunded. Allow four to six weeks for delivery.

© 1983 SCIENTIFIC AMERICAN, INC

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

LISP 1.5 PROGRAMMER'S MANUAL. John McCarthy. The MIT Press, 1962.

THE LITTLE LISPER. Daniel P. Friedman. Science Research Associates, Inc., 1974.

LET'S TALK LISP. L. S. Siklóssy. Prentice-Hall, Inc., 1976.

LISP. Patrick Henry Winston and Berthold K. Horn. Addison-Wesley Publishing Co., Inc., 1981.

THE FUTURE OF AMERICAN AGRICULTURE

FARM AND FOOD POLICY: ISSUES OF THE 1980s. Don Paarlberg. University of Nebraska Press, 1980.

THE FUTURE OF AMERICAN AGRICULTURE AS A STRATEGIC RESOURCE. Edited by Sandra S. Batie and Robert G. Healy. The Conservation Foundation, 1980.

THE LATTICE THEORY OF QUARK CONFINEMENT

THE CONFINEMENT OF QUARKS. Yoichiro Nambu in *Scientific American*, Vol. 235, No. 5, pages 48–60; November, 1976.

THE BAG MODEL OF QUARK CONFINEMENT. Kenneth A. Johnson in *Scientific American*, Vol. 241, No. 1, pages 112–121; July, 1979.

GAUGE THEORIES OF THE FORCES BETWEEN ELEMENTARY PARTICLES. Gerard 't Hooft in *Scientific American*, Vol. 242, No. 6, pages 104–138; June, 1980.

LATTICE GAUGE THEORIES AND MONTE CARLO SIMULATIONS. Claudio Rebbi in *Non-Perturbative Aspects of Quantum Field Theory*. World Scientific Publishing Co., Singapore, 1982.

SYNTHETIC VACCINES

CHEMICAL SYNTHESIS OF A POLYPEPTIDE PREDICTED FROM NUCLEOTIDE SEQUENCE ALLOWS DETECTION OF A NEW RETROVIRAL GENE PRODUCT. J. G. Sutcliffe, T. M. Shinnick, N. Green, F.-T. Lie, H. L. Niman and R. A. Lerner in *Nature*, Vol. 287, No. 5785, pages 801–805; October 30, 1980.

ANTIBODIES TO CHEMICALLY SYNTHESIZED PEPTIDES PREDICTED FROM DNA SEQUENCES AS PROBES OF GENE EXPRESSION. Richard A. Lerner, J. Gregor Sutcliffe and Thomas M. Shinnick in *Cell*, Vol. 23, No. 2, pages 309–310; February, 1981.

TAPPING THE IMMUNOLOGICAL REPERTOIRE TO PRODUCE ANTIBODIES OF PREDETERMINED SPECIFICITY. Richard A. Lerner in *Nature*, Vol. 299, No. 5884, pages 592–596; October 14, 1982.

THE OPTICAL COMPUTER

OPTICAL BISTABLE DEVICES: THE BASIC COMPONENTS OF ALL-OPTICAL SYSTEMS? H. M. Gibbs, S. L. McCall and T. N. C. Venkatesan in *Optical Engineering*, Vol. 19, No. 4, pages 463–468; July/August, 1980.

OPTICAL BISTABILITY. Edited by Charles M. Bowden, M. Ciftan and H. R. Robl. Plenum Press, 1981.

DYNAMIC NON-LINEAR OPTICAL PROCESSES IN SEMICONDUCTORS. Alan Miller, David A. B. Miller and S. Desmond Smith in *Advances in Physics*, Vol. 30, No. 6, pages 697–800; November/December, 1981.

BISTABLE OPTICAL DEVICES: PHYSICS AND OPERATING CHARACTERISTICS. David A. B. Miller in *Laser Focus with Fiberoptic Technology*, Vol. 18, No. 4, pages 79–84; April, 1982.

HIDDEN VISUAL PROCESSES

EYE TORSION AND VISUAL TILT ARE MEDIATED BY DIFFERENT BINOCULAR PROCESSES. Jeremy M. Wolfe and Richard Held in *Vision Research*, Vol. 19, No. 8, pages 917–920; 1979.

