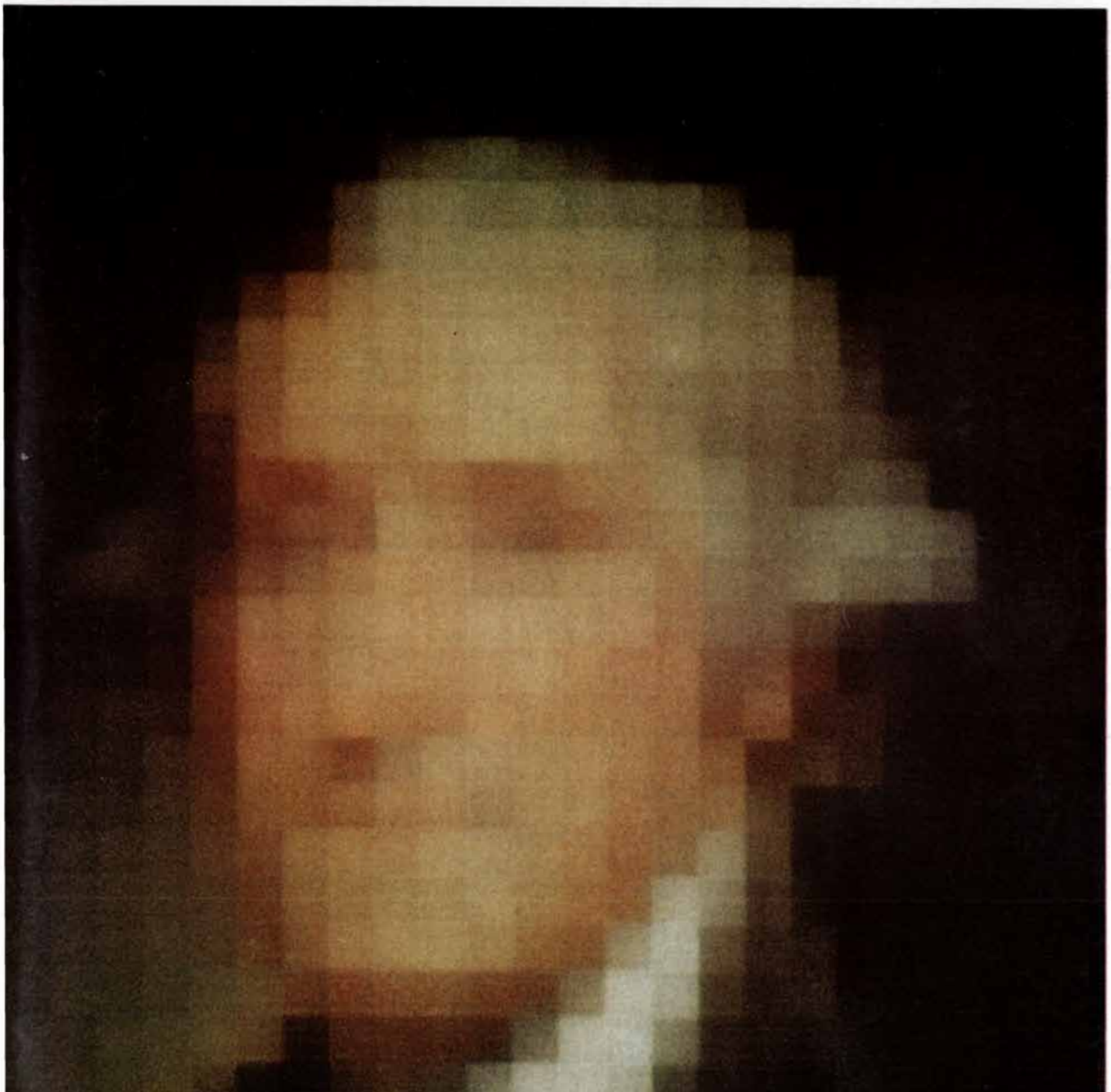
SCIENTIFIC AMERICAN



THE RECOGNITION OF FACES

ONE DOLLAR

November 1973

Scientists and engineers realize that knowledge has no boundaries.

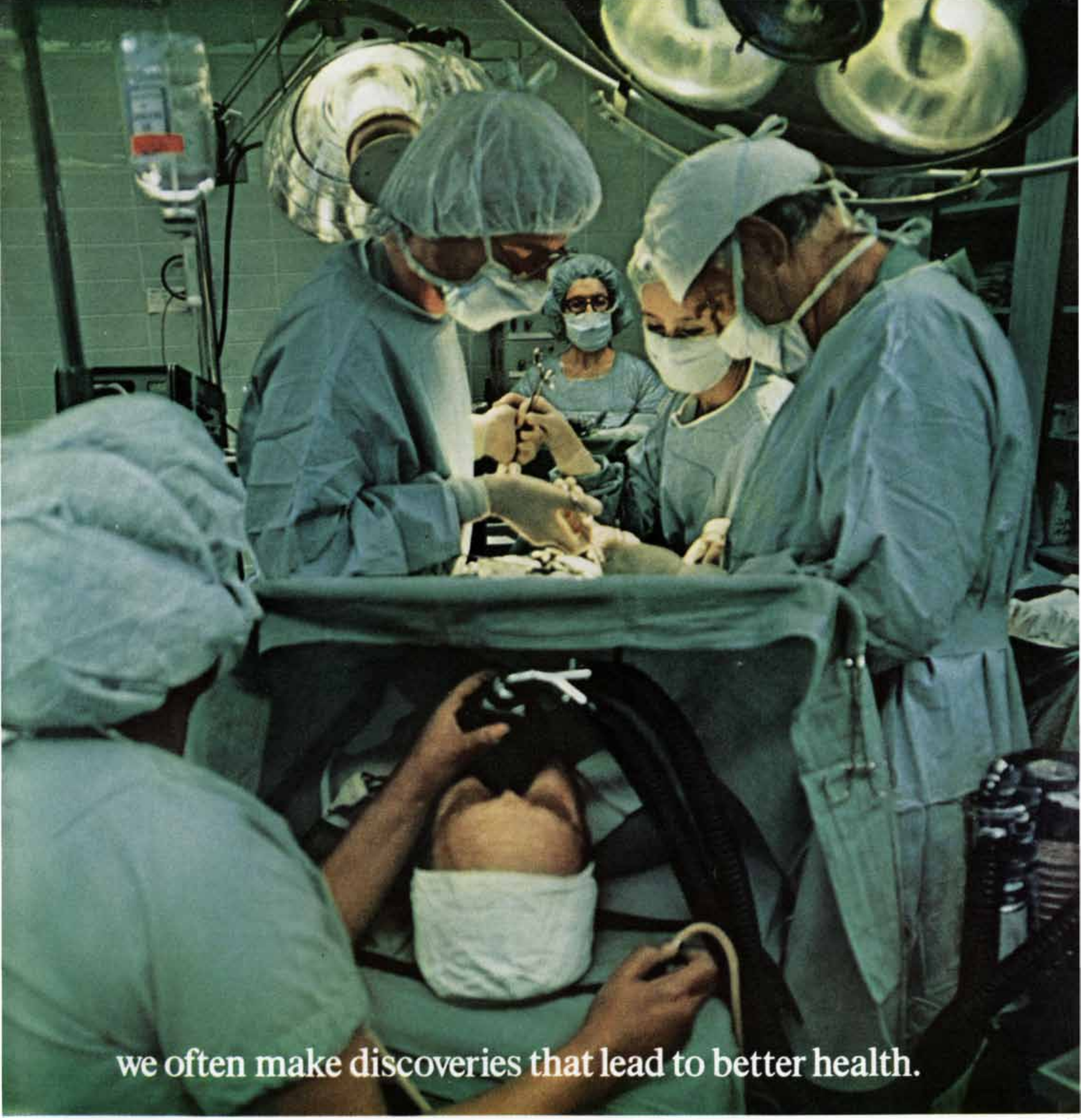
At Bell Labs and Western Electric, our efforts in communications stimulate breakthroughs in other important areas. Including medicine.

Our invention and manufacture of the transistor helped make it possible to regulate heartbeats, to make sensitive internal tests and even to use computers for diagnosis and treatment.

Research into the use of the laser as a communications carrier has led to such experimental applications as welding retinas and performing bloodless surgery.

Our people develop and manufacture specialized communications equipment to help the deaf, the blind and other handicapped.

They made the hearing aid and artificial larynx medical facts. Through the Data-Phone® Medical Set they allow doctors to read EKGs, even though the patients are miles away.



The Bell System's tradition of helping with medical projects dates back to Alexander Graham Bell, whose work with the deaf led to the invention of the telephone. Today, many of our people also spend free time working on medical problems.

The men and women of the Bell System have a commitment to communications.

And sometimes this helps people in other ways, too.

Bell Labs
Western Electric


GM
MARK OF EXCELLENCE
Camaro Type LT.
Camaro has a ride and a feel you'll appreciate, too. Steel-belted radial tires are available for both the Sport Coupe and Type LT.
Camaro has a new standard 350-cu.-in. V8. And all V8 Camaros have power steering standard. Tachometer, ammeter and temperature gauge are standard with the Camaro Type LT.
Camaro Type LT.
Camaro not only looks quick, sleek and nimble. It is. Camaro gives you comfortable seats for four people. Rich upholstery. New cut-pile carpeting.
Camaro has a very sensible new aluminum bumper system that helps cushion minor impacts. New taillights wrap around for visibility. You've called sporty cars impractical for the last time.
1974 Camaro. The way it looks is the way it goes.
Chevrolet. Building a better way to see the U.S.A.
Chevrolet

Established 1845 **SCIENTIFIC AMERICAN** November 1973 Volume 229 Number 5

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1973 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: $10 PER YEAR, U.S., POSSESSIONS AND CANADA; $12.50 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O INTERNATIONAL DISTRIBUTORS, KOEPOORTBRUG 1, B 2000 ANTWERP, BELGIUM.

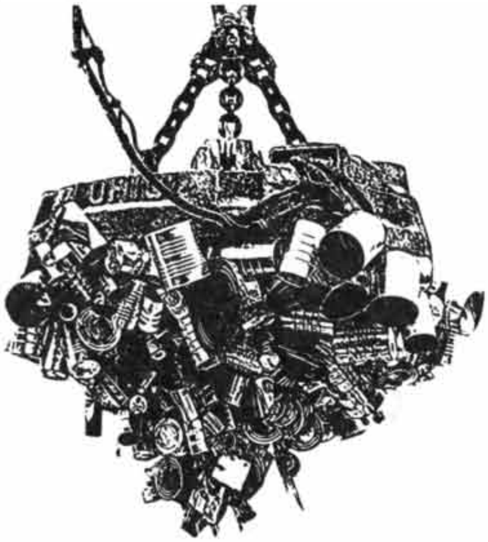
THE COVER

The picture of George Washington on the cover was generated by an optical process from a photographic copy of Gilbert Stuart's painting. The high resolution of the original image has been reduced to an array of 625 squares, each of which has a roughly uniform color and brightness throughout its area. In spite of this loss of detail the subject of the portrait is easily recognized. Perception of the figure can be enhanced by moving one's head, by squinting at the image, by jiggling it or by viewing it from a distance of 10 to 15 feet. The smaller reproduction above can readily be identified at ordinary reading distance. Pictures such as this one have shown that one can recognize faces in spite of a severe reduction in the information conveyed by the image; their role in the study of perception is explained by Leon D. Harmon in "The Recognition of Faces" (*page 70*).

THE ILLUSTRATIONS

Cover courtesy of Leon D. Harmon

Page	Source	Page	Source
19–27	Gabor Kiss	73	Leon D. Harmon (*top*), George V. Kelvin (*bottom*)
29–31	F. W. Goro	74–75	Leon D. Harmon
32–35	Dan Todd	76	George V. Kelvin
37	European Organization for Nuclear Research	77–78	Leon D. Harmon
38	Allen Beechel	79–80	George V. Kelvin
39	European Organization for Nuclear Research	81–82	Leon D. Harmon
40–44	Allen Beechel	84–91	Jerome Kuhl
55	E. A. Munn, Institute of Animal Physiology, Cambridge, England	93	Miriam Rothschild (*photographs*), Tom Prentiss (*drawings*)
56	George V. Kelvin	94–97	Tom Prentiss
57	Emma Shelton, National Cancer Institute	103	*Glomar Challenger*
58–60	George V. Kelvin	104–112	Eric Mose
61	Bernhard Cinader, University of Toronto	117	Michael Beeler, Artificial Intelligence Laboratory, Massachusetts Institute of Technology
62–63	George V. Kelvin	118–120	Ilil Arbel
64	Robert R. Dourmashkin, Clinical Research Centre, Harrow, England	121–122	Michael Beeler, Artificial Intelligence Laboratory, Massachusetts Institute of Technology
65	Burton D. Goldberg, New York University School of Medicine	123	Ilil Arbel
70–72	Leon D. Harmon	124–128	Roger Hayward



St. Louis Converts Garbage Into Energy

In *St. Louis, enough electric power to serve 25,000 homes is generated by mixing garbage with coal in a conventional boiler. This reduces air pollution, saves landfill and conserves coal.

The city's garbage is first mined for steel cans. St. Louis is one of twenty cities reclaiming three billion steel cans annually by magnetic separation—the leading resource recovery method. Steel's unique magnetic property makes it possible.

What happens to the reclaimed steel cans? St. Louis sells them to one of our steel companies for remelting. In other cities the can scrap goes to the copper industry, detinning plants or ferroalloy producers as well as steel mills.

For more information, write Tinplate Producers, American Iron and Steel Institute, 150 East 42nd Street, New York, New York 10017.

Steel — the recycled material

Tinplate Producers/American Iron and Steel Institute

*St. Louis Solid Waste Demonstration Project is funded and operated by U.S. Environmental Protection Agency, City of St. Louis, Union Electric Company, and American Iron and Steel Institute.

The driv

This is the new 3-liter BMW, from the famed Bavarian Motor Works of Munich, Germany. It is a very different concept of what a fine car should be.

That difference lies, essentially, in its purpose. Because above all else, a BMW is built to be *driven*. It is not intended merely to be impressive, or ostentatious, or stately.

Instead, the BMW is designed and constructed, with an almost fanatical dedication, for what it will do on the road. And especially what it will do in the hands of a man who knows and loves great machinery.

The result is a car of extraordinary capabilities. It is a car that will go, and stop, and corner, and maneuver, and respond in a way that, to the uninitiated, is nothing less than astonishing.

Its engine features triple-hemispherical combustion chambers. Such is its output that Road and Track states flatly that "It is unquestionably the most efficient inline six in the world."

Fully independent suspension combines with a multi-jointed rear axle to provide uncanny tire-to-

er's car.

road adhesion, the secret of BMW's legendary roadholding.

In sum, it is a car of tremendous *refinement*. Because to each aspect that is truly important to the car's function, in handling, in safety, in comfort, in quality of construction, we have brought standards born of a half-century tradition of engineering leadership.

Visit your BMW dealer soon, and see this remarkable car. Test it. The more you like to drive, the more you'll appreciate the new BMW.

BMW automobiles are available in four or six cylinder models. Sole U.S. Importer: **Hoffman Motors Corporation.** 375 Park Avenue, New York, N.Y. 10022; 12541 Beatrice Street, Los Angeles, California 90066. Dealers Coast to Coast. Advantageous European Delivery Program. Write for free brochure.

Canada: East: BMW Distributors Eastern Canada, Ltd. 801 Progress Avenue, Scarborough, Ontario.
West: BMW Distributors (Western) Co. 2040 Burrard Street, Vancouver 9, B. C.

Bavarian Motor Works

BMW

LETTERS

Sirs:

On reading "Public Policy on Fertility Control," by Frederick S. Jaffe [SCIENTIFIC AMERICAN, July], the question in my mind is: How is a Government program kept from having "information" and "availability" slide by rapid degrees into indoctrination and coercion?

This past summer a good deal of media attention has been given to the activities of the Montgomery (Ala.) Community Action Committee Inc., an agency funded by the U.S. Office of Economic Opportunity. The agency is currently being sued for damages on behalf of two young black sisters aged 12 and 14 who were sterilized "because boys were starting to hang around them." It should be pointed out that neither of the girls has ever been pregnant and that there is little likelihood that they understood the nature of the operation. The Montgomery Community Action Committee's family-planning clinic states that they really are not sure how many others have been similarly sterilized. One thing they are sure of, however, is that all have been young black girls.

In Aiken County, S.C., three local obstetricians have a policy that requires welfare mothers having their third pregnancy or more to agree to sterilization in order to receive treatment. One of these physicians received $60,846.40 in 1972–1973 for treating welfare patients. Nearly half of the welfare mothers who had babies at Medicaid expense in Aiken County so far this year have been sterilized. The physicians have readily stated that their policy was based on their social views. Women who did not want to take advantage of their "service" could drive many miles to find another doctor and hospital. The local hospital administration can find nothing wrong with the policy of these doctors. Also mail to the local newspaper is heavily in favor of the policy. . . .

To those who assume that the Alabama and South Carolina cases are regrettable but isolated instances, I can only warn that the worst barbarisms often have small beginnings.

CAROLE K. LEVISON

Wexford, Pa.

Sirs:

Ms. Levison raises a number of important issues that certainly need to be addressed:

To begin with the South Carolina case: Some private physicians, based on their own social and economic biases, refused to provide maternity care to pregnant welfare recipients unless the women agreed to sterilization following delivery; if the women agreed to the doctors' conditions, the doctors provided the services (in this case maternity care and sterilization), which were paid for by Medicaid, a joint Federal-state health-financing mechanism. Medicaid is not a Government *program* in the sense that a Government agency makes rules setting forth required standards of professional practice or treatment of recipients, but is an instrument for *financing* specified medical services for specified eligible people. In this regard Medicaid functions exactly like Blue Shield or private health insurance: virtually all of the decision making is decentralized; whether or not Medicaid patients will be served and how they will be treated is determined by the individual physician and hospital. Even decisions on the quality of care the doctor will provide are constrained only by reference to an ambiguous concept of "prevailing standards of medical practice," which can vary from one community to the next, and not by relatively clear regulations, guidelines and monitoring that characterize (or *should* characterize) Government *programs*. Precisely because private medical practice is unsupervised and unmonitored, there is no information on the extent of the abuse surfaced by the Aiken case in regard to Medicaid patients, or other poor patients who may receive charity care from private physicians or hospitals, or even to nonpoor patients whose care is financed by private health insurance or their own funds. Nor is there any systematic information on other similar abuses—such as physicians or hospitals that refuse to treat certain patients at all because of the patient's color or economic status, or to provide patients with certain types of care (e.g., contraception) because of the doctor's or hospital's philosophy or biases. Abuses such as these did not originate with the entry of the Government into the medical economy and constitute a central problem underlying the recent efforts to make the health system more accountable. Many health workers share my view that the incidence of such abuses will not be reduced if the Government role were to be removed and the medical economy left to its own devices. Indeed, only if the Government takes on *more* responsibility for monitoring and evaluating the health programs it agrees to finance, is it possible to reduce the incidence of such abuse.

The Alabama case is different in that a Federally funded, locally administered program was involved: the family-planning project referred the two minors for sterilization to the local hospital (which apparently financed the operations through Medicaid). There are disagreements as to the facts: The project states that the minors are retarded, that protection against pregnancy was sought from the project by the girls' mother and that the mother consented to the sterilizations. In the lawsuit on behalf of the parents and the minors different facts are asserted, but it is significant that the suit for damages against the family-planning program has been dismissed. As a result of the case, the Department of Health, Education, and Welfare has formulated regulations to govern sterilization in Federally funded projects and to more explicitly safeguard the rights of minors and legally incompetent individuals. Family-planning organizations (including the Planned Parenthood Federation of America) have fully supported this

Scientific American, November, 1973; Vol. 229, No. 5. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

Advertising correspondence should be addressed to Harry T. Morris, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 25 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

☐ Please enter change of address.

☐ Please enter subscription.
U.S. and Canada:
☐ 1 year, $10 ☐ 2 years, $18 ☐ 3 years, $25
All other countries:
☐ 1 year, $12.50 ☐ 2 years, $23 ☐ 3 years, $32

NAME

NEW ADDRESS

OLD ADDRESS

effort to develop effective and clear regulations.

The legal, ethical and practical issues posed by the Alabama case (Who is "retarded"? When is consent "informed"? How best to serve the sometimes conflicting rights and interests of the minors and parents involved in cases such as this?) are considerably more complex than Ms. Levison's letter implies. It is noteworthy that the four women members of the Black Congressional Caucus (Representatives Burke, Chisholm, Collins and Jordan) issued a statement on July 10 asserting that "the heart of the issue is how to make family planning information and services available to all those who want and need them and at the same time ensure that no element of coercion creeps into programs which Congress has specifically mandated must be voluntary in nature." They called on the Department of Health, Education, and Welfare to develop detailed regulations and guidelines with the assistance of experts in retardation, law, medicine and other relevant fields. They also noted that the Administration is opposing continuation of the national family-planning program under the Family Planning Services and Population Research Act of 1970 in favor of 50 separate state programs supported by Medicaid and stated: "To turn all family planning programs over to the states will merely compound the problems raised by the [Alabama] case.... Clearly if we are to protect children...and other citizens seeking assistance, there is need for uniform Federal guidelines not only for the OEO and HEW family planning programs but also for the state-delegated Medicaid programs."

Most of us who work in family planning agree with the black congresswomen that the best protection against coercion lies in *more* Federal leadership, participation and supervision, not *less*. We believe that while the record of publicly funded family-planning programs has been relatively free of abuses since 1966 (especially in view of the very large numbers of patients served), there is considerable potential for abuse if the Administration prevails in applying its overall decentralization philosophy to the family-planning program.

FREDERICK S. JAFFE

Center for Family Planning Program Development
Technical Assistance Division of Planned Parenthood–World Population
New York


REYNOLDS ALUMINUM
METAL RECYCLING CENTER
Reynolds
Aluminum
Recycling
Center
SGD-691

...and the cans keep marching in.

Started in 1967, Reynolds Aluminum Recycling programs are still growing, helping save energy and resources, reduce litter and solid waste.

Last year, Reynolds recycled more used aluminum cans than ever before . . . nearly a billion cans, or 35 million pounds of aluminum; and we've learned that these programs do more than collect scrap. Aluminum recycling also:

1. encourages people to do something about litter
2. helps pay for programs that will turn tons of garbage into useful materials and energy
3. saves natural resources
4. conserves energy.

Big and complex jobs, but the empty aluminum can is helping to accomplish more every year.

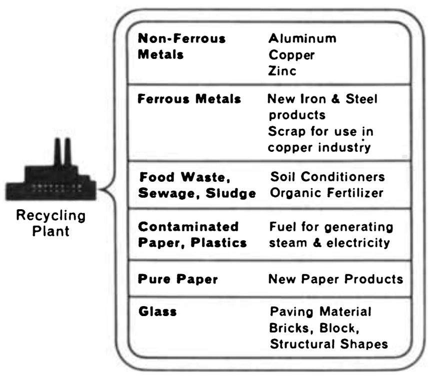
Reynolds first started recycling all-aluminum cans in 1967, so by now we have some fairly sophisticated programs in full swing. These programs prove that the principle, which we call recycling, is one effective tool in tackling many environmental problems.

The key: Aluminum's Value The key is aluminum's intrinsic, virtually indestructible value. To the consumer it is worth $200 a ton as scrap, compared to $16 or $20 a ton for other common packaging materials.

So, although aluminum makes up less than 1% of the nation's garbage, its value can pay a major share of the cost of programs that collect and reclaim other waste materials.

Getting People Involved Reynolds has found, too, that the aluminum can is a powerful incentive for people to get involved in neighborhood cleanup projects. We've been paying 10¢ a pound (about ½¢ per can) for household aluminum scrap brought to collection centers—and that offer has drawn millions of pounds back into the production system.

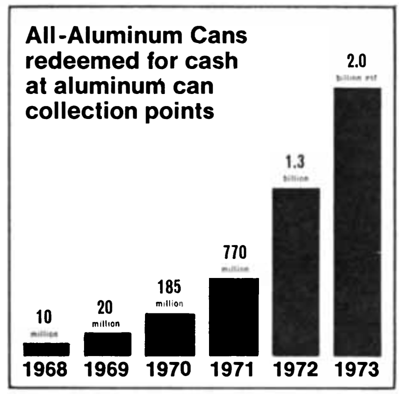
Thanks to cooperating beer and soft drink industries, there are now more than 1200 aluminum collection points in 44 states and the District of Columbia, and the program is growing. This year an estimated 2 *billion* cans will be collected.



Conserving Resources and Energy The used cans can be converted into new cans and other useful aluminum products, helping to save natural resources. And since it takes less than 1/20 of the energy to make new aluminum from scrap as compared to ore, recycling helps stretch our power supply.

Mining Municipal Waste Litter has been called just the tip of the solid waste iceberg. Reynolds has been working on the bigger "hidden" portion as well—municipal solid waste.

We've teamed with other companies and local and Federal government in the planning of soon-to-be constructed plants in several cities. These will reclaim metals, glass and other usable materials and energy from municipal refuse.



We're also working on the theory that junked automobiles will be less of a problem as more and more high-scrap-value aluminum is used in cars. We've even built a home from recycled materials to demonstrate there are vast new markets for recycled products and that recycled materials can look as good and function as well as products from virgin sources.

For details on the value of aluminum, new and old, or on any of these important projects, write Dr. R. F. Testin, Director, Environmental Planning, Reynolds Metals Company. P.O. Box 27003-LS, Richmond, Va. 23261.

REYNOLDS
where new ideas take shape in
ALUMINUM

Pitch In!

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

NOVEMBER, 1923: "We have gradually learned that electricity exists in two forms, the negative form, which is called an electron, and the positive form, which is now beginning to be called a proton. There is no other kind of electricity so far as we know. The material universe seems to be built of these two elements. Both the electron and the proton are exceedingly small, very much smaller than an atom of matter. Both probably have weight, though one is much heavier than the other. The proton weighs as much as 1,830 electrons. But it is not appreciably bigger, and some even think that it may be smaller than an electron. The fact is, we do not know much about it, except that it is the unit of positive electricity, just as an electron is the unit of negative electricity. Whether the proton is an ultimate unit, or whether it can be resolved into a close-packed assemblage of simpler ingredients, which would account for its remarkable weight or massiveness, remains for future discovery."

"What may prove to be an epochal move, bringing about extensive changes in laws relating to motor vehicle operation and, through these, important changes in motor vehicle design, is suggested by a Yale professor as a result of his chemical investigations of the air in our large cities. It was shown that the air is being heavily polluted with carbon monoxide gas. Investigations recently made by Professor Henderson show that the contamination of the air in the more congested streets of New York often reaches the upper limit of a well-founded health standard, and sometimes exceeds it. To a great measure the carbon monoxide is generated as the result of carburetor adjustments that give maximum engine pull. The rich mixture wastes a third of the fuel value of the gasoline and generates much excess carbon monoxide in the exhaust."

"The possible use of oxygen or oxygenated air in its application to the blast-furnace smelting of iron and the bessemer and open-hearth processes for steel is forecast by the Committee for the Application of Oxygen in Metallurgical and Allied Industries. The trend of development in smelting and refining is in the elimination of inert matter that does not enter into the reactions in the furnace and that simply passes through, absorbing and carrying away heat and reduced metal. In the production of one ton of pig iron three tons of nitrogen is passed through the furnace. To diminish this loss oxygen might be substituted, or a mixture of oxygen and air. The application of oxygen will make possible the use of cheaper materials, ore of lower content and coke of a higher ash."

"The trench hat, or 'tin hat,' is coming into quite extensive use as a means of head protection against small falls of rock in mines. Many of the miners have been soldiers and came back with a wholesome respect for the trench hat. It was argued that if, in fun, a comrade could strike the top of the tin hat a strong blow with the butt of a rifle, without local injury to the wearer and no further inconvenience than the sudden vision of a few stars and a seemingly telescoped neck, then the accidental fall of a 10-pound piece of ore, a frequent occurrence in the mines that may seriously wound or kill the miner, would do no damage to him."

SCIENTIFIC AMERICAN

NOVEMBER, 1873: "Professor James Clerk Maxwell lately delivered an interesting lecture on molecules before the British Association for the Advancement of Science. Air confined in a vessel presses, as we say, against the wall thereof. What we term pressure is simply the impact of the moving molecules against the interior surfaces of the vessel. Dr. Joule has calculated the velocity of hydrogen molecules, at the temperature of melting ice, at a little over 6,000 feet per second. Professor Maxwell has calculated the size and weight of the hydrogen molecules, and he finds that about two million of them, placed side by side in a row, would occupy a length about one twenty-fifth of an inch, and that a package containing a million million million million of them would weigh 62 grains, or not quite one-eighth ounce."

"The periodical fluctuations in the light of the star Algol have been accounted for in two different ways: first, by supposing that a nonluminous body revolves around this star, and second, on the hypothesis that Algol is a secondary body, revolving around a dark primary. It seems that if we admit the existence of a dark companion, it would be more correct to say that both bodies revolve around the center of gravity between them, rather than to say that either revolves around the other. There is a method of observation, however, by which the truth of this theory can be put to the test. We refer to spectroscopic observation. In case Algol moves in such an orbit, it is obvious that at times it must be approaching our system and at other times receding from it. If, therefore, the orbital velocity of the star is sufficiently great, displacement in the lines of its spectrum will result. By observing the amount of their displacement at different times the rate at which the star moves in its orbit could be determined."

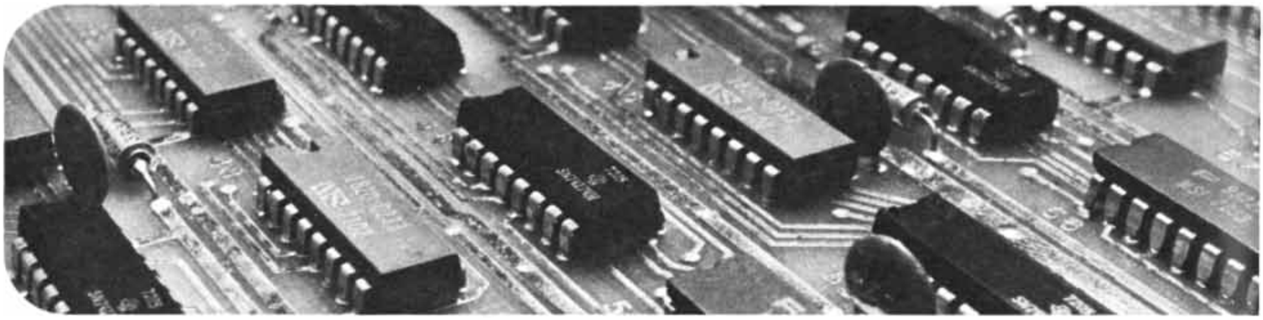
"The yearly rings, shown when the trunk of a tree is transversely divided, by which, as is well known, the age of the tree can be determined, do not diminish in relative thickness by a constant law. M. Charles Gros seeks a cause for the irregularity, and he has arrived at the conclusion that the data, mean and extreme, of meteorological phenomena, when known and tabulated, might be compared year by year with the annual ligneous layers formed during such periods in many different varieties of trees. From the comparison it is not impossible that some interesting ideas relating to the laws of development of trees might be obtained. Moreover, once these laws were established, the trees in turn might become precious collections of meteorological evidence for places and times where observations cannot be made."

"The project of a colossal telescope, which is to be the largest and most complete instrument that modern scientific knowledge can suggest or ingenuity devise, is in progress of elaboration. A telescope of the largest size and most consummate workmanship, properly located 10,000 feet above the sea in the clear skies of the Sierra Nevada, has been the lifelong object of James Lick. The scheme is being quietly perfected, and the geological, meteorological and other peculiarities of various sites of the mountain range above named are soon to be carefully scrutinized and reported upon. A peak will be selected that, from its high altitude and clear surrounding atmosphere, will afford the finest possible view of the heavens."

How to live more comfortably in an increasingly digital world.

In the continuing technological evolution, television sets, automobiles and even kitchen appliances are beginning to use digital electronic circuits once limited to computers.

Having come up against these complications in our own manufacturing and service operations, we've developed a family of instruments that complement the traditional oscilloscopes, DVMs and counters for digital circuit test and analysis. As you read on, you'll learn how these instruments provide easier, faster and more accurate solutions to digital problems.



In the digital world, it's the mere presence or absence of electrical pulses — or their on-off pattern in time — which conveys the information needed for computation, control or communication. In analog circuits, the data conveyed comes from the amplitude or rate of change of electrical signals. Because of this fundamental difference, the many benefits of digital electronics are accompanied by a new set of problems for designers and troubleshooters.

If you've ever had to troubleshoot a digital circuit with traditional analog instruments, the experience probably convinced you there should be a better way.

Take voltage measurement. In an analog circuit, it's important that you know the absolute voltage of a test point, whether it's between 5.015 and 5.018 V., for example. But in a digital circuit, the important thing is to know the logic state of a node, whether it's above the threshold voltage and therefore a logic high, or below the threshold voltage and therefore a logic low.

Time presents another problem. Absolute time measurements are unnecessary in digital systems. Things don't happen after a certain amount of time has elapsed but rather after a certain number of clock pulses, regardless of their duration. The circuit troubleshooter needs to know, for example, if a data bit occurred 1024 clock pulses after an event — and couldn't care less if it took .509 ms to get there.

Furthermore each IC in a digital circuit has 14 or 16 leads and you may need to know what's going on at all the leads simultaneously, not just at two or three input and output leads as with the components of an analog circuit.

The IC troubleshooters pinpoint the circuit problem.

Used singly or in combinations, these instruments are unique in their ability to detect in-circuit logic failures, analyze the cause of the failure and isolate it to a particular IC node.

When the logic clip is attached to an in-circuit IC, it indicates on a bright LED display the logic states at all 14 or 16 pins simultaneously. The clip automatically locates the test IC's ground and power supply, borrows power from it and thus requires no cable connections of any kind. $125.*

The logic probe indicates the condition of a specific IC node and tells you if it's high, low, bad-level or inactive. It captures single pulses as narrow as 10 ns and stretches them to a bright blink at the probe tip. $95* for TTL/DTL or high level logic. $95* for ECL.

The logic pulser teams up with the probe or clip for stimulus-response testing. Typically, you use the pulser to inject reset, shift and clock signals into a circuit, and then observe the effects at a particular IC with the clip, or at an individual node with the probe. The pulser's output is high enough to overcome "clamped" IC outputs yet its pulse is brief enough to prevent circuit damage. $95.*

The logic comparator is a self-contained functional tester that borrows power and input signals from an in-circuit test IC. It instantly compares the logic states of the test IC and those of a reference IC of the same type. Differences are displayed on one of 16 LEDs. Even brief or intermittent failures are detected. $375.*

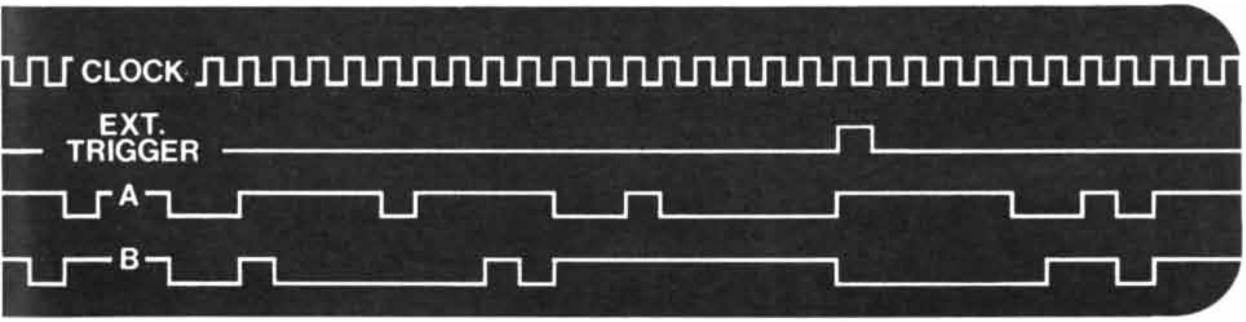
The logic analyzers detect bit-stream problems.

With the instruments described so far, you can determine the logic state of a digital circuit at one particular clock cycle. But if you want to know overall circuit performance and compare it against truth tables or system timing diagrams, you have to learn how the procession of logic states changes with each pulse of the circuit clock.

Two new HP instruments greatly simplify this task. Both extract from long bit patterns the essential information required by the digital troubleshooter: for each successive bit time, is the data (logic state) high or low? The Model 5000A captures 32 successive bits from two parallel channels and displays them as logic highs or lows on LEDs. The Model 1601L does the same job for twelve parallel channels at a time and displays 16 successive bits in each channel, expressed as 1's and 0's on a cathode ray tube.

Both instruments can operate at speeds up to 10 megabits per second, capturing the desired portions of long, non-repetitive bit stream with great flexibility and complete repeatability. You can trigger the display — and therefore the bit-capture — to occur after a single unique event or multiple simultaneous events. You can dial the precise amount of delay you want and thus move the display window forward from the trigger event as many as 10^6 clock pulses (10^5 for the 1601L). With negative delay, you can even look backward in time. You can, for example, trigger the display on a fault condition and see the bit pattern that led to and caused the fault.

The 5000A and the 1601L are compatible with all digital logic families. 5000A, $1900,* 1601L, $2650.*



Automatic circuit tester mass-tests complete digital assemblies.

When you're dealing with large numbers of digital circuit assemblies, manual testing of any sort can be prohibitively expensive or even impossible. This is where the Model 9560B/D Digital Circuit Test System fits in.

This new modular system can simultaneously test all the pins of a complex digital assembly at rates as high as 22,000 truth table patterns per second. Results can be displayed as a simple pass/fail, or extended to specific fault diagnosis using logic probes according to computer-generated instructions. Test programs are written in easy-to-use ATS BASIC; a software package for computer-aided generation of test programs is also available. Prices start at $53,900.*

Our digital capabilities don't stop at the end of this page. The range of HP digital test and analysis includes digital communications testing systems as well as digital pulse generators. For more information on all of the above write to us. Hewlett-Packard, 1502 Page Mill Road, Palo Alto, Calif. 94304.

*US Domestic Prices Only. 00318

Please send me more information on:

( ) IC Troubleshooters
( ) 5000A Logic Analyzer
( ) 1601L Logic Analyzer
( ) 8015A Digital Pulse Generator
( ) 9560B/D Automatic Circuit Tester
( ) 3760A/3761A Digital Communication Test System

Name__________

Title__________

Company__________

Address__________

City__________ State__________ Zip__________

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.

THE AUTHORS

HERBERT F. YORK ("Multiple-Warhead Missiles") is professor of physics and director of the Program on Science, Technology and World Affairs at the University of California at San Diego. He is also an associate of the Center for the Study of Democratic Institutions and a member of the Executive Committee of the Federation of American Scientists. York was graduated from the University of Rochester in 1942, receiving his master's degree there in 1943 and his Ph.D. from the University of California at Berkeley in 1949. From 1952 to 1956 he was director of the Lawrence Livermore Laboratory; he then served successively as Chief Scientist of the Advanced Research Projects Agency (ARPA) of the Department of Defense, Director of Defense Research and Engineering in the office of the Secretary of Defense and Chancellor of the University of California at San Diego. He has twice been a member of the President's Science Advisory Committee (PSAC). From 1962 to 1969 he was a member of the General Advisory Committee of the U.S. Arms Control and Disarmament Agency.