CYCLOPEAN STIMULATION CAN INFLUENCE SENSATIONS OF SELF-MOTION IN NORMAL AND STEREOBLIND SUBJECTS. Jeremy M. Wolfe and Richard Held in *Perception and Psychophysics*, Vol. 28, No. 2, pages 139–142; 1980.

INTEROCULAR TRANSFER OF VISUAL AFTEREFFECTS. R. Blake, R. Overton and S. Lema-Stern in *Journal of Experimental Psychology: HPP*, Vol. 7, pages 367–381; 1981.

IS ACCOMMODATION COLORBLIND? FOCUSING CHROMATIC CONTOURS. Jeremy M. Wolfe and D. Alfred Owens in *Perception*, Vol. 10, pages 53–62; 1981.

A PURELY BINOCULAR MECHANISM IN HUMAN VISION. Jeremy M. Wolfe and Richard Held in *Vision Research*, Vol. 21, No. 12, pages 1755–1759; 1981.

THE ACTIVE SOLAR CORONA

THE SOLAR CORONA. Jay M. Pasachoff in *Scientific American*, Vol. 229, No. 4, pages 68–79; October, 1973.

A NEW SUN: THE SOLAR RESULTS FROM SKYLAB. John A. Eddy. Scientific and Technical Information Office, National Aeronautics and Space Administration, U.S. Government Printing Office, 1979.

DART-POISON FROGS

PRELIMINARY EVALUATION OF SKIN TOXINS AND VOCALIZATIONS IN TAXONOMIC AND EVOLUTIONARY STUDIES OF POISON-DART FROGS (DENDROBATIDAE). Charles W. Myers and John W. Daly in *Bulletin of the American Museum of Natural History*, Vol. 157, Article 3, pages 173–262; 1976.

CLASSIFICATION OF SKIN ALKALOIDS FROM NEOTROPICAL POISON-DART FROGS (DENDROBATIDAE). John W. Daly, George B. Brown, Monica Mensah-Dwumah and Charles W. Myers in *Toxicon*, Vol. 16, No. 2, pages 163–188; 1978.

MORPHOLOGY OF THE GRANULAR SECRETORY GLANDS IN SKIN OF POISON-DART FROGS (DENDROBATIDAE). Maria Neuwirth, John W. Daly, Charles W. Myers and Lois W. Tice in *Tissue & Cell*, Vol. 11, No. 4, pages 755–771; 1979.

TAXONOMY AND ECOLOGY OF *DENDROBATES BOMBETES*, A NEW ANDEAN POISON FROG WITH NEW SKIN TOXINS. Charles W. Myers and John W. Daly in *American Museum Novitates*, No. 2692, pages 1–23; January 29, 1980.

THE PHARMACOLOGY OF PUMILIOTOXIN-B, 1: INTERACTION WITH CALCIUM SITES IN THE SARCOPLASMIC RETICULUM OF SKELETAL MUSCLE. E. X. Albuquerque, J. E. Warnick, M. A. Maleque, F. C. Kauffman, R. Tamburini, Y. Nimit and J. W. Daly in *Molecular Pharmacology*, Vol. 19, No. 3, pages 411–424; May, 1981.

FAN VAULTING

MEDIEVAL STRUCTURE: THE GOTHIC VAULT. James J. Acland. University of Toronto Press, 1972.

FAN VAULTING: A STUDY OF FORM, TECHNOLOGY AND MEANING. Walter C. Leedy, Jr. Arts and Architecture Press, Santa Monica, 1980.

EXPERIMENTS IN GOTHIC STRUCTURE. Robert Mark. The MIT Press, 1982.