J. S. COOK ("Communication by Optical Fiber") is head of the optical systems research department of the Crawford Hill Laboratory of Bell Laboratories. "I have been with Bell Laboratories since 1952," he writes, "when I left Ohio State University with bachelor's and master's degrees in electrical engineering. My work has been with electron devices, simple classical electromagnetics and communication systems of various kinds. The alternate focus in my life is my theologian wife (truly grace), our children and our home. When we are not quite so much at home, our favorite activities are travel, skiing and tennis, in about that order; we would class ourselves as enthusiastic intermediates in each. I get into my workshop at home occasionally to build or rebuild a bit of furniture or almost anything of reasonable scope."

UGO AMALDI ("Proton Interactions at High Energies") is a member of the staff of the physics laboratory of the Istituto Superiore di Sanità, an agency of the Italian national research council; at present he is on leave to work with the European Organization for Nuclear Research (CERN). "In connection with the responsibility of our institute in the field of public health," he writes, "I have been the representative of the minister of health in the Italian commission for nuclear safety and radiation protection, which examines and licenses nuclear plants in Italy. Since 1959 I have taught a course in physics at the School of Radiology of the University of Rome. In 1971 I published a book on radiation physics directed to physicians, biologists, engineers and protectionists who need to understand the physical foundation of their work. I find that work in the field of radiation protection is interesting and useful, although for me it has meant less time devoted to 'pure' research. Apart from physics, I like skiing and classical music. With time I am becoming more and more interested in the philosophical implications of scientific culture." Amaldi was born in Rome, the son of the physicist Edoardo Amaldi, and took his degree in physics in Rome in 1957.

MANFRED M. MAYER ("The Complement System") is professor of microbiology at the Johns Hopkins University School of Medicine. Born in Germany, he was graduated from the College of the City of New York in 1938 and obtained his Ph.D. from Columbia University in 1946, which was the year he began his affiliation with Johns Hopkins. "My nonprofessional interests and activities are family-oriented," he writes. "My wife and I share a strong interest in music. Both of us play the piano, and in addition she plays the cello and I play the viola in local chamber groups and orchestras. We share an interest in art and archaeology and are amateur collectors. On vacation trips we paint primitive 'primitives.' My four sons and I used to go boating in the summer and skiing in the winter, but now that they are grown and have left home, athletics are limited to swimming, hiking and ball-playing with Jäger, my talented German short-haired pointer."

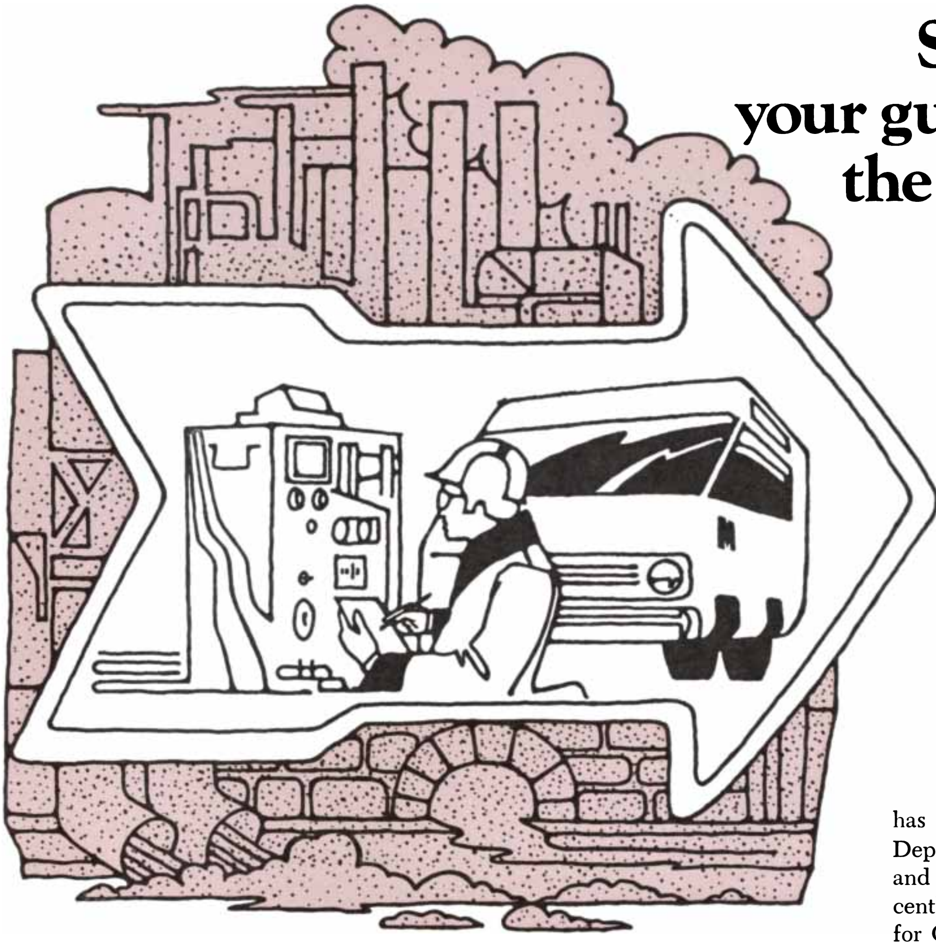
LEON D. HARMON ("The Recognition of Faces") is professor of biomedical engineering and head of the department of biomedical engineering at Case Western Reserve University. He went there a year ago after 16 years as an investigator at Bell Laboratories, where his studies included information processing in the nervous system, sensory perception and pattern recognition by computers. "My principal scientific interest," he writes, "is in understanding intelligent machines. Nothing is more fascinating than the human brain, and attempting to understand it and to build machines that aim to simulate it are the most challenging tasks I know. On the other hand, enjoying Bach, mountains, radio-controlled model airplanes, photography and other nonprofessional diversions offers strong competition."

MARTIN DAVIS and REUBEN HERSH ("Hilbert's 10th Problem") are respectively professor of mathematics at the Courant Institute of Mathematical Sciences of New York University and professor of mathematics at the University of New Mexico. Davis, who writes that he spent most of his childhood "in a slum neighborhood in the East Bronx," was graduated from the College of the City of New York in 1948 and received his Ph.D. from Princeton University in 1950. Hersh, who was graduated from Harvard College with a degree in English literature, spent four years as an office boy and editorial assistant at SCIENTIFIC AMERICAN and four years as a machinist before turning to mathematics; he obtained his Ph.D. in mathematics from N.Y.U.

MIRIAM ROTHSCHILD, Y. SCHLEIN, K. PARKER, C. NEVILLE and S. STERNBERG ("The Flying Leap of the Flea") are respectively an amateur entomologist, a zoologist from the University of Jerusalem, an investigator with the physiological-flow-studies unit at the Imperial College of Science and Technology, an insect biophysicist who has done research at Imperial College, the University of Copenhagen, the University of Oxford and the University of Bristol, and an insect physiologist at the Institute of Biological Research at Ness Ziona in Israel. Miss Rothschild, who notes that she has "published a quarter of a million words on the subject of fleas," was the author of the article "Fleas" in the December 1965 issue of SCIENTIFIC AMERICAN.

BRUCE C. HEEZEN and IAN D. MACGREGOR ("The Evolution of the Pacific") are respectively associate professor of geology at the Lamont-Doherty Geological Observatory of Columbia University and associate professor of geology at the University of California at Davis. Heezen, who was graduated from the University of Iowa in 1948 and received his Ph.D. from Columbia in 1957, has been at the Lamont-Doherty Observatory since it was founded. MacGregor, who was born in India, was graduated from the University of Aberdeen in 1957 and obtained his master's degree at Queen's University in Ontario and his Ph.D. from Princeton University.

Environmental Services: your guide through the pollution jungle.



Baffling – to management. Frightening – to stockholders. Frustrating – to industrial chemists. These are the first reactions to the gathering storm of Environmental Protection Agency and Occupational Safety and Health Act regulations. Although no one quarrels with the necessity.

Despite the best will to conform, manufacturers surveying their complex processes in finely-tuned plants ask: where do you start?

For example, take a paper mill. Its staff of scientists might be the industry's best – on any problem from pulping to coating. But how do they cope with quantifying solids, sulfur and nitrogen oxides in the power plant stacks? They can search out the know-how, but can the paper mill afford to buy them computer-operated gas samplers and a spectrometer? Or a detergent manufacturer with five plants. On quality control and new product research, the company technologists can't be beat. But micro-elemental waste water analysis, trace elements in the stack gases, and the particle count of dusts breathed at the package filling machines – these are new, unplowed fields.

Sampling, Identifying, Quantifying

The technology of pollution abatement. It starts with sampling identifying, analysis, and quantifying. It is a body of know-how completely apart from what it takes to batch-process asphalt, gin cotton, or build radial tires. And – like any other discipline – it requires its own expensive equipment.

This is the reason Monsanto offers Environmental Services. It is a special department of Monsanto Research Corporation, operating out of Dayton, Ohio. MRC's business for fifteen years has been contract research. For the Department of Defense, AEC, NASA, and private industry. And most recently for EPA and National Institute for Occupational Safety and Health – under contracts to identify hazards, establish industry sampling procedures, and to develop a number of the now standard tests for air, gases, water, particulates.

Technology Plus Sophisticated Equipment

This technology (and hardware) form the core of Environmental Services. Over the years, some eleven million dollars worth of Dayton laboratory facilities and analytical equipment have been invented, adapted, or simply bought and installed. It ranges from a tandem coupled gas chromatograph/mass spectrometer to stereo scanning electron microscopes. And much technology. This was harder won than apparatus. It includes methods for computer-controlled gas sampling, elemental analyses of tissue from a fish kill, the identification of photo-reactive constituents of plant

Monsanto Research Corporation scientists, retained on contract by various government agencies, have developed analytical methods and procured data relating to acceptable standards of operation in pollution abatement.

EQUIPMENT FOR SPECIALIZED ANALYSES

MONSANTO RESEARCH CORP. USES THIS EQUIPMENT —	FOR ANALYSIS OF THESE MEDIA						
	Gases	Mists and Vapors	Partic-ulates	Trace Metals	Hydro-carbons	Waste-water	PCB, Pesti-cides
Gas Chromatograph with Various Detectors	■	■			■		■
Spectrometers							
Mass	■	■					
GC/Mass Spec	■	■					
Emission	■	■	■	■			
X-Ray Diffraction			■				
Atomic Absorption			■	■		■	
X-Ray Fluorescence				■		■	
Ultratrace	■	■	■	■	■	■	■
Spectrophotometers							
Ultraviolet-Visible	■	■		■		■	
Infrared	■	■			■		■
Nondispersive Infrared	■	■					
Microscopes							
Scanning Electron			■				
Quantitative Television			■				
Optical			■				
Electron Microprobe			■				
Total Organic Carbon Analyzer					■	■	
Bomb Calorimeter			■			■	
Specific Ion Electrodes		■	■			■	
Orsat Apparatus	■						
Radiation Counters & Analyzers (all types of radiation)	■	■	■			■	
Samplers							
Continuous	■	■	■	■	■	■	■
Grab	■	■	■	■	■	■	■
Impaction (sizing)	■	■	■				
Isokinetic	■	■	■				
Composite				■	■	■	■

stack gases. In fact, the laboratories can handle most existing analyses — from flameless atomic absorption determinations to micron-range particulate size distribution. If necessary, precise, reliable methods are developed to meet unique analytical requirements.

On-Site Analyses

Environmental Service. It is available to you as the place to start on pollution problem-solving. You can use it at any level — from on-site, start-to-finish, effluent analyses (carried out by small specialist teams with fully equipped laboratory vans) to simply sending in your own samples for analysis. The service tells you if you have a problem, and exactly where you stand. It will supply the data needed for your National Pollutant Discharge Elimination System forms. *Firm data* will be provided you *in the form* you want it. If needed, it will provide you a *continuing* sampling, analytical or monitoring service. It's flexible.

And the cost is quite reasonable — because virtually every piece of needed equipment has been already paid for, because it would be hard to find a condition where the basic sampling and test procedures have not been encountered in earlier work. Results are reliable, they are delivered on time. And you receive a knowledgeable proposal and a cost quotation right at the start.

MRC Environmental Services will handle any physical, chemical, or biochemical analysis that's alien to your own discipline or would be a drain on your own capabilities. You can think of it as your guide through the pollution jungle. It's sound, thorough, efficient.

For more information, request a copy of:

Monsanto Environmental Services. *Write or call: Monsanto Research Corporation, Station B, Box 8, Dayton, Ohio 45407. Tel.: (513) 268-3411.*

Monsanto
the science company.

Established 1845 **SCIENTIFIC AMERICAN** November 1973 Volume 229 Number 5

Multiple-Warhead Missiles

MIRV's increase the number of strategic nuclear weapons and now threaten the stability of the nuclear balance of power. Their history shows why they present special problems of arms control

by Herbert F. York

From 1945 to 1970 the number of nuclear warheads in the U.S. strategic arsenal went from zero to about 4,000. From 1970 to mid-1975 the number will increase to almost 10,000. This increase, about half of which has already been achieved, is made possible by a single development in military technology: the multiple independently targeted reentry vehicle (MIRV).

The MIRV system enables a single rocket to launch several warheads, which can be aimed at separate targets or made to approach the same target on different trajectories. Two of its characteristics may have important consequences for military strategy. First, MIRV's are better able to penetrate an anti-ballistic-missile (ABM) system than missiles with a single warhead. Second, because MIRV warheads are numerous and can be guided with great accuracy, MIRV's could lead to an effective "counterforce" weapon, one capable of destroying a very large part of an adversary's retaliatory forces if used in a surprise attack. Counterforce weapons are usually considered the most disruptive to the strategic balance, because they threaten the deterrent on which national security is presumed to depend. A measure of how threatening MIRV's can seem was provided in August, when the U.S. Department of Defense announced that the U.S.S.R. had made its first successful test of a MIRV missile. The Secretary of Defense, James R. Schlesinger, cited the flight test as evidence that the Russians "are seeking a strategic advantage."

In actuality it is highly unlikely that MIRV's could affect the ultimate outcome of a nuclear war between the U.S. and the U.S.S.R. It seems clear that each has ample power to destroy the society of the other, with or without multiple-warhead missiles and whether or not a surprise attack destroys the opposing force. By generating fears of the other side's intentions, however, MIRV's could affect the likelihood that a nuclear war will occur. Thus in the peculiar logic of the nuclear arms race the possession of a potent weapon may diminish national security rather than enhance it. In this context I shall discuss what the MIRV system is and how it works, what its antecedents were in the U.S. arms program and why the decision to build it was made. Finally I shall examine why efforts to stop its development and deployment failed.

Few details of the operation of multiple-warhead missiles have been made public. The dimensions and mass of the individual warheads, for example, have not been revealed, nor has the maximum separation of targets been stated. A general, nonnumerical explanation of the MIRV system can nevertheless be given. It is based on statements made by officials of the Department of Defense, on testimony given before Congressional committees and on data published by nongovernment organizations such as the Stockholm International Peace Research Institute (SIPRI). In addition certain characteristics of the MIRV missile can be inferred from the operation of related devices known to have preceded MIRV's in the weapons and space programs. With this information it is possible to describe a hypothetical flight of a MIRV missile [*see illustration on pages 20 and 21*].

For this purpose let the missile be a land-based intercontinental ballistic missile (ICBM) carrying three independently aimed warheads. (Minuteman III is such a missile.) In the initial stages of flight the main rocket motors put the missile on a trajectory calculated to terminate near the first of the three selected targets. When the fuel of the last of the main rocket stages is exhausted, the final stage containing the warheads and their associated apparatus separates from the rocket and coasts on toward apogee, the point of greatest altitude. The apogee for ICBM's is typically about 800 miles.

Once the stage containing the warheads has separated from the larger stages, propulsion and guidance are provided by the post-boost control system (PBCS), the most important component of the MIRV system [*see illustration on next page*]. The final stage, consisting of the warheads and the PBCS, is often called the "bus"; it carries the warheads as passengers to be discharged at intervals. The bus has an inertial guidance system and small rockets that can modify its velocity and attitude.

After a brief period of coasting the bus refines its trajectory until it is aimed as precisely as possible at the first target. It then gently ejects one of the warheads. In satellite programs that were precursors of MIRV this ejection was accomplished with small springs in compression; the MIRV system could have a similar mechanism.

When the first warhead has been released, the bus alters its course in preparation for releasing the second. It can do this in several ways. First, the bus can increase its speed in the direction of its original orbit, causing the second warhead to impact "downrange," or farther away than the first. A decrease in speed, brought about by firing the rocket in a direction opposite to the direction of motion, will cause the warhead to land "uprange," or closer to the launching site. Another mode of deflection is perpendicular to the plane of the original trajectory. Impulses in this direction will aim the warhead at targets in an arc on each side of the initial target. Finally, by giving the bus an impulse in the plane of the original trajectory but roughly perpendicular to its direction, the second warhead can be aimed at the same target as the first. The second will approach from a higher or lower angle, however, and its arrival will be delayed by as much as several minutes. Ordinarily there would be movement on all three axes.

When the second weapon has been released, the bus is reoriented once again, this time on a new trajectory terminating at the third target. If more than three warheads are incorporated in the bus, the process is repeated until all have been launched at their targets. Several kinds of decoys and other "penetration aids" might also be released with some or all of the warheads.

Numerous engineering compromises are necessary in the design of the bus. The maximum separation between the targets, for example, depends primarily on the total impulse available from the bus propulsion system and therefore on the payload weight allotted to it. MIRV's now in use by the Navy and the Air Force can evidently reach targets separated by distances on the order of a few hundred miles, or a few percent of the total range of the missiles.

A more obvious compromise is that between the number of warheads and their aggregate explosive power, or "yield." Some yield is always lost in the transition from single to multiple warheads. Even in multiple-warhead missiles that do not provide separate guidance for each reentry vehicle much weight, and therefore yield, must be sacrificed to extra heat-shielding for reentry into the atmosphere and to the diseconomies of smaller scale.

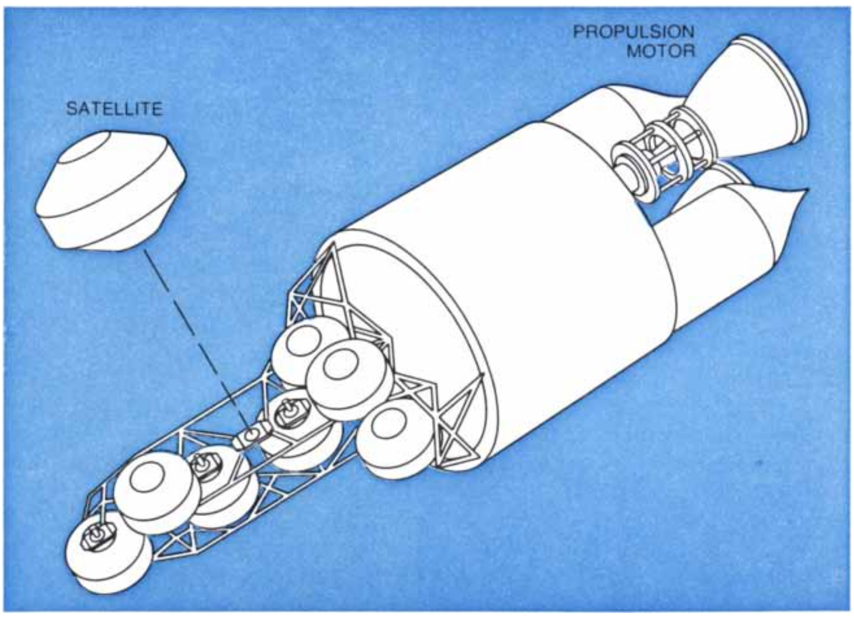
For true MIRV's the ratio is even less favorable, because the weight of the post-boost control system as well as that of the shielding must be subtracted from the payload available for the weapons themselves. The engineering is also more complicated, because the missile-maker has more options: he can choose to emphasize wider separation of targets, higher multiplicity of warheads, greater overall range, more penetration aids or higher yield. Presumably, optimum compromises have been found for all these choices; they depend on the nature of the target being attacked and on the defenses to be overcome.

The Department of Defense has not disclosed the yield of the warheads carried by the U.S. missiles now equipped with MIRV's. Estimates have been made, however, that for the purposes of this discussion are sufficiently reliable. (The figures given here are those published by SIPRI.)

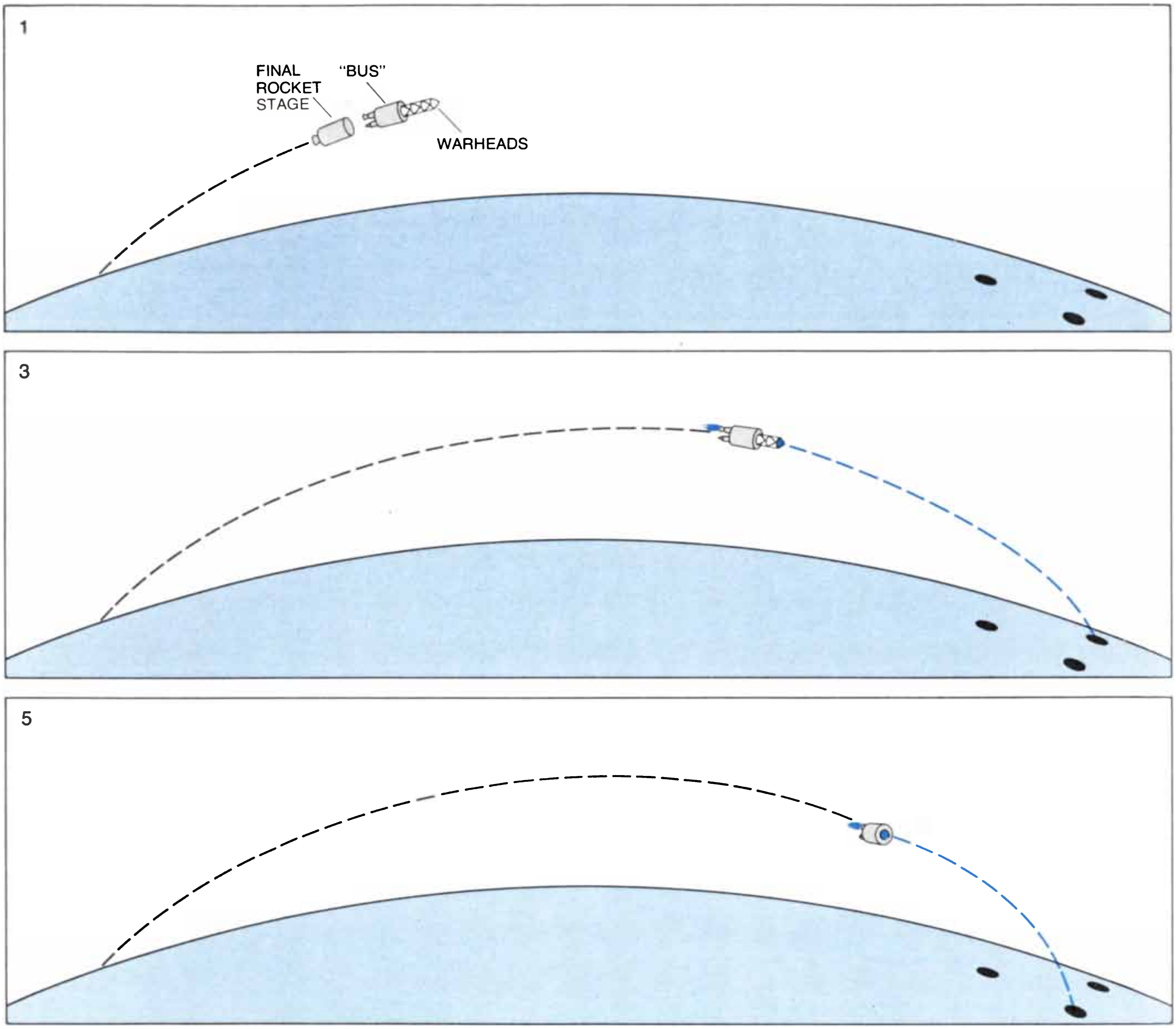
Even if the precise yield of a warhead were known, it would be difficult to calculate the amount sacrificed in converting to multiple warheads, since in each case MIRV's were introduced in new missiles that had never carried single warheads. The Navy's Poseidon, for example, is a larger missile than the Polaris it is replacing, and it has twice the payload. Nevertheless, it is estimated to have a smaller yield. The last Polaris to carry a single warhead is reported to have had a yield of about one megaton. The Poseidon is usually said to carry 10 warheads of 50 kilotons each, for a total yield of about half a megaton.

A better indication of the compromises required may be given by the Minuteman III, an Air Force ICBM. The single-warhead versions of Minuteman had yields estimated at from one megaton to two megatons; Minuteman III is said to have three independently aimed warheads of 200 kilotons each, suggesting again that the aggregate yield is one-half or less.

It is important to note that a reduction in total yield does not necessarily imply a reduction in destructive effect. Megatonnage and "throw weight" are not reliable indicators of the destructive capability of nuclear weapons. A better measure is the circular area within which a given warhead will cause some specified degree of damage; this in turn is a function of the "blast overpressure." For overpressures high enough to be effective against military targets the radius of destruction increases roughly as the cube root of the explosive yield; the area of destruction, therefore, increases as the two-thirds power of the yield. For example, if a one-kiloton device could destroy a particular target within an area of one square mile, then to produce



TRANSTAGE, a post-boost control system used with the Titan III booster rocket, is shown carrying defense communication satellites, which it was used to launch beginning in 1966. The satellites are mounted in a tubular frame and are ejected by small springs. The Transtage was the immediate predecessor of the MIRV system used in the Minuteman III ICBM and has been said to incorporate all the essential technology of a MIRV warhead.



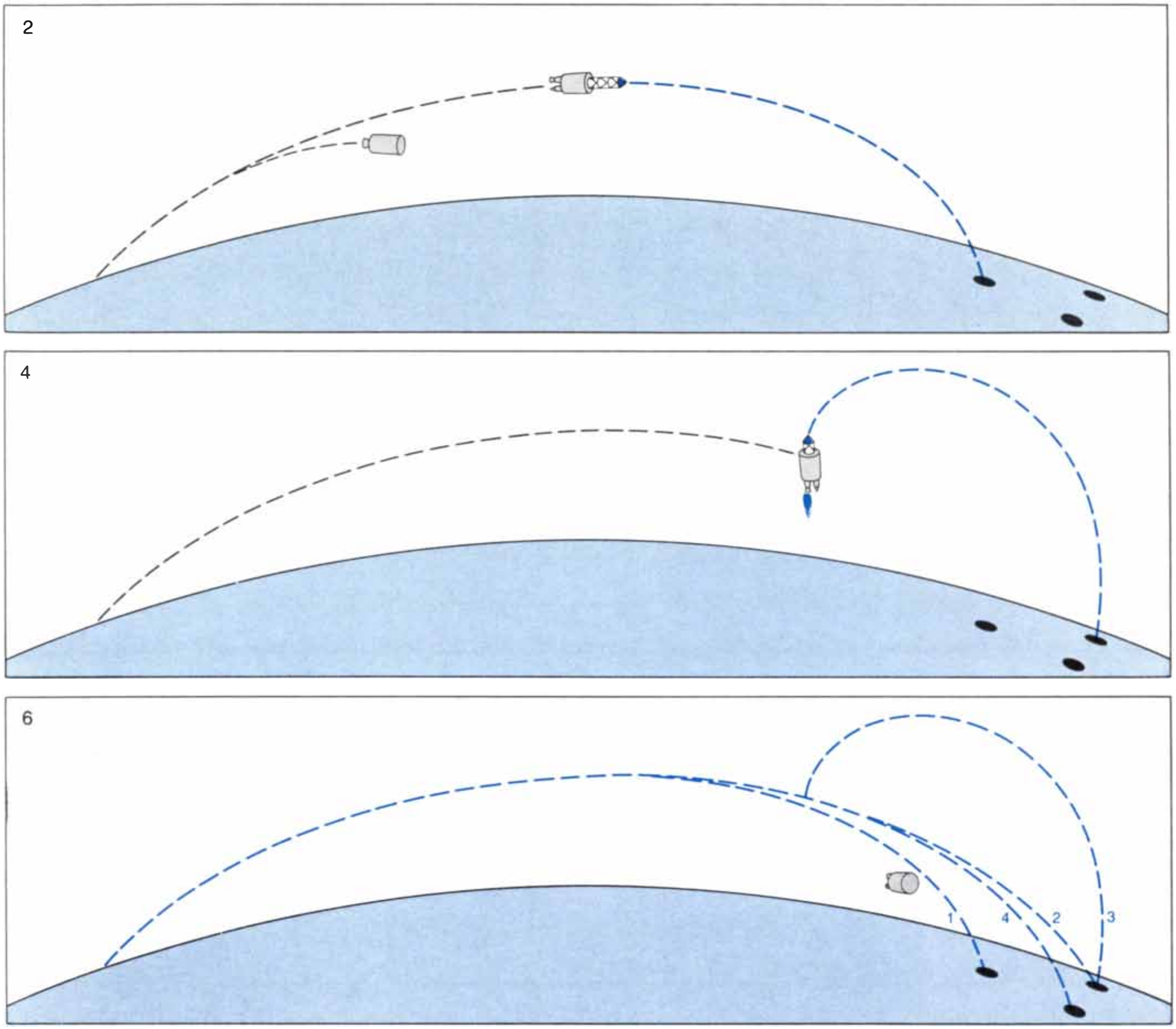
POST-BOOST CONTROL SYSTEM, or "bus," is at the heart of MIRV; its operation is illustrated schematically in these drawings, which show a hypothetical flight of a bus carrying four warheads. When the last of the main rocket stages is exhausted (*1*), the bus, with the warheads attached, separates from it. After coasting briefly the bus adjusts its trajectory until it is aimed as accurately as possible at the first of the selected targets. The first of the warheads is then gently ejected (*2*) and will continue to follow the ballistic trajectory to the target. After firing its rocket motor to add an increment of speed the bus ejects another warhead (*3*), which will

an area of destruction of two square miles would require a warhead of about 2.8 kilotons. Because of this exponential relation the potential for destruction is greater with many small weapons than it is with a single large one of the same total yield.

The history of how multiple-warhead missiles came to be developed in the 1960's and how they came to be deployed in the early 1970's provides an interesting lesson in the structure and operation of the military and military research organizations. The technology necessary for MIRV's evolved from research directed toward several independent and quite disparate goals. Ideas and personnel were exchanged among the various programs, so that the course of development became not a thread but a fabric. It could have been cut in any number of places without seriously impeding the progress of the MIRV system.

Once the technology was developed MIRV assumed a momentum of its own; the chances of halting it were by then slim. In addition a number of quite different arguments were presented in favor of deployment, and apparently any one of several might have been sufficient to gain the necessary Congressional and Department of Defense approval. Many of the development decisions could have been countermanded on several occasions and the result would have been about the same: MIRV's on U.S. missiles at the beginning of the 1970's.

A convenient moment at which to begin an examination of the history of MIRV's is the launching of the first Russian artificial satellite in October, 1957. About a month after the satellite went into orbit, and at least partly in response to the launch, William M. Holaday, director of guided missiles in the Office of the Secretary of Defense, established

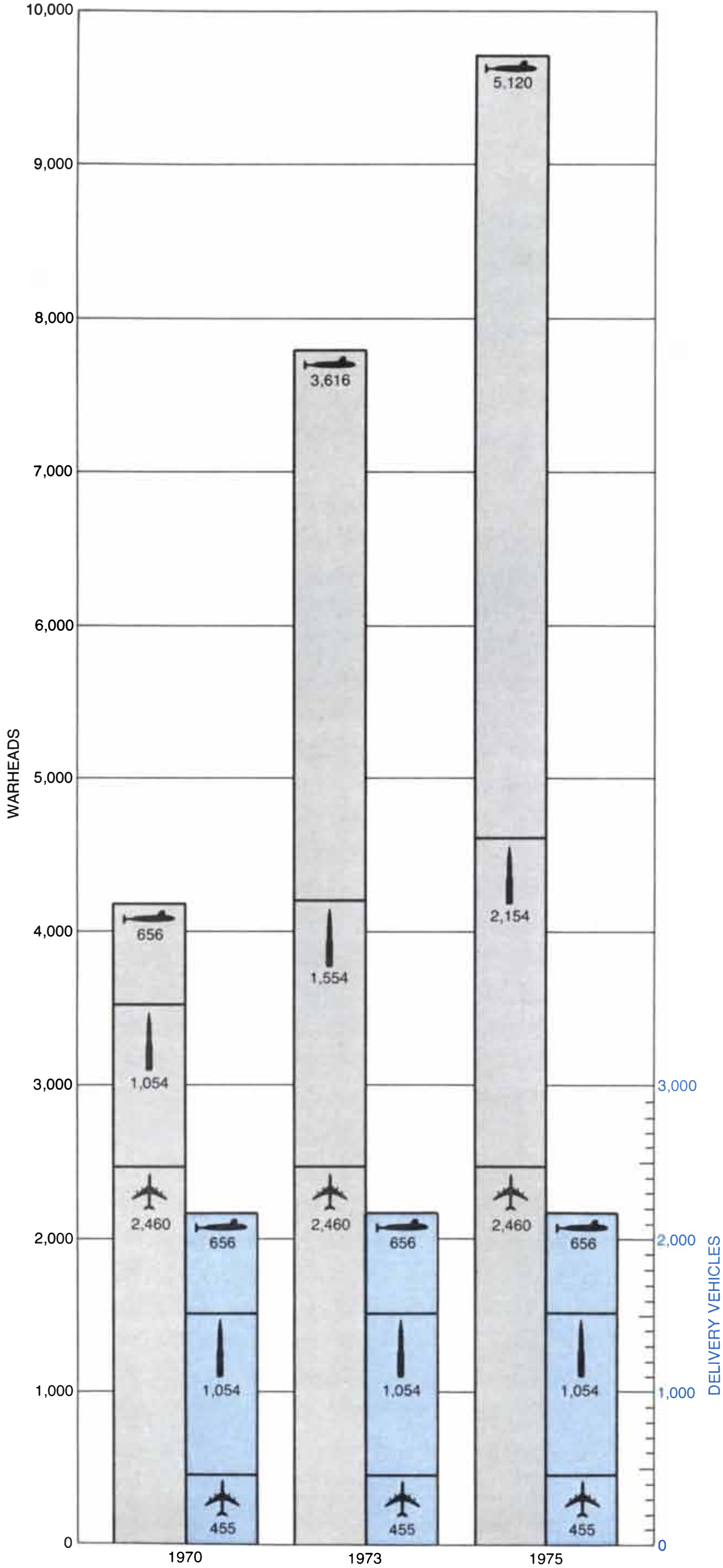


strike a target farther downrange. The bus next increases its velocity in a direction roughly perpendicular to the direction of the trajectory (*4*), placing the third warhead on a path that will lead it to the same target as the second but will delay its arrival by as much as several minutes. The last warhead is released after the bus has executed a final maneuver, adding a velocity increment in the plane that in this illustration is perpendicular to the page (*5*). Many other combinations of these movements are possible. Finally, the bus disintegrates on reentering the atmosphere and the warheads reach their assigned targets (*6*), although not simultaneously.

the Reentry Body Identification Group, with representatives from several agencies of the Department of Defense, from industry, from academic research departments and from such consulting firms as the Rand Corporation. The committee was formed to determine whether or not the designers of offensive ballistic missiles should consider seriously the possibility that defensive missiles might be built by the opposition. In early 1958 the committee reported that missile defense should be given consideration; it also described, however, a number of countermeasures available to the offense.

All but one of the countermeasures proposed by the Reentry Body Identification Group were intended to confuse the radars of the defenders. They included decoys, objects that to radar would resemble a warhead; booster fragments, pieces of the rocket-motor fuel tanks used as decoys but available at no weight penalty; chaff, small lengths of wire dispersed in space to act as a radar reflector; a reduced radar cross section for the warhead, and radar blackouts produced by exploding thermonuclear devices in the upper atmosphere. The remaining proposal, and the most important for this discussion, was the use of multiple warheads. Rather than confuse the defense, multiple warheads simply exhaust it. These countermeasures and others are what are collectively called penetration aids.

Decoys, even cheap and light ones, can approximate the characteristics of a warhead as long as both the decoys and the real weapons are in space, where all objects, regardless of mass and shape, follow ballistic trajectories. Once the decoys enter the atmosphere, however, they are slowed by air resistance to a greater extent than the heavier warheads, and the difference in velocity becomes progressively greater as the reentry bodies reach denser regions of



the atmosphere. Booster fragments, chaff and such light decoys as balloons covered with metal foil will disintegrate and burn up in the upper atmosphere. Thus the differential effects of the atmosphere on the reentry bodies enable the defense to discriminate between decoys and weapons (if it is willing to wait until well after they have entered the atmosphere) and to allocate defensive missiles only to the real weapons.

The solution for the offense is to use heavier decoys that will mimic the flight characteristics of real warheads to lower altitudes. As the weight of the decoys approaches that of the warhead itself, however, it becomes more efficient simply to use several warheads.

This was in fact the strategy adopted for the Polaris A-3, the last of the submarine-launched ballistic missiles (SLBM's) in the Polaris series. The A-3 is a multiple-reentry-vehicle (MRV) system with three warheads. The cluster is launched as a unit, on a trajectory chosen to guide it as accurately as possible to the single target. When the rocket has burned out, the three weapons are separated and given small additional impulses. They continue on separate but close trajectories and impact in a triangular pattern, presumably centered on the target. The missile might be compared in principle to a shotgun.

The dimensions of the triangle formed by the impact points of the warheads have never been made public, but the military situation as it was perceived in the late 1950's set obvious limits. The separation had to be more than a few tenths of a mile or all three warheads could have been destroyed by the explosion of a single antimissile missile. On the other hand, they could be no farther

NUMBER OF WARHEADS in the U.S. strategic force will increase almost two and a half times in the first half of the 1970's, even though the number of delivery vehicles (bombers, ICBM's and SLBM's, indicated in the graphs by symbols) will remain constant. The statistics are those published by the Stockholm International Peace Research Institute. The graph marked 1970 represents the force level before any MIRV's were installed; 1973 bars are for early in that year; figures for 1975 show the number of warheads when all Poseidon and Minuteman III MIRV's now planned are deployed. In these charts it is assumed that the Minuteman III carries three warheads and the Poseidon 10 warheads. MRV's are considered single warheads since each MRV assembly could be directed to only one target. The figures given for weapons carried by bombers are approximate and can change quickly.

apart than a few miles or the dimensions of the pattern would have exceeded the size of most cities in the U.S.S.R. The yield of each warhead is estimated to have been about 200 kilotons.

The first Polaris A-3's were deployed in 1964. Even before they joined the strategic-weapons force it was recognized that the MRV warhead would not be able to cope with improved ABM systems. By 1962 or 1963 progress in U.S. defensive missiles and the knowledge that Russian defenses were also being improved made it clear that the separation of warheads in the A-3 MRV was much too small to thwart any but a primitive ABM. Moreover, it was seen that the solution could not be found by increasing the horizontal spread of the impact points; to do so would make the pattern larger than most targets. Another solution was soon found: the MIRV.