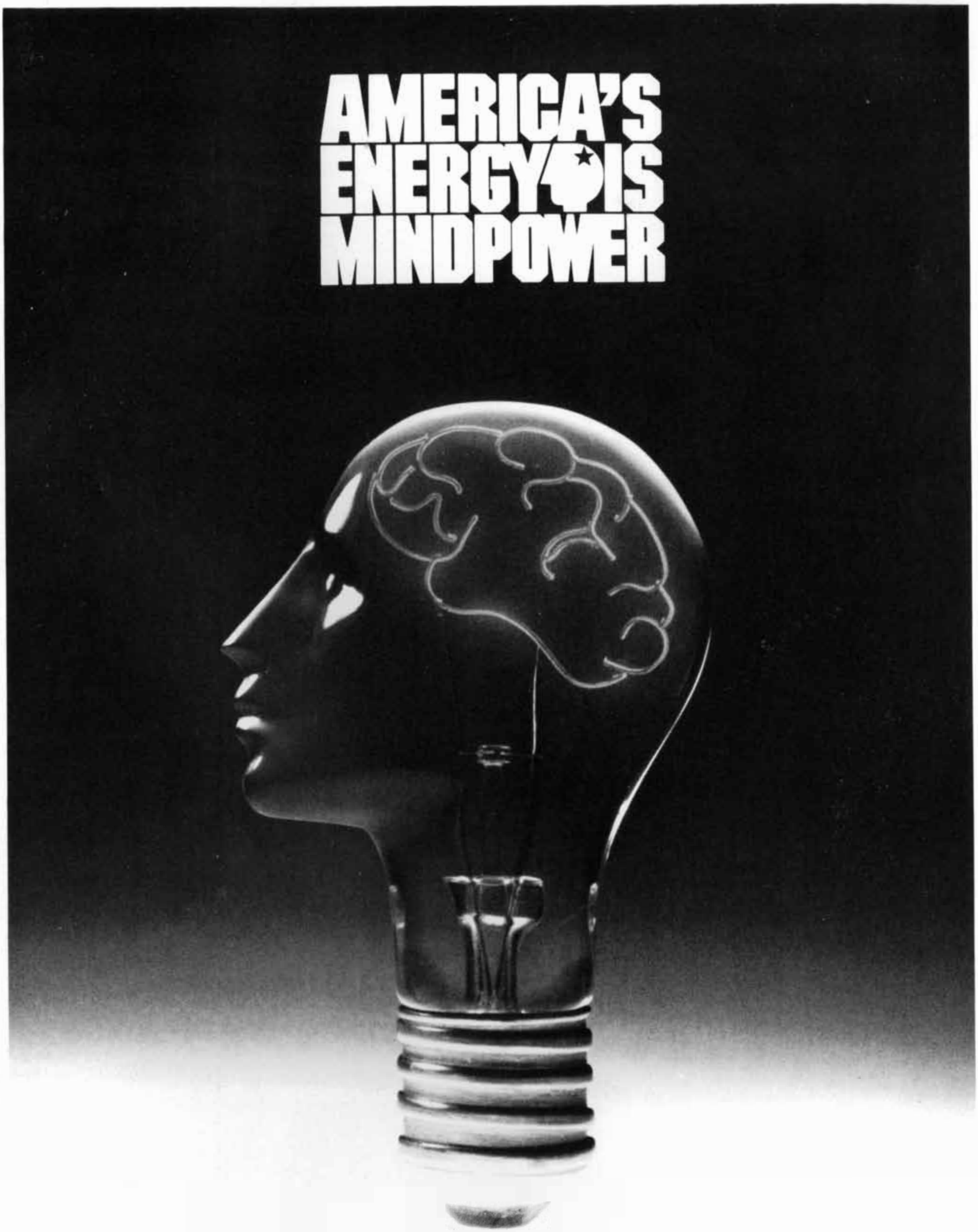
THE AMATEUR SCIENTIST

HEAT-MAINTAINED SOUNDS. E. G. Richardson in *Sound: A Physical Text-book*. Edward Arnold & Co., 1935.

THEORY OF RESONATORS. John William Strutt, Baron Rayleigh, in *The Theory of Sound*. Dover Publications, Inc., 1945.

OBSERVATION OF OSCILLATORY ICE STATES IN A SIMPLE CONVECTION EXPERIMENT. Pierre Welander in *Journal of Geophysical Research*, Vol. 82, No. 18, pages 2591–2592; June 20, 1977.