The launching of the first Russian satellite and the launching of the first Russian ICBM (which had preceded the satellite by two months) stimulated in the U.S. an outburst of ideas about how to make and use satellites and missiles. The use of penetration aids, as we have seen, was one of the results of this process. Another was the concept of launching more than one satellite with a single rocket.

The earliest proposal of multiple satellite launchings of which I am aware was directed to defense against missile attack. This was the ballistic antimissile boost interceptor (BAMBI). The objective was to intercept an enemy's missiles during the first few minutes of flight, while the booster motors were still operating. Missiles were thought to be particularly vulnerable during this period because simply puncturing their propellant tanks could make them fall thousands of miles short of their target.

Two versions of BAMBI were proposed; both would have placed large numbers of satellites in orbits from which missiles could be detected and destroyed in the early stages of flight. One version, the random-barrage system, would have used many thousands of small satellites, launched by, say, hundreds of boosters. The other, called space patrol air defense (SPAD), would have deployed the small "killer" satellites in a "mother ship" equipped with central guidance and detection devices. On command the mother ship would have oriented itself and determined when, at what rate and in what direction to launch its subsatellites.

Neither of these systems was built, but they were studied on paper by the Advanced Research Projects Agency (ARPA) of the Department of Defense with help from the Rand Corporation. During the early years of ARPA (it was founded in 1958) there was an active interchange of technical personnel between the agency and industry and between the various companies most heavily involved in missile and space technology. As a result many of those who helped to fashion these early proposals were later members of the organizations that designed real multiple-warhead devices.

A quite unrelated development whose basic technology was later adapted to MIRV's was the Able-Star, a second-stage vehicle designed to be used with the Thor booster. It was the first spacecraft where the main propulsion rocket could be shut off and later restarted. The Able-Star used hypergolic propellants (substances that ignite on contact) and incorporated restart, guidance and control devices, a programmer and an accelerometer—all necessary to the operation of MIRV's.

The Able-Star was first tested in space in April of 1960. Two months later it was used in the first multiple satellite launch, in which a Transit II-A satellite and a Naval Research Laboratory solar radiation satellite were placed in near-circular orbits 500 miles above the earth. Once the Able-Star achieved the proper orbit the satellites were detached and separated by a compressed spring, giving the smaller satellite an additional velocity of 1.5 feet per second.

In a subsequent launch the Able-Star was used to place three satellites in similar orbits, although the procedure was only partly successful. In 1963 the Atlas-Agena rocket was used in a more difficult maneuver: placing a pair of satellites in very different orbits. Later versions of the Agena second stage, like the Able-Star, could be stopped and restarted during flight. The satellites, called Vela, were used to monitor compliance with the Limited Test-Ban Treaty of 1963. They were placed 180 degrees apart in orbits from 62,000 to 72,000 miles high.

The immediate technological ancestor of the Air Force version of MIRV was Transtage, a highly flexible post-boost control system. It was crucial in the development of the components and techniques used in MIRV's, yet it was devised for reasons unrelated to the effort to improve missiles and missile warheads.

Transtage was used with Titan III, which in the early 1960's was the largest of the U.S. booster rockets. Transtage had a propulsion system capable of coasting and restarting, like the Able-Star and the Agena, but it carried a larger payload and was capable of more complex and more extensive maneuvers. It was conceived without a specific mission in mind, and it was first used to launch a series of defense communication satellites called IDCSP (for initial defense communication satellite program).

The special requirements of defense communication demanded that the satellites be many and that their orbits be quite high. On June 16, 1966, a Titan III-C and Transtage placed eight 100-pound satellites in eight different equatorial orbits, all at an altitude of about 21,000 miles.

The operation of Transtage was comparable in almost all respects to that of the MIRV bus. Using its ability to coast and restart, it first achieved a near-circular orbit at the proper altitude with a period of 1,334.2 minutes. It gently nudged off one of the subsatellites with compressed springs. Then, with four vernier motors of 50 pounds' thrust (whose main purpose was controlling pitch and yaw), it added a small increment of velocity and ejected a second satellite.

This one would orbit at essentially the same altitude, but with a period of 1,334.7 minutes. The maneuver was repeated for each satellite, until the last was dropped off three minutes after the first in an orbit with a period of 1,347.6 minutes.

Three more groups of IDCSP satellites were launched, using the same technique, at intervals of about six months. The importance of the program in the development of MIRV's was indicated in 1968 by John S. Foster, Jr., the director of Defense Research and Engineering. Asked during his testimony before a Senate subcommittee why he was so confident the Minuteman III and Poseidon MIRV's would work, he cited the successful operation of Transtage as proof that all the essential engineering problems had been solved.

Another way to pursue the relation between these projects and the later MIRV missiles is through the contractors who produced them. For example, Agena was designed by the Lockheed Missiles and Space Company (a subsidiary of the Lockheed Aircraft Corporation), which later designed the Poseidon missile. Similarly, systems engineering for the Titan III and Transtage was done by the Aerospace Corporation, which later did the concept engineering for the

Minuteman III MIRV. Another company involved in the MIRV program for Minuteman was the Space Technology Laboratory of the Thompson Ramo Wooldridge Corporation, which had earlier participated in the Able-Star and Vela satellite programs. In the Air Force itself the Space and Missile Systems Office, which supervised the Transtage program, was soon to begin development of the Minuteman MIRV.

In addition to all these programs, bits and pieces of MIRV technology were invented or reinvented independently in the course of unrelated endeavors. One of them was a study by the Rand Corporation of what advantage the U.S.S.R. might gain from its relatively large missiles. Among other conclusions of the report was the possibility that such missiles might be used to deliver multiple warheads; a MIRV-like device was hypothesized as the means for doing so. Another was a study of orbital offensive weapons, conducted for the Air Force by seven competing firms. In an orbital bombardment system nuclear weapons would be placed in permanent orbit and brought down on their targets from space on receipt of coded instructions. Some of the proposals involved what was essentially a one-passenger bus. The system was never developed, but if it had been, the technology of MIRV's might well have been derived from it.

Another program that could have been used as a point of departure for MIRV was the sequential-payload-delivery system (SPD), used to deliver unarmed warheads from California to Kwajelein Atoll in the Pacific, where they served as test targets in the ABM program. The justification for this system was economic: it is cheaper to attack several targets with a single rocket than to use a separate rocket for each target. The design and construction of the sequential-payload-delivery system was supervised by the Aerospace Corporation, which had been responsible for Transtage.

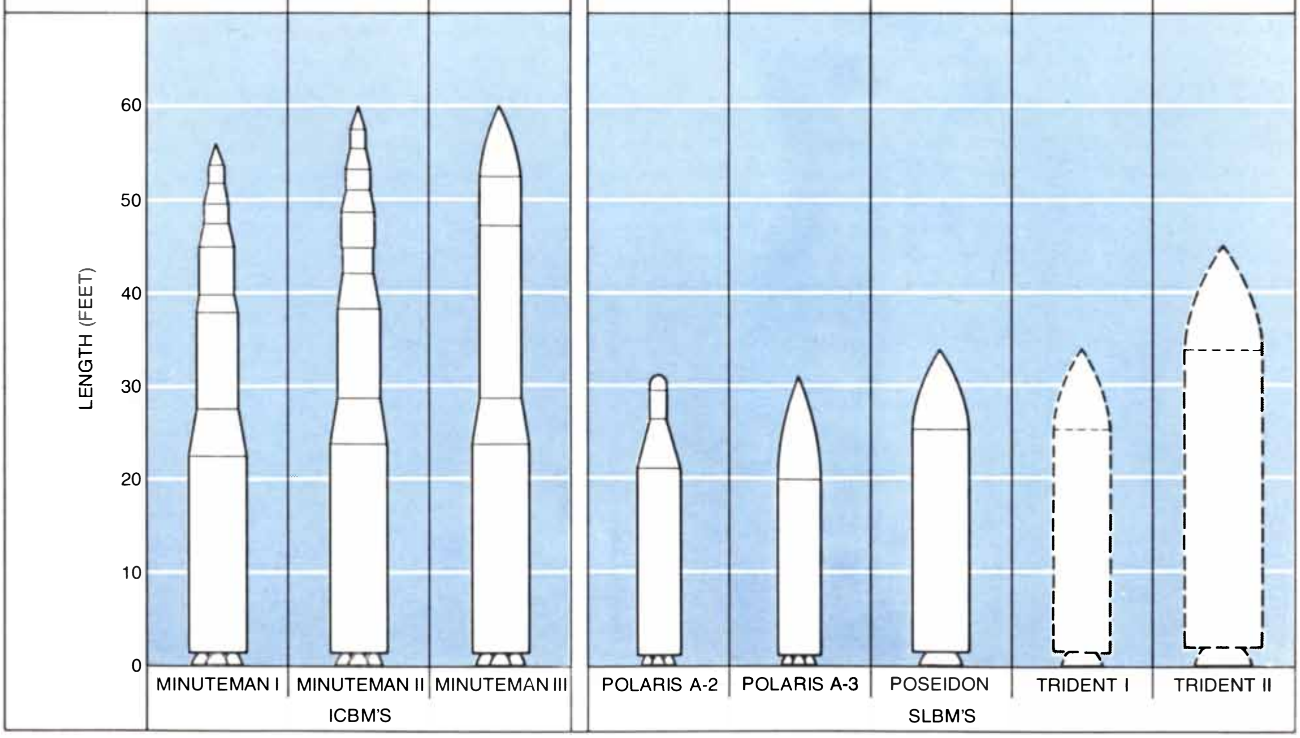
Almost all the mechanisms and techniques used in MIRV's could also have been derived from the civilian National Aeronautics and Space Administration manned space program, particularly the systems used in lunar exploration. I have described here, however, only those programs that were addressed to some military purpose.

By the mid-1960's it was clear to military planners that a MIRV missile could be built. Even before then, in 1962 and 1963, two independent arguments had been put forward in support of deployment.

One was embodied in the "counterforce" speech given by Secretary of Defense Robert S. McNamara in Ann Arbor, Mich., in 1962. The notion of counterforce did not begin with McNamara; it was a part of military strategy before nuclear weapons were invented. It holds that one should plan to attack an adversary's weaponry, the "counterforce targets," rather than his cities and industries, the "countervalue targets." It is the strategy usually adopted by those who favor expanded deployment of weapons, and it is necessarily the strategy of those who contemplate making a preemptive surprise attack.

McNamara subsequently modified his position on counterforce, but the speech nevertheless stimulated the proposal of MIRV's as a means of increasing the number of targets that could be attacked. Indeed, there can be no doubt that in a counterforce strategy MIRV is a powerful weapon. In a preemptive at-

	MINUTEMAN I	MINUTEMAN II	MINUTEMAN III	POLARIS A-2	POLARIS A-3	POSEIDON	TRIDENT I	TRIDENT II
WARHEADS	1	1	3 MIRV	1	3 MRV	UP TO 14 MIRV	UP TO 24 MIRV	UP TO 24 MIRV
YIELD (KILOTONS)	1,000	1,000–2,000	600	800	600	500	NO DATA	NO DATA



MIRV'S ARE DEPLOYED on two of the six U.S. missiles that are now operational, the Air Force Minuteman III and the Navy Poseidon. Although both have larger payloads than the missiles that preceded them, the total yield of their warheads is less. The Trident I, which could carry as many as 24 warheads, is expected to be introduced by 1978; the Trident II is planned for the mid-1980's.

tack almost perfect system reliability is essential, since the failure to destroy even only a few of the enemy's missiles would result in great harm to the aggressor. Because the reliability of individual missiles is not likely to even approach 100 percent, the first-strike force must be substantially larger than the force being attacked. Even if the adversary offers no defense, several warheads must be allotted to each of his missile silos. If missiles with single warheads are used, the military and economic advantage is clearly with the defender.

MRV's offer no advantage over single warheads; they merely scatter weapons over the target area and give no assurance of hitting a small, specific site. MIRV's, however, substantially improve the chances of the offense. With three or more warheads per missile, designed so that each can be assigned a separate target, a nation could attack an opponent's fixed-base forces without first enlarging its own arsenal. By aiming many weapons at each silo it could greatly increase the probability that all the missiles would be destroyed.

MIRV's also satisfy another major requirement of a first-strike force: high accuracy. If a missile is to be destroyed inside a hardened silo (a buried tube armored with steel and concrete), a warhead must be exploded at quite close range, perhaps even within the "cratering radius" of the weapon. Because the radius of destruction against a hardened target increases very slowly with increasing yield it is generally considered more profitable to improve accuracy than to increase explosive power.

The customary measure of accuracy is the "circular error probable," a circle centered on the target with a radius such that half of the warheads aimed at the target will fall outside the perimeter. To be useful as a counterforce weapon a warhead should have a radius of destruction against its particular target somewhat larger than the circular error probable. For example, if a given warhead has a circular error probable of one mile and a radius of destruction against a given silo of half a mile, far fewer than half of the warheads could be expected to destroy the missiles they are aimed at. To achieve 50 percent reliability the yield could be increased eightfold; the same result could be obtained by a twofold improvement in accuracy.

The second argument favoring the deployment of MIRV emphasized its ability to penetrate missile defenses. Russian nuclear tests at high altitude in 1961 and 1962 and Premier Khrushchev's famous boast that "You can say our rocket hits a fly in outer space" led many U.S. planners to ascribe to the Russian ABM performance at least as good as that which could in theory be achieved by the U.S. system. In particular it was concluded that a single large antimissile missile could simultaneously destroy all three warheads of the Polaris A-3.

For the Navy there were at least two additional motives for building MIRV's. First, the Polaris A-3 development program was nearing the end, and it was almost automatically assumed that there would be a new generation of sea-launched missiles. Second, improvements in submarine navigation for the first time made high accuracy possible in SLBM's. (In order to launch a missile with a specified accuracy the submarine must know its own position and orientation to at least that degree of accuracy.)

The new missile was at first designated the B-3, but the name was later changed to Poseidon. From the beginning it was evident that it would be a more powerful rocket than the A-3. The increase in power was obtained largely by making the rocket bigger.

At the outset it was not at all clear that MIRV's would be used in the Poseidon. The desire to compete with the Air Force in the counterforce role led some on the Navy staff to favor a large, accurate, single warhead, whereas others, including Harold Brown, who was then the director of Defense Research and Engineering, sought continued emphasis on MRV's as a means of ABM penetration.

The Navy Special Projects Office soon proposed a compromise, a device that was never built but that served as a link between MRV and MIRV. In this concept a bus would have delivered several warheads to a single target but on trajectories that had different apogees. Each warhead would be aimed at the target as accurately as guidance technology would allow; they would arrive at different times, however, and would be spaced as much as 100 miles apart. They would have the same capacity for exhausting a newer, more sophisticated ABM as the Polaris MRV had against a crude ABM. Such a delivery system requires a bus capable of making modest changes in its velocity in the plane of the trajectory and roughly perpendicular to its direction. To transform this proposed bus into one suitable for MIRV's it is necessary only to provide a means of adding small increments of velocity in the other two directions.

Continuing progress in missile guidance and anticipation of even more improvement soon led the Special Projects Office to propose the high-multiplicity MIRV deployed today. The Poseidon is almost twice as heavy as the Polaris A-3 and carries about twice the payload. Its nominal range is about the same, 2,500 miles, but this could be increased by sacrificing payload. It is reported to be capable of carrying as many as 14 warheads, but in most analyses it is usually considered to carry 10. The warheads are estimated to have a yield of 50 kilotons each.

The first boatload of Poseidon missiles was deployed at sea in April, 1971, aboard the missile submarine *James Madison*. Converting from Polaris to Poseidon requires the installation of new launching tubes to accommodate the greater length and girth of the new missile. At least 17 boats have been converted so far, and eventually 31 of the nation's 41 missile submarines will carry Poseidons. Thus if each missile does indeed carry 10 warheads, the number of independently targeted reentry vehicles in the nation's sea-launched fleet will increase from 656 to 5,120.

Although the Poseidon will not be fully deployed for another two years, its successor is already being planned. It is the Trident, a still larger missile that in its final version will have a range of 6,000 miles. Each missile will be armed with as many as 24 independently aimed warheads. The Navy intends initially to build 10 Trident submarines, carrying 24 missiles each.

In the Air Force the argument over single v. multiple warheads persisted into the mid-1960's. Concerned, like the Navy, over progress in missile defenses and over the Russian high-altitude test series, the Air Force began to consider MIRV's as a means of improving penetration. At the same time a steady increase in the perceived number of military targets in the U.S.S.R. stimulated interest in the use of multiple warheads to improve force effectiveness.

On the other hand, the Air Force had for some time been committed to large, single warheads protected against interception by decoys, chaff and other penetration aids. At one point the factional dispute reached the missile-industry press when a contract for the Minuteman III warhead was delayed following the intercession of the director of Defense Research and Engineering.

Early in the debate two versions of MIRV were considered. One employed a bus virtually identical with the Transtage already under development. In the alternative system a rocket booster would have launched a cluster of small, single-stage missiles, each incorporating a propulsion and guidance system. After the cluster as a whole was placed on a trajectory leading to the area of the targets the individual missiles would adjust their velocities separately to impact on their particular targets. The extra weight and extra cost of the individual guidance and control systems made this proposal less attractive than the bus, and its development was never authorized.

The bus type of post-boost control system was selected for Minuteman III in 1964; the first flight of 10 missiles was turned over to the Strategic Air Command in June of 1970 at Minot Air Force Base in North Dakota. Minuteman III carries three warheads, usually said to have a yield of about 200 kilotons each.

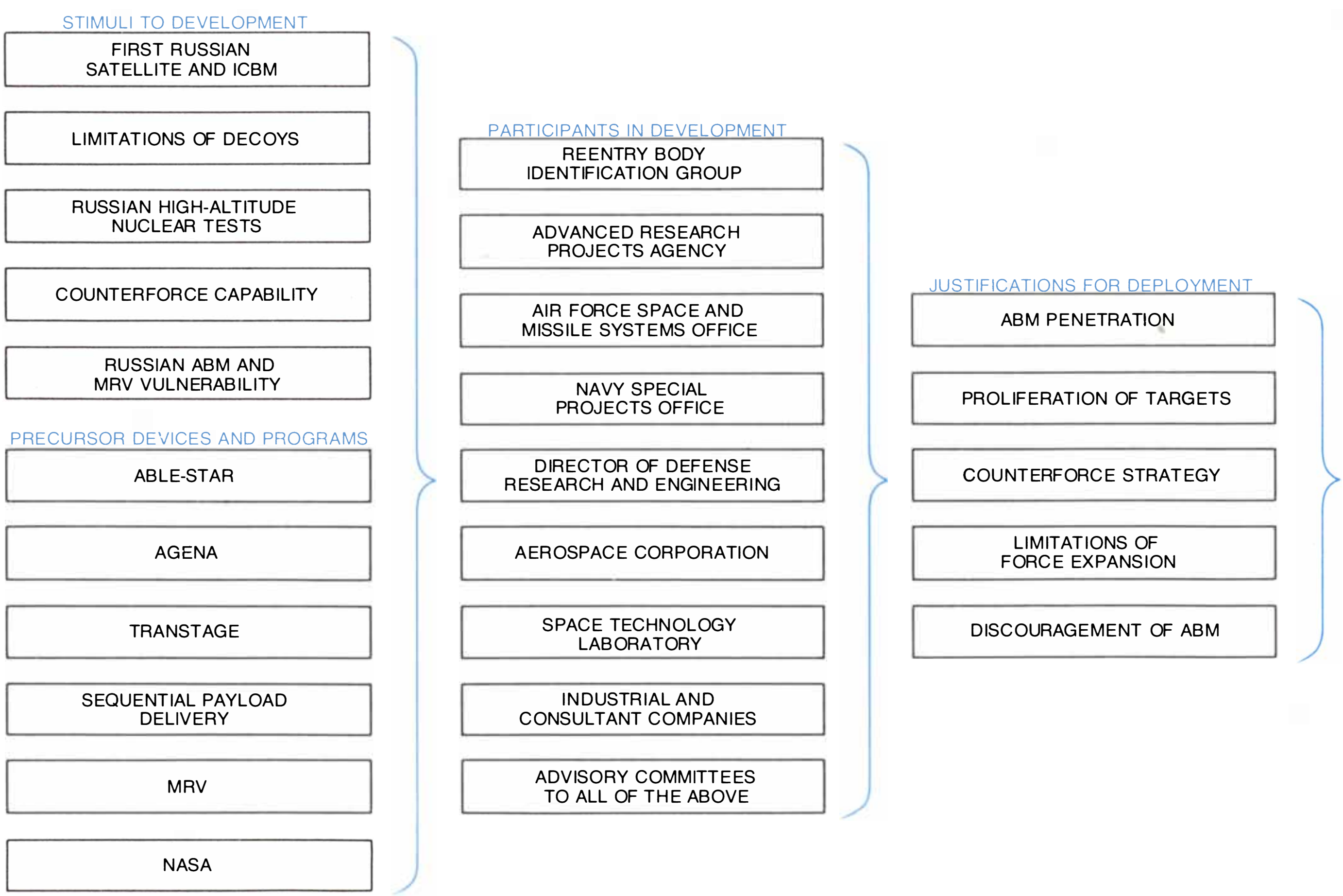
The decision to deploy MIRV's was made all but inevitable by the decision to develop them. Weapons systems, once proved feasible, assume a momentum of their own, under the rationale of "If it can be done, it must be done, because if we don't, they will."

Even so, the deployment of MIRV's was debated after the development decisions were made, and the matter was not settled until the missiles were in place in 1970 and 1971. As we have seen, the arguments usually cited in favor of MIRV deployment are improved penetration of missile defenses and an increase in the number of targets that could be attacked. The penetration of an anti-ballistic-missile system was stressed by those who perceived a tradition of defensive measures in Russian history and who therefore believed that the U.S.S.R. would construct such a system. The proliferation of potential targets was emphasized by those who advocated a counterforce mission for U.S. nuclear weapons.

Secretary McNamara had other, primarily political motives for the deployment of MIRV's. Multiple warheads, he contended, offered a less costly way than the addition of more missiles to expand the strategic force and maintain at least some counterforce capability against growing Russian forces. Thus the potential powers of MIRV's were invoked in the arguments of McNamara and his staff against strategic-force expansion.

McNamara also mentioned MIRV's in arguing against deployment of missile defenses. He doubted that the proposed anti-ballistic-missile network would work and believed it might bring on a new cycle in the arms race. He opposed its deployment in the U.S. and tried to persuade Premier Kosygin (at the conference in Glassboro, N.J.) that the U.S.S.R. also should forgo antimissile systems. A U.S. commitment to deploy MIRV's was among his arguments, since MIRV's represent a relatively inexpensive means of overcoming any conceivable antimissile system. Thus, from the point of view of McNamara and some of his immediate associates in the Office of the Secretary of Defense, the deployment of MIRV's could benefit the cause of arms control.

Indeed, the limitation of ABM systems achieved as part of the agreements made in the strategic-arms-limitation talks (SALT I) of 1972 might in part be attributed to the existence of MIRV's. It could be argued that the U.S. and the U.S.S.R. were willing to renounce extensive antimissile systems because MIRV's



MULTIFARIOUS PATHS OF DEVELOPMENT, culminating in MIRV warheads, are suggested by a chart showing some of the motives for development, the projects and devices that preceded the weapons, the organizations that participated and the justifications proposed for deployment. The elimination of any one, or even several, of these programs would not have halted MIRV, since develop-

ensured the futility of building them. MIRV's are themselves excluded from the regulation of the SALT I agreements. They are considered a "qualitative" refinement in weaponry and therefore are not subject to restriction. They are expected to be a major topic in the second round of the talks (SALT II), scheduled to conclude by the end of 1974.

If McNamara believed that the deployment of MIRV's could slow the arms race, officials of the U.S. Arms Control and Disarmament Agency saw it differently. As early as 1964 Herbert Scoville, Jr., and George W. Rathjens noted ways in which the deployment of MIRV's could upset the "balance of terror." In particular, they predicted that an adversary could construe the deployment of MIRV's as possible preparations for making a preemptive strike. If a MIRV force is capable of destroying an adversary's ICBM's, that is, if it is an effective and reliable counterforce weapon, it could be considered a threat to the nation's deterrent and to that extent would be provocative. Of course, even a perfect MIRV force could not make a first strike a rational policy, since sea-launched missiles and perhaps bombers would still be able to retaliate, but MIRV's do contribute some increment of instability. For example, in a crisis where nuclear war seemed imminent MIRV's might be perceived as giving an advantage to the aggressor and therefore could encourage a first strike.

Whether or not MIRV's are now sufficiently accurate and reliable to serve as counterforce weapons, they are considered menacing by those responsible for military planning. The MIRV missiles tested in August by the U.S.S.R., for example, were described by Secretary of Defense Schlesinger as leading the U.S.S.R. to "a clear advantage in counterforce capability." Under the SALT I agreement the U.S.S.R. is allowed about 25 percent more offensive missiles than the U.S.; with the deployment of MIRV's, Schlesinger suggested, this superiority in launchers and throw weight could be "married" to equality in technological sophistication.

During most of the period in which MIRV's were under development the program was kept secret and the controversies it entrained were unknown to the public and to many members of Congress. It did not become a political issue until late in 1967, when public debate over deployment began. In the 1968 Presidential campaign Senator Eugene McCarthy echoed the view of the Arms Control and Disarmament Agency, noting that "the introduction of sophisticated anti-ballistic-missile systems and new missiles equipped with multiple warheads threatens to make the situation unstable. With the deployment of such weapons systems, each side will become concerned as to whether in the event of a preemptive attack it will be able to inflict sufficient damage in retaliation—if not, its deterrent will not be credible. The arms race will thus be impelled to a new intensity. In crises, there could be an incentive to launch a first strike."

Later Senator Edward W. Brooke introduced a resolution calling for the suspension of testing of multiple-warhead missiles. Some other senators and a number of other people supported the resolution, but the effort was in vain. The first deployment of MIRV's took place while the issue was being debated.

The MIRV program had many roots and branches. Decisions were made by many people, some only loosely connected with one another, over the course of more than a decade. Of all the stimuli that led to the development of MIRV the most important was the perceived need to penetrate ABM systems whose theoretical capabilities were slowly but steadily improving. Even without the stimulus of ABM, however, MIRV's would probably have been devised, and probably at about the same time. All their essential technology was developed in unrelated programs and for unrelated reasons: Able-Star, Agena, Transtage, sequential payload delivery. During the early phases of development the progress of MIRV was determined largely by the decisions of technologists who were attempting to solve problems presented by nature or responding to their perception of the technological challenges of the Russian missile and space programs.

Once proved feasible, MIRV's were also proved necessary, at least to the satisfaction of those who made the decision to deploy them. That decision was compelled by three factors: (1) the participants in and sponsors of the development program urged application of the new weapon; (2) MIRV's promised to thwart any ABM system the U.S.S.R. might construct and at the same time served as an argument against the deployment of missile defenses by either nation; (3) MIRV offered a relatively inexpensive way to increase the number of Russian targets that could be attacked, and thus ended debate over strategic-force expansion.

Opposed to these arguments were the predictions of the Arms Control and Disarmament Agency that deployment of MIRV's would be perceived as seriously disturbing the strategic balance and in crises would make the most dangerous policy the most profitable one. The first of these predictions has already turned out to be right. The number of people holding such views was negligible, however, and their voices were not powerful. Plans to deploy proceeded while the arguments went on.

In the development of multiple warheads there is a lesson for those who would reduce the world's armaments. Some programs, such as the B-70 bomber and ABM, are expensive, are addressed to a clearly evident and single purpose and depend on what might be called a unitary decision-making process. In principle they can be stopped by direct confrontation. Other programs, however, and MIRV is an example, evolve from many independent and seemingly unrelated goals and decisions. They are too diffuse, too protean, too difficult to define and delimit to be stopped by confrontation. They can be slowed or stopped only by slowing or stopping the arms race as a whole.

MIRV

ment could have taken alternative pathways. Similarly, had some arguments for deployment been refuted, others would have served.

Communication by Optical Fiber

Rapid progress is being made toward a system in which a light signal, generated and repeated with small solid-state devices, will be transmitted through a hairlike fiber with little loss

by J. S. Cook

The quantity of information that can be carried by a transmission channel, for example a pair of telephone wires or a microwave radio beam, depends on the frequency of the signal; the higher the frequency, the greater the amount of information. Since the frequency of light is roughly 1,000 times the frequency of the shortest radio waves, communication engineers have always dreamed of using light as a means of transmitting information. The advent of the laser raised their hopes, because the coherent waves of laser light can be sharply directed and in principle might also be modulated like radio waves. But what kind of "channel" could be used to transmit the modulated light signal? Today the leading candidate for that function is the optical fiber: a hair-thin, glassy filament of extremely high transparency. Indeed, given the present state of development, it is not at all reckless to hope for an optical fiber that, over a distance of a mile, will lose no more than half of the light it is carrying.

Bundles of glass or plastic fibers have for some time been a means of directing illuminating light from one place to another. In many automobiles, for example, such fibers serve to pipe light from a single lamp to various parts of the instrument panel. These fiber bundles carry light very well over their length, which can be up to several meters, but they are not nearly transparent enough to be practical for communication over significant distances. For example, the fibers are typically less transparent than water is.

The recent work that has brought about the greatly enhanced transparency of optical fibers, making their use in communication possible, represents a genuine technological breakthrough. A feature that is attractive about optical fibers in communication is their small size. One hundred individual, signal-carrying fibers bundled together might occupy a space that has the diameter of the lead in an ordinary wooden pencil. Each fiber might carry 1,000 telephone conversations or several television programs at a time.

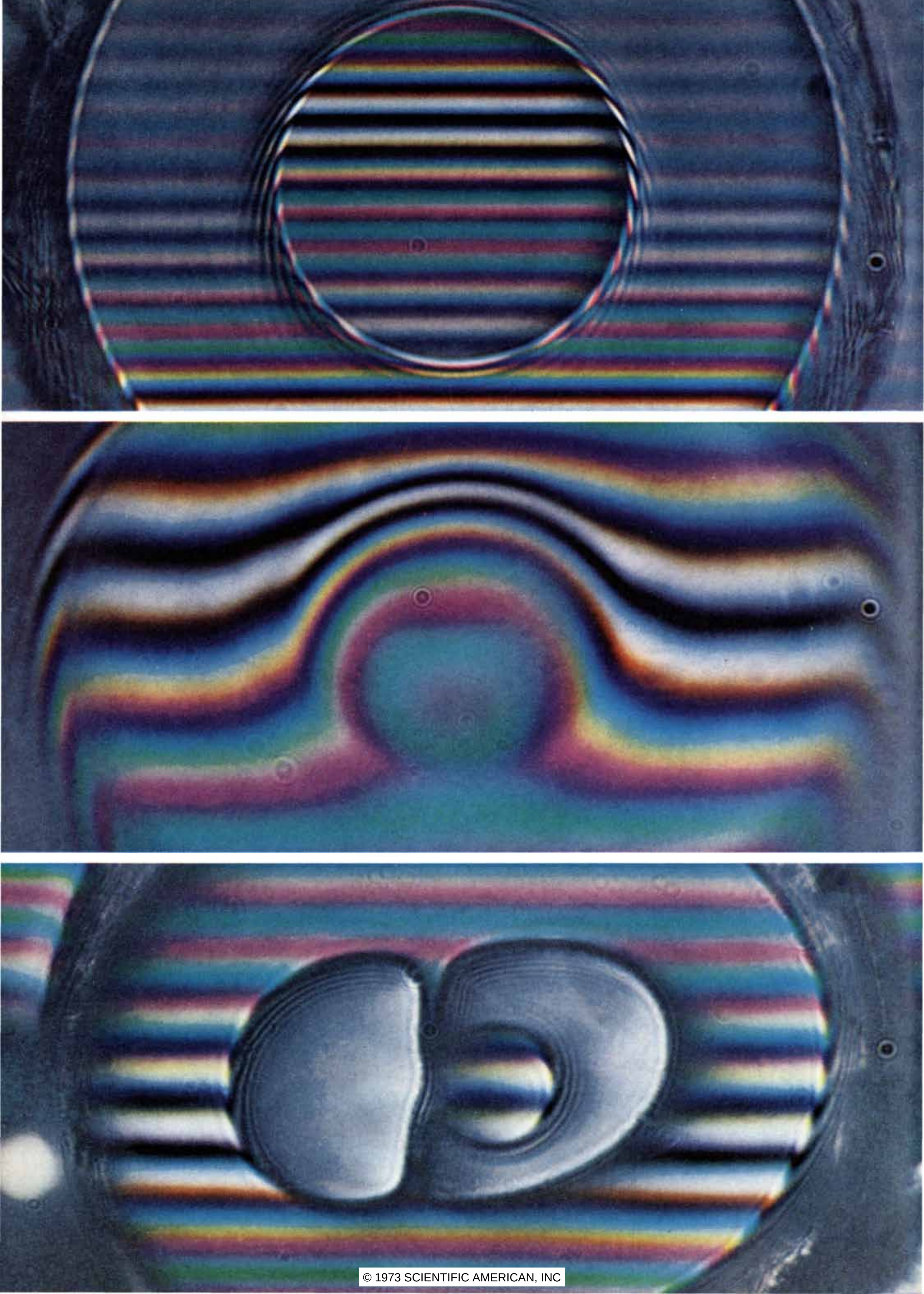
The principle underlying the guidance of light in optical fibers is quite simple. A fiber typically has a core, which is a central section that carries the light rays, and a cladding, which is an outer layer that traps the rays in the core. The materials of which the core and the cladding are made have slightly different indexes of refraction, that is, light travels through them at different speeds. The index of refraction is a bit less in the cladding than it is in the core. Because of this small difference light rays in the core that strike the interface between the two materials at a grazing angle are reflected back into the core. The very simplicity of the light-guiding mechanism leads at once to a possible drawback (known as differential delay) in sending signals on fibers and to a solution to the drawback. The problem and its solution should become clear if I first discuss certain general features and characteristics of communication systems.

Every communication system consists of a signal source, a signal-carrying medium (the transmission channel) and a receiver. Information is sent by varying the intensity of the source either continuously (analogue modulation) or impulsively (digital modulation). The medium carries the modulated signal to the receiver, which reconstructs the original information from the modulated pattern it receives. It is the task of the medium (in this case the fiber) to deliver the modulated signal to the receiver—not perfectly, but with enough fidelity so that a reasonably well-designed receiver can tell how the signal source was modulated.

The receiver's function is best seen with digital signals. Assume that the information consists simply of the pattern of light impulses generated in a continuous series of uniform time slots, some of which contain light flashes and some of which do not. To determine what the sent message was the receiver, which is designed to step along exactly in time with the transmitter (the source of the modulated light), must simply discover in which time slots the light source was flashed. Because the receiver "knows" how to interpret the on-off pattern, it can then reconstruct the original message.

A problem confronting the receiver is that the medium (in the current example the optical fiber) degrades the pulses of light in two ways as they travel along it: it attenuates, or dims, them and also makes them less impulsive, or broader in time. If the pulses become too dim, the ubiquitous noise in the receiver masks them. If they become too broad, they begin to spread into the adjacent time slots. Eventually they are washed out over so many slots, some of which

LIGHT-CARRYING FIBER made of fused silica and wound on a drum conducts light from a helium-neon laser, which is visible at left in the photograph on the opposite page. This fiber, which is some 100 meters long, was made so that light could be seen along the drum; in a fiber of the highest quality nearly all the light would emerge at the end of the fiber. Here the light emerging at top right of drum strikes a screen and is reflected as a red spot.



were flashed and some of which were not, that the receiver can no longer tell what the transmitter was doing.

The question of why the pulses broaden brings us back to the problem of differential delay—the trouble caused by the guiding mechanism. Some of the light rays entering the fiber from the light source travel parallel to the axis of the fiber, but other rays enter at an angle to the axis and thus reflect back and forth down the core. It is apparent that the coaxial rays will travel more rapidly than the reflecting ones because they do not travel as far.

If the difference in arrival time between the slowest rays and the fastest ones is comparable to the interval between impulses, the received signal pulses begin to overlap. The farther the signal travels, of course, the larger the differences in arrival time are and the more dispersed in time the pulses become. Moreover, the faster the light is modulated—the closer together the time slots are—the less the pulse-broadening needs to be to cause signal confusion.

The guiding mechanism offers a means of controlling this differential-delay problem as follows. Recall that the reflection at the interface of core and cladding depends on the fact that the two materials possess slightly different indexes of refraction. The situation is analogous to what a skin diver sees from below the surface of a quiet body of clear water: the surface is either a mirror or a window, depending on the angle from which it is viewed. Directly overhead it is a bright window to the upper world, whereas some distance away from the diver it is a mirror, reflecting only the light from underwater objects.

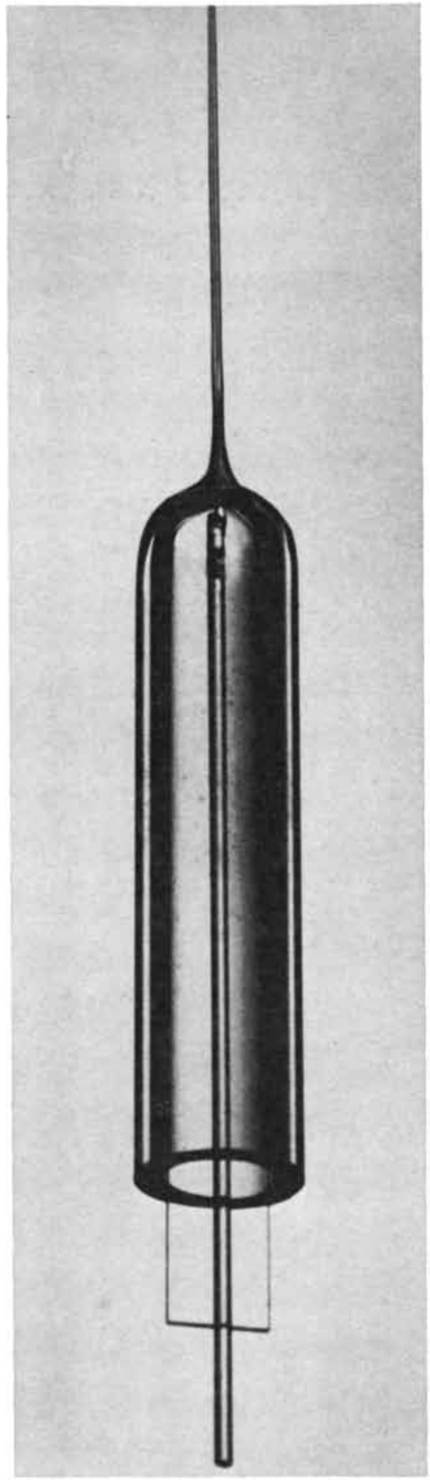
The reason is that light rays that strike the air-water interface from inside the higher-index water at a shallow angle are reflected back from the surface, whereas the rays that strike the surface more nearly perpendicularly to it go on through. The angle at which the surface becomes transmitting or transparent rather than reflecting depends on the difference in the refractive indexes of the two materials at the interface. If the difference is small, which is not the case with air and water, the critical angle is small too. Hence a solution to the differential-delay problem is to maintain the indexes of refraction of the core and cladding materials close together, so that the light rays that travel down the core all proceed almost axially. Rays that travel at slightly larger angles simply pass through the interface and are lost in the surrounding medium.