© 1983 SCIENTIFIC AMERICAN, INC

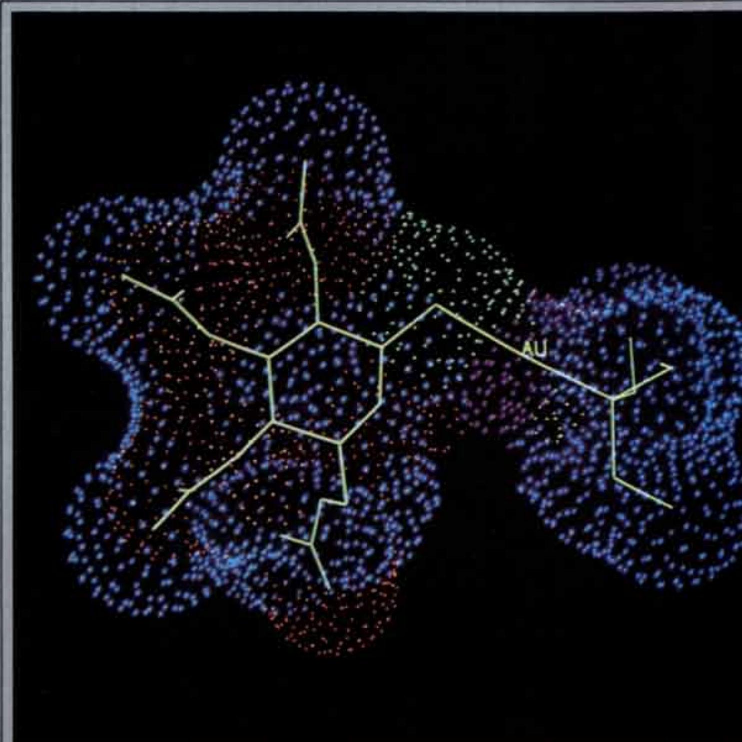


MAKE HIGHER EDUCATION YOUR PRIORITY
Support our colleges: They provide our nation's greatest resource.

© 1983 SCIENTIFIC AMERICAN, INC

SmithKline

SmithKline Beckman Corporation/World Headquarters



Before SmithKline united with Beckman Instruments to become SmithKline Beckman Corporation, those two world leaders in technology had made independent advances for diagnosis and treatment of rheumatoid arthritis. We now continue our complementary efforts.

Immunology and Therapeutics. The SmithKline research program that yielded our promising oral gold agent now under investigation (above, a computer model of the auranofin molecule) has been extended to study on the molecular level how the body's immune system operates and how changes in immunity may affect autoimmune diseases and alter susceptibility to infection.

Expanding our research programs in

Converging

© 1983 SCIENTIFIC AMERICAN, INC

Beckman

One Franklin Plaza/Philadelphia, PA 19101

cellular and molecular biology, we are also investigating new aspects of gene regulation and function to develop new therapeutic agents targeted to key cellular macromolecules.

Immunochemistry and Diagnostics. The capabilities of the Beckman rate-nephelometric Immunochemistry System (above) now include precise analysis of the RF protein marker correlative with rheumatoid arthritis.

Advances in our understanding of similar biochemical markers in other diseases promise new opportunities for development of sensitive assay methods and sophisticated instrumentation.

Look into SmithKline Beckman. And look into the future.

Technologies

© 1983 SCIENTIFIC AMERICAN, INC

© 1982 Toyota Motor Sales, U.S.A., Inc.

When racing driver Dan Gurney drove a Toyota Supra around Willow Springs International Raceway, he was impressed.

As Dan put Supra through its paces, he commented, "This Supra handles better than some race cars I've driven." It felt, as he put it, "like it had been there before."

What makes Supra such a superb performer? Features like 4-wheel independent suspension. Variable assist power rack and pinion steering. Extra wide 225/60HR14 tires on 14"x7" aluminum alloy wheels. And 4-wheel ventilated disc brakes.

Supra's Twin-Cam Six also made Dan say "Wow!" Until now, double overhead cams were found only on the most exotic sports cars. Supra's electronically fuel-injected power plant gives you soul-stirring performance, without constant maintenance.

Inside, the "Wow!" continues. With a driver's Sport Seat that custom-forms to your body with 8 different adjustments — even pneumatically regulated lumbar support! Supra comes standard with a 5-speed overdrive transmission. And its optional 4-speed electronically-controlled automatic overdrive transmission is truly revolutionary! Shift points are selected by a micro computer for maximum performance and economy. You'll see this "next generation" automatic on other cars years from now — Supra has it today!

The Toyota Supra. The car that shook up the automotive world last year returns with even more of the right stuff!

BUCKLE UP—IT'S A GOOD FEELING!

OH WHAT A FEELING!
TOYOTA

"ON THE TRACK, SUPRA FEELS LIKE IT'S BEEN THERE BEFORE." —DAN GURNEY

© 1983 SCIENTIFIC AMERICAN, INC