Indeed, it is possible to avoid the differential-delay problem altogether. As with other kinds of transmission line, such as a wave guide or a coaxial cable, the dimension of the guide—in this case the diameter of the fiber core—can be chosen in such a way as to eliminate the propagation of all but one electromagnetic mode. The strategy results in the capturing by the fiber core of only the axial rays. Such fibers have been made; the core in this case is typically a few microns in diameter. Launching light into such a thin core is a special problem to which I shall return.

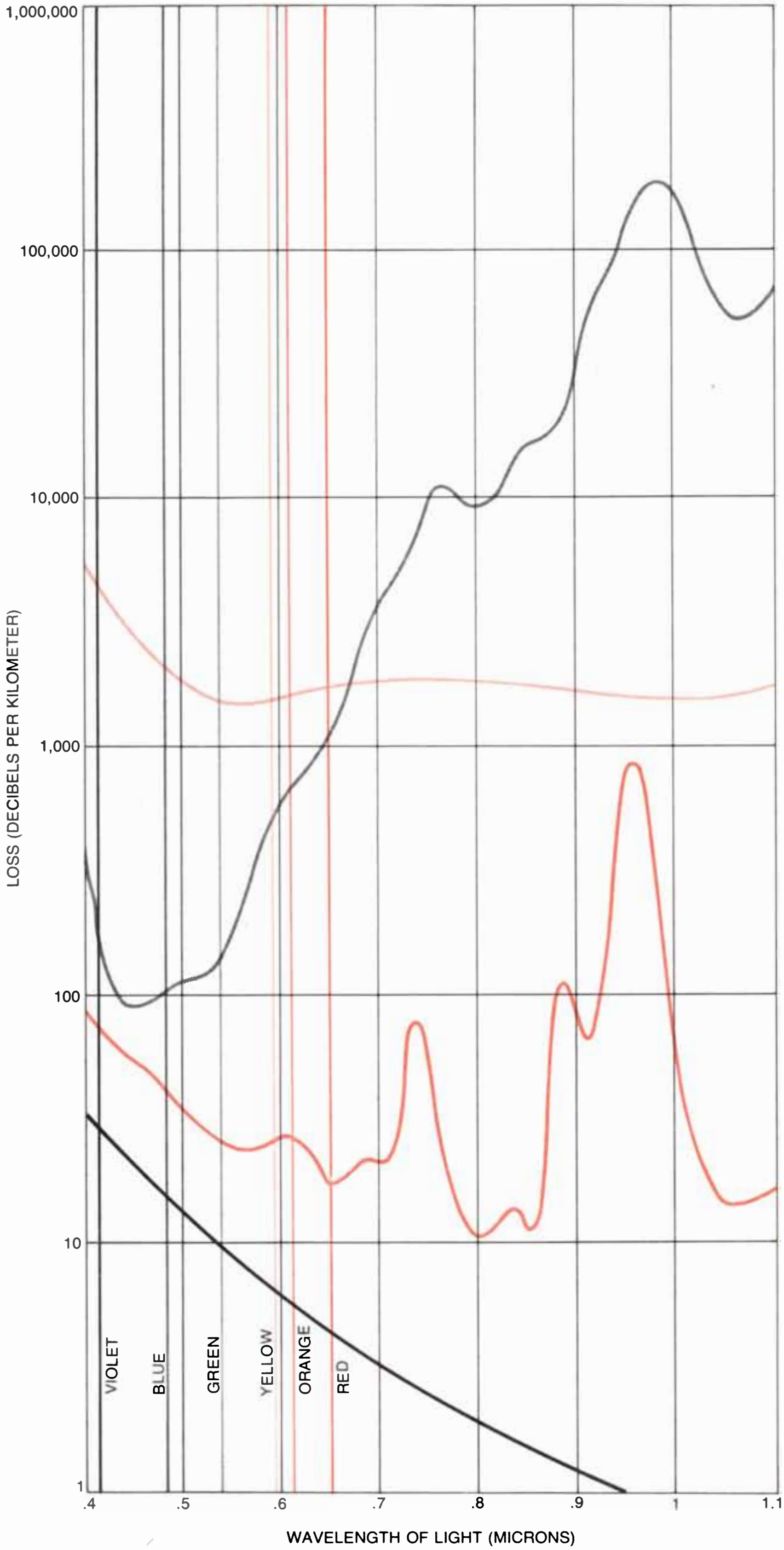
A third way of controlling differential delay is to employ a parabolic-index fiber. The Nippon Sheet Glass Co., Ltd., has made such fibers, which the firm calls Selfoc (for "self-focusing"), with remarkable accuracy. Stretching the concept of core and cladding, the parabolic-index fiber consists of a single material "loaded" with another material in such a way that the refractive index decreases with increasing rapidity away from the axis of the fiber. In this fiber light rays cross and recross the axis repeatedly, traveling not in straight lines but along snakelike sinusoidal paths. The fiber has the remarkable characteristic of speeding up the light rays that travel the longer distance, that is, the ones that make the greatest excursion from the axis, so that all the rays travel at nearly the same net axial velocity.

In addition to differential delay another reason for pulses to increase their width as they travel along the fiber is the phenomenon termed material dispersion. The index of refraction of the known low-loss fibers is slightly different for different wavelengths of light, that is, different colors. The result is that light of one color travels in the fiber at a

INTERFERENCE FRINGES in the photographs on the opposite page show the optical refractive-index characteristics of three kinds of fiber. Index difference is indicated by displacement of fringes. The fibers, which are seen in cross section at a magnification of about 500 diameters, are a fiber with an inner core and an outer cladding (*top*) having different indexes of refraction, so that the transmitted light tends to stay in the core; a parabolic-index fiber (*middle*), in which the refractive index decreases with increasing rapidity away from the fiber axis, and a single-material fiber (*bottom*), made as shown in the photograph on this page. The gray shapes are air. They surround the thin rod that carries the light. The continuity of the fringes from outside through the rod shows single-material nature of fiber.



SINGLE-MATERIAL FIBER is fabricated entirely from fused silica of high purity. Thick outer tube, which has an outside diameter of 10 millimeters, contains a rod and a thin plate. As the tube is heated at the bottom the rod is fused to the plate and the plate to the tube. The fiber drawn from the bottom of the assembly has an outside diameter of about 150 microns and the same cross section as the original assembly. Light travels through the rod, which serves as the core, and other parts of the structure serve as cladding, with effective refractive index determined by thickness of supporting plate.



TRANSMISSION OF LIGHT is charted. Pure water (*gray*) is most transparent at visible wavelengths and almost opaque in the infrared. A "light pipe" (*light color*), which is a bundle of fibers for transmitting light over short distances, is clearer than water at longer wavelengths. Fused silica (*dark color*) serves in fibers carrying light over long distances. Rayleigh scattering limit (*black*) is set by fundamental characteristics of pure fused silica.

velocity different from that of light of a slightly different color.

All optical sources radiate light at more than one frequency at a time, which is to say that the light is multicolored. The different colors from the source enter the fiber together in a single impulsive flash, but they arrive at the far end in a sequence of colors spread out in time. A detector there, being color-blind, simply "sees" a wider (more time-dispersed) pulse than the one sent by the source. A laser radiates nearly monochromatic light, so that typically the variation of the fiber's refractive index over the spectrum of the laser is too small to be a problem. For some sources of light, however, spectral width can be important.

Until recently the loss of light in optical fibers was so large that light pulses died away long before measurable pulse-broadening could take place. Light loss, in fact, is still the most important problem to be dealt with in fiber design. It comes about through absorption of the light energy and through scattering of the waves as they propagate through the fiber.

The glasses of which the best fibers are made absorb very little light. If as little as one part per million of certain metals (or, strangely, water) is included in the glass, however, the fiber will absorb a significant amount of light. The key to low absorption therefore lies in being vigilant about the cleanliness of the conditions under which fiber materials are compounded and processed.

Scattering of the light can be caused by nonuniformity of the fiber, particularly in the interface between core and cladding, by particles or bubbles in the core material, and even by the random arrangement of the molecules of the core material, which makes the material look slightly granular at light wavelengths. It is this third phenomenon, which is named Rayleigh scattering after the British physicist Lord Rayleigh, that sets a lower bound to the loss of light in glassy or fluid materials. Since the scattering caused by this granularity decreases rapidly as the wavelength becomes longer, infrared light is much to be preferred to visible light for low-loss transmission in fibers.

Two infrared lasers are particularly promising as sources of radiation for transmission through optical fibers: the neodymium-doped yttrium aluminum garnet laser, which is commonly called the "neodymium YAG" laser, and the gallium arsenide laser. The neodymium

YAG laser, which radiates at a wavelength of about 1.06 micrometers, is attractive because it is almost indestructible and because it is efficient in converting drive power into coherent light. The drive power is also light, which need not be coherent; it can be derived from several kinds of source, including the devices I shall describe below.

A complication for the neodymium YAG laser is that once it is turned on, its coherent radiation cannot be turned off quickly. If rapid changes of signal are wanted, the light beam sent out by the laser must be modulated by another device. Fortunately several types of light modulator are available.

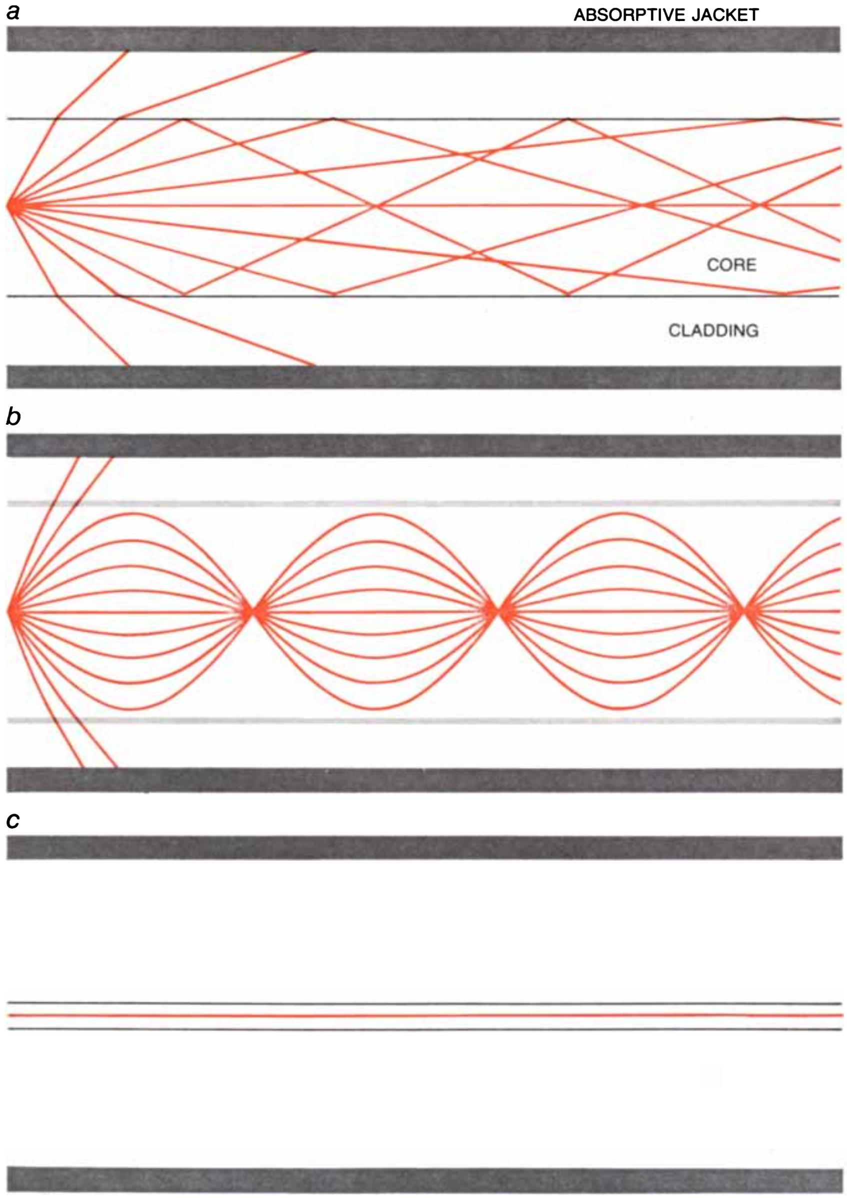
The gallium arsenide laser is an injection-diode device that has as its key element a tiny diode made in layers of semiconducting material [see "A New Class of Diode Lasers," by Morton B. Panish and Izuo Hayashi; SCIENTIFIC AMERICAN, July, 1971]. Such a diode has one region in which negatively charged electrons are the predominant current-carriers (*n*-type material) and another region in which positively charged holes predominate (*p*-type material). Between them is a *p-n* junction across which current is caused to flow. The resulting electrons are injected into the junction region, where they readily combine with holes, releasing their energy in the form of light. When the density of recombination events is sufficiently high, that is, when high current is flowing through the junction, the resulting optical electromagnetic field influences the process so that further radiation of light tends to reinforce the light wave already present. In short, laser action takes place.

The resonant cavity of an injection laser is formed by limiting the width of the high-current region in the junction to between 10 and 20 micrometers, forming what is called a stripe, and cleaving the junction at each end of the stripe to form light-reflecting surfaces. Light from this cavity can be coupled easily into a multimode fiber by simply butting one end of the laser stripe against the end of the fiber core, which has a much larger diameter (about 100 micrometers). The light from an injection laser can also be coupled to a single-mode fiber, even though the diameter of the core may be only three or four micrometers. The main problem is that the laser junction tends to be so thin (a half-micrometer or less) that the light diverges as it leaves the end of the stripe. This problem can be overcome by a tiny cylindrical lens that redirects the light along the fiber.

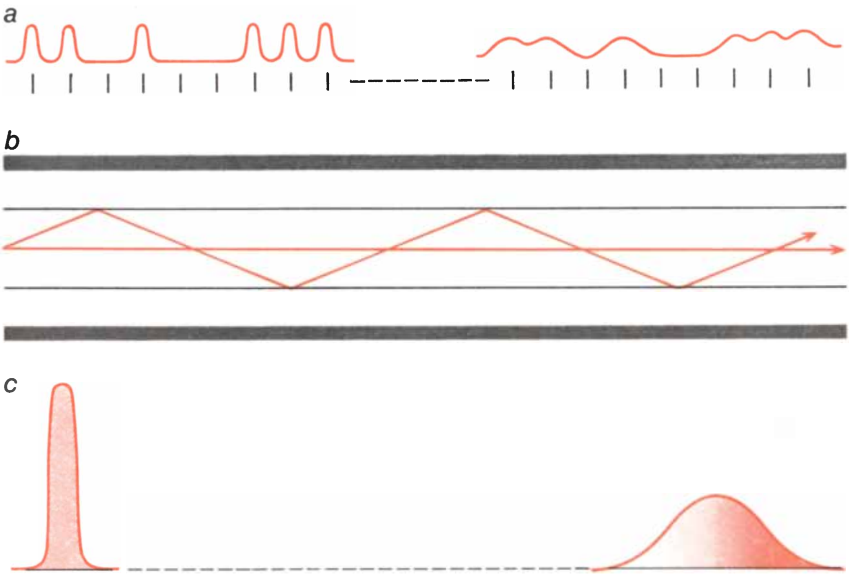
Depending on the amount of aluminum that is included with the gallium arsenide (in a demanding fabrication procedure), wavelengths anywhere between .8 and .9 micrometer can be generated. (The spectrum of any one diode is fixed and quite narrow.) These diodes can be modulated rapidly by simply modulating the electric current that powers them. They can be made to pulse up to 500 million times per second.

Lasers, then, are ideal for communication by optical fiber for two reasons: they can focus light into a narrow beam that can be directed into the fiber efficiently, and their spectral output is narrow. A practical hitch, however, is that injection laser diodes have not been made to radiate continuously for more than a few thousand hours at room temperature without failure. Although there is no apparent intrinsic cause of failure, such lasers cannot easily be employed in a commercial communication system until their durability has been improved.

A possible alternative to the laser as



FOCUSING OF RAYS in various kinds of optical fiber is depicted. In a fiber having a core and a cladding (*a*) rays that enter the fiber at a low angle with respect to the axis of the fiber are trapped in the core by reflection from the cladding, which has a lower index of refraction than the core. High-angle rays pass into the cladding and may, as in this case, be absorbed in an outer jacket. A parabolic-index fiber (*b*) focuses low-angle rays again and again along its length by means of a carefully devised refractive-index gradient. Here too high-angle rays are lost. A single-mode fiber (*c*) eliminates the problem of delayed rays by allowing propagation of only one electromagnetic mode, in effect only the axial ray.



DISTORTION OF SIGNAL passing through an optical fiber takes place in two ways. At top left (*a*) the original signal appears, together with short vertical lines indicating the time interval. At the far end of the fiber the signal is less clear. One reason (*b*) is differential delay. The axial ray and the angular one are the same length, illustrating the difference in distance traveled in a given time by the rays in a multimode fiber. As a result the light energies traveling the two routes arrive at a given point down the fiber at different times. Material delay (*c*) results because index of refraction and hence propagation velocity of known low-loss fibers change with the frequency of light. Original distinct signal (*left*) reaches end of fiber with the source spectrum spread out in time, as indicated by colored shading.

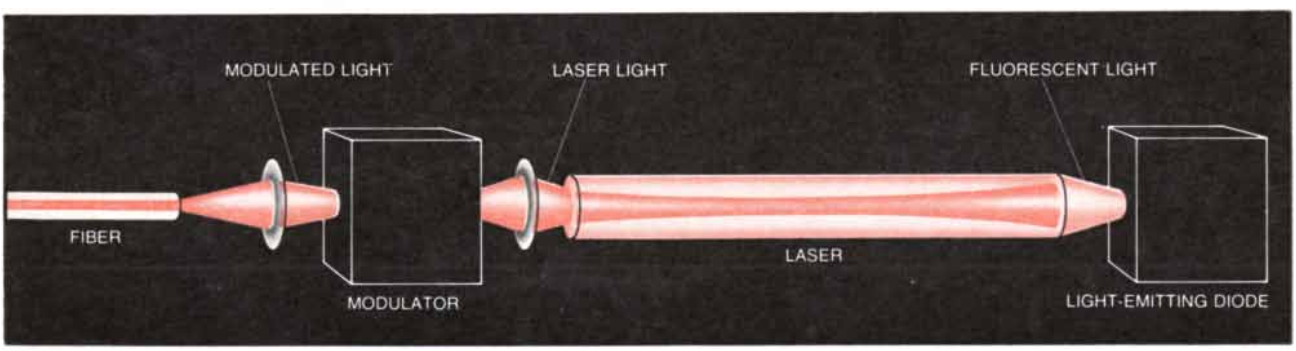
a signal generator is the gallium arsenide light-emitting diode. Although it is made of the same materials as the injection laser, it radiates a wider spectrum of light frequencies over a wide angle. Typically a gallium arsenide or aluminum gallium arsenide light-emitting diode radiates light throughout a range of wavelengths as much as .04 micrometer wide. Light from the diode can be filtered so that only part of the total spectrum will pass into the fiber, but this is done at the cost of a commensurate decrease in the total available light-source power.

A more severe penalty results from the fact that the light from the intensely bright fluorescent-spot source in the light-emitting diode radiates in all directions. Only a small part of it can be captured by a fiber, namely the rays that enter the fiber at angles (with respect to the axis of the fiber) less than the critical angle that is determined by the difference in refractive index between the core and the cladding. Notwithstanding these difficulties, the light-emitting diode is attractive in its simplicity and durability. The question is whether or not it is usable.

It is here that the transmission engineer must begin making choices. He would like to be able to send his signals as far as possible along a fiber before they must be either received or regenerated. The received pulses must be neither too weak nor too broad. If he is to use a light-emitting diode as a source, he would like the difference in the index of refraction between the core and the cladding to be as large as possible in order to trap as much light as possible in the core. On the other hand, if the index difference is too large, the pulse will be broadened out by the differential delay between the axial rays and the high-angle rays. Clearly the engineer would like to choose an index difference that causes the pulses to reach both limiting breadth and dimness at the point of regeneration.

But how dim is dim? Where does the noise in the receiver mask the incoming pulses to the point at which the receiver is incapable of ascertaining which time slots contained flashes? Of the number of photons, or units of light energy, transmitted in each pulse, enough must be still traveling with the pulse when it reaches the receiver so that the receiver can detect it (with a finite rate of error of, say, one mistake in a billion). The requirement can be translated into average power and might be as little as one billionth of a watt for a pulse rate of 1.5 million per second, the rate at which certain digital telephone transmission systems work today.

Taking into account that requirement of correctness, together with the signal power available at the transmitting end from a light-emitting diode (less the power needed by the receiver), and assuming that a loss of five decibels per kilometer is reasonable for future mass-



"NEODYMIUM YAG" LASER is employed as a source of nearly monochromatic light for transmission through optical fibers. It is basically a crystal of yttrium aluminum garnet (YAG) doped with neodymium and ground to the shape of a rod, with mirrors deposited on the ends. In this configuration pumping light supplied by a light-emitting diode enters the laser through a frequency-selective mirror at one end, and the coherent light generated in the laser rod passes out the other end. The light is then modulated by any one of several light-modulating devices that are suitable for the purpose to create information-carrying signal for the optical fiber.

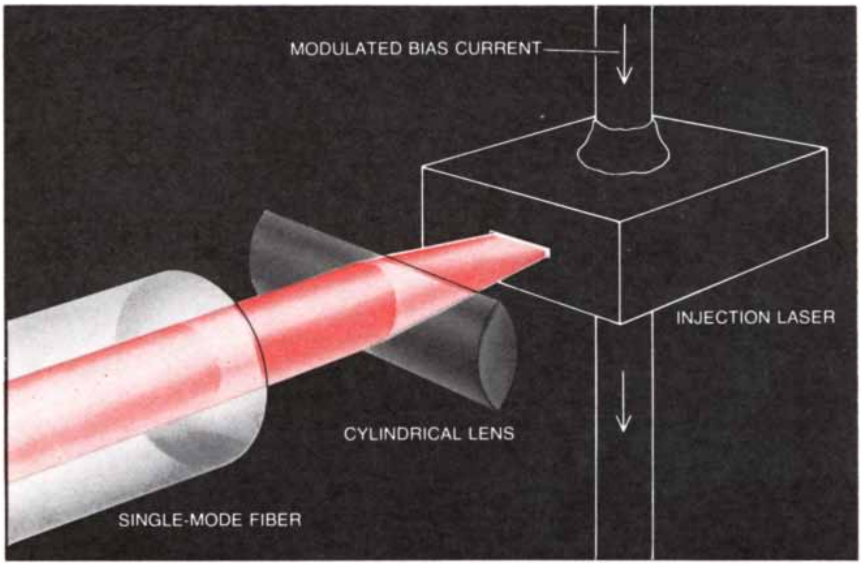
produced fibers, the distance between repeaters works out to about eight kilometers (five miles). The optimum core-cladding index difference, as a balance between the power that can be directed into the fiber from a gallium arsenide light-emitting diode and the allowable differential delay, is about 2 percent. An index difference of .5 percent, more typical of current low-loss fibers, would reduce the repeater spacing by about a kilometer.

Material dispersion does not appear to present a problem. One of the most transparent of all materials is fused silica (silicon dioxide or, in the crystalline form, quartz), which is one of the commonest minerals. It is also one of the least dispersive of optical-fiber glasses. A signal with a spectral width of .04 micrometer, pulsing at 1.5 million times per second, will not exhibit any overlap of the pulses due to material dispersion alone for nearly 100 miles.

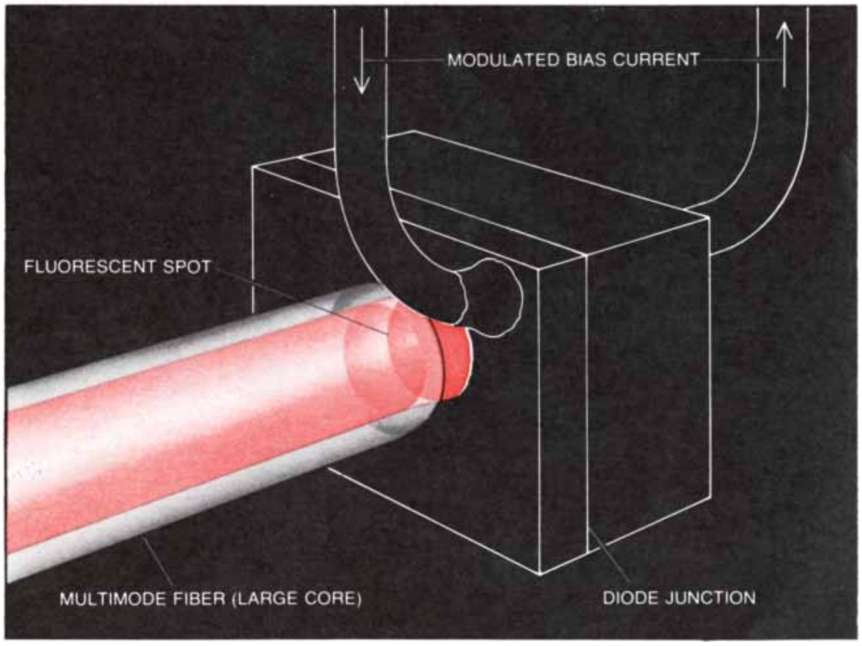
One emerges, then, with the prospect that a long-lived fluorescent aluminum gallium arsenide diode will be able to send 24 telephone conversations (consistent with the current digital trunks) by pulse-code modulation for five miles along a low-loss, fused-silica fiber to a simple receiver that will reproduce the voices accurately. Present digital trunk-carrier telephone systems that send their pulses along standard telephone cable wires must amplify and reshape the pulses about once each mile.

These estimates are not intended to suggest the structure of an optical-fiber communication system. It is true that the essential components of such a system have been demonstrated in the laboratory: gallium arsenide light-emitting diodes of the proper intensity have operated at Bell Laboratories for several thousand hours with little degradation; the Corning Glass Works has announced that a loss of four decibels has been measured on a one-kilometer fiber of fused silica, and receiver sensitivity of the kind I have discussed can be demonstrated with any of several commercially available light detectors. At this point in the development of this new and different technology, however, the system that can be inferred from this simple calculation can serve only as a kind of engineering symbol for future transmission-system possibilities.

The technology programs directed toward communication by optical fibers are in an exploratory period that can be described as being almost feverish. Discoveries, new understanding and technological advances are coming too rapidly for anyone to be sure about the course that the development of commercial-scale systems will take. Indeed, many important questions remain to be answered. What materials and designs are best for fibers? How should the fibers be manufactured? How should they be put into cable? How should the cables be connected and terminated? A great deal of work will be required to answer these questions, but the communications promise held out by optical fibers makes the effort well worth while.



INJECTION LASER of aluminum gallium arsenide can provide signals for both multimode and single-mode fibers. Its key element is a tiny diode (*right*) consisting of layers of semiconducting material separated by a junction. Current is constrained to pass through the junction only along a narrow stripe between 10 and 20 micrometers wide. The laser cavity, which is only half a millimeter long, is formed by cleaving the diode perpendicularly to the stripe. Light can be put into a multimode fiber simply by butting the end of the fiber against one end of the laser stripe. For a tiny single-mode fiber a lens is required to direct the light.



LIGHT-EMITTING DIODE of aluminum gallium arsenide is a simple nonlaser light source for an optical fiber. Although the light issuing from the diode in the form of a fluorescent spot, which has a diameter of approximately 50 micrometers, radiates in all directions, a significant amount of it can be captured by a large-core fiber bonded close against the spot.

Proton Interactions at High Energies

Experiments performed in the new CERN colliding-beam accelerator have yielded an unexpected finding: In a certain high-energy range the chance that two passing protons will interact increases with energy

by Ugo Amaldi

In the effort to penetrate the structure of matter, physicists have built particle-accelerating machines of ever increasing energy. By bombarding matter with the energetic particles produced by such machines new species of particles have been created. The list of "elementary." particles discovered in this way is already so long that no one believes the particles on it are really elementary. The large number of the particles and the striking patterns evident in their properties are clear indications that they have some kind of internal structure.

Most of the current research in high-energy physics is directed toward investigating how the particles interact with each other. The key question here is: What is the nature of the forces that act between particles and that give rise to the three main types of particle behavior observed in nature: the decay of one particle into other particles, the scattering of one particle from another particle and the transformation of the energy set free in the collision of two particles into new particles?

The interactions observed in matter can be grouped into four distinct categories. In order of increasing strength they are the gravitational interaction, the "weak" interaction, the electromagnetic interaction and the "strong" (or nuclear) interaction. Our knowledge of the force of gravity comes mainly from astrophysical studies; at present the gravitational interaction does not seem to play any role in particle physics. The weak interaction, which is observed in the radioactive decay of atomic nuclei, is more than a billion times less intense than the electromagnetic interaction, which binds electrons to the nuclei in atoms and molecules and determines the chemical properties of all substances. The strong interaction, which binds the protons and the neutrons in the nucleus, is extremely intense and acts only within distances on the order of a fermi (a trillionth of a millimeter).

In this article I shall describe the background, the results and the meaning of two recent experiments that have revealed a new and unexpected property of the strong interaction at extremely high energies. The experiments, which involved collisions between oppositely directed beams of high-energy protons, were performed with a new kind of accelerator: the intersecting storage rings at the European Organization for Nuclear Research (CERN) near Geneva. The results obtained so far with this machine have shown that above a certain energy the probability that two passing protons will interact increases with energy, a finding that may require a major revision in the accepted description of high-energy nuclear interactions.

The many hundreds of "elementary" particles that are subject to both weak and strong interactions are classed as hadrons; the few particles that feel only weak or electromagnetic forces are classed as leptons. Electrons, muons and neutrinos are the only known leptons, whereas protons, neutrons and pions are just three examples of the very large family of hadrons. Many physical quantities are used to characterize a particle, the simplest being its mass. According to the famous Einstein equation $E = mc^2$, which relates mass (m) to energy (E) in terms of the fundamental constant c (the speed of light in a vacuum), the mass of a particle can be expressed in terms of its equivalent energy. Thus it can be said that the proton has a mass of .94 GeV, whereas the pion, the lightest of the hadrons, has a mass of only .14 GeV. (GeV stands for gigavolts, or billions of electron volts.)

The strong forces acting between two hadrons that pass close to each other (at distances on the order of a fermi or so) lead to a localized burst of energy, which often materializes so that a fraction of the energy is transformed into the mass of a number of newly created hadrons. In such "inelastic" processes the rest of the energy appears as the kinetic energy of the outgoing particles. In a comparatively small percentage of the collisions no fraction of the available energy is transformed into mass; the incoming hadrons are simply deflected and are said to undergo "elastic" scattering. The recent experiments that have been conducted at CERN address the question of how the overall probability of elastic and inelastic encounters between two hadrons depends on the energy of the colliding particles.

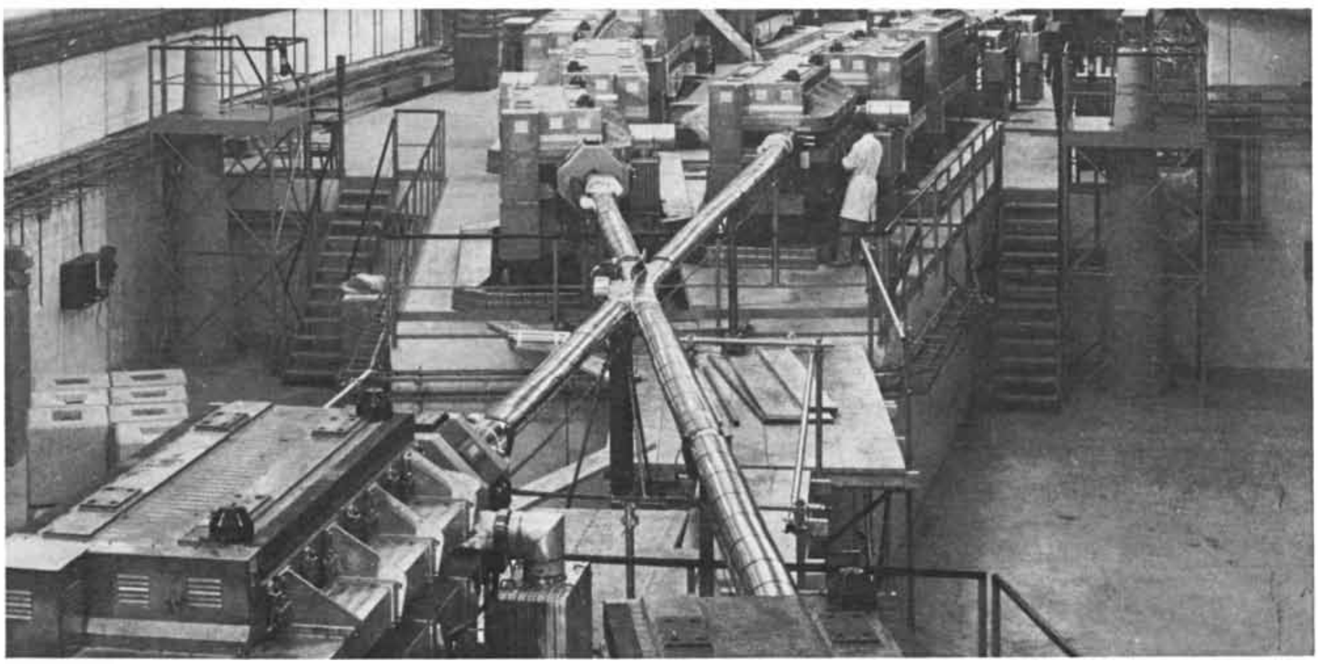
The quantity best suited to express the energy at which two hadrons collide is the energy available in the center-of-mass system. (In this reference system the center of mass of the collision is assumed to be at rest, so that no energy is going into its movement and hence the total energy of the colliding particles can be transformed into the mass and the kinetic energy of the outgoing hadrons.) At any given center-of-mass energy the probability of observing an elastic or an inelastic interaction when a moving hadron passes through a thin target is expressed in terms of the elastic and inelastic "cross sections." These quantities are measured in square centimeters and can be defined as the equivalent target areas offered by one of the colliding hadrons to the other as far as elastic or inelastic processes are concerned [*see illustration on page 40*]. The elastic and inelastic cross sections depend on both the type of the colliding hadrons and the energy available in the center-of-mass system. The total probability of interaction (which takes into account both elastic and inelastic processes) is obviously measured by the sum of the elastic and inelastic cross sections. This sum is called the total cross section of the interaction.

In conventional accelerators total cross sections are determined by letting a beam of high-energy hadrons impinge on a target and measuring the fraction of the incoming particles that pass through the target without being either deflected or absorbed. By "conventional" accelerators I mean devices such as synchrotrons in which protons, accelerated to some tens or hundreds of GeV, interact with a solid target to produce, through inelastic processes, the hadrons that are then used in the cross-section measurement. Of course, one can get a beam only of those hadrons that are either stable (particles such as protons and antiprotons) or metastable (particles with a lifetime long enough for them to be detectable many meters away from the production point, before they decay into other particles). The only metastable hadrons that live long enough for practical cross-section measurements are the positively and negatively charged pions and the positive and negative kaons. All the other hadrons, whose cross section would be very interesting to measure, decay in such a short time that they cannot form beams.

The choice of hadrons that can be employed as targets in such experiments is even more restricted: only protons (the nuclei of hydrogen atoms) and neutrons contained in nuclei will serve the purpose. When the results of all the cross-section experiments performed before 1972 with protons as targets are plotted as a function of the energy available in the center-of-mass system, one finds that, whereas below about 4 GeV the cross sections build up a complex landscape of peaks and valleys, above that value they become smooth functions of the energy and tend toward constant values [*see illustration on page 41*]. Indeed, until a few months ago extrapolation of the data suggested that all hadron-hadron total cross sections tend toward constant values with increasing energy. At first this seems to be a relatively simple physical situation: As long as the energy is high enough, its particular value is irrelevant in determining the probability of interaction between two hadrons; hence in this view increasing the energy changes nothing. It has been known for many years, however, that this simple interpretation cannot be correct. Measurements of the angular distribution of elastically scattered hadrons have shown that in the same energy range where the total cross section is constant (from 4 to 10 GeV) the distance at which two interacting protons "feel" each other increases with energy. To appreciate the significance of this statement we must take a closer look at the phenomenon of elastic scattering at high energies.

The language used so far in this article to describe hadronic processes is consistent with the notion that hadrons resemble tiny billiard balls, which (in elastic collisions at least) collide and ricochet away from each other. Quantum mechanics tells us, however, that this picture is wrong; each moving hadron behaves more like a wave than like a little ball. The wavelength (λ) of this wave is related to the energy (E) of the particle by the well-known de Broglie relation $\lambda = h/E$, which stipulates that for very energetic particles (that is, for particles moving with velocities close to the velocity of light) the wavelength is inversely proportional to the energy, the constant of proportionality being Planck's constant (h). The numerical value of this constant is 1.2 when the wavelength is measured in fermis and the energy of the particle in GeV.

The phenomenon of two hadron waves moving toward each other and



INTERSECTION REGION of the new colliding-beam proton accelerator at the European Organization for Nuclear Research (CERN) is one of eight such experimental areas where interactions between the two oppositely directed beams of high-energy protons can be studied. The photograph was made before the installation of the complex detection equipment that was used by a team of investigators from the University of Pisa and the State University of New York at Stony Brook to measure the proton-proton "total cross section," or overall interaction probability, in a high-energy range never before explored. Thousands of sputter-ion and titanium pumps maintain an ultrahigh vacuum of a few times 10^{-11} torr inside the two intersecting stainless-steel chambers in which the beams travel. The vacuum chambers cross at an angle of 15 degrees. The large blocky structures in the foreground and background are magnets that direct the beams along their roughly circular paths. A plan view of the intersecting storage rings appears on the next page.

interacting is difficult to visualize. It is completely equivalent, however, to the more familiar phenomenon of a single wave that impinges on an absorbing object fixed in space. Consider the hypothetical hadron-absorbing disk in the illustration on page 42. The disk is an absorber because the interacting hadrons give rise to many inelastic processes that have the effect, in the wave description, of subtracting energy from the incoming wave and transferring it to the outgoing hadrons. The phenomenon can be characterized by the "absorptiveness profile," a curve that represents the distribution of the absorption as a function of the distance from the center of the disk. In general this profile varies with the energy available in the center-of-mass system, because the probability of the various inelastic processes is a function of the energy.

The far-reaching consequences of the wave nature of the interacting hadrons can best be appreciated by comparing the two simple situations depicted in this illustration. In the first situation a large number of pointlike particles move along parallel paths and are absorbed by the disk. In the second situation a plane wave, that is, a wave with a flat wave front, is absorbed by the disk.

In the first case all the particles hitting the disk are absorbed as a result of the inelastic processes by which other hadrons are produced. The inelastic cross section is clearly equal to the area of the disk, and as a result a "hole" appears in the incoming beam and extends an infinite distance. Since the pointlike particles behave independently, and since each of them either is absorbed by the disk or is unaffected by it, there is no elastic scattering and the elastic cross section is equal to zero.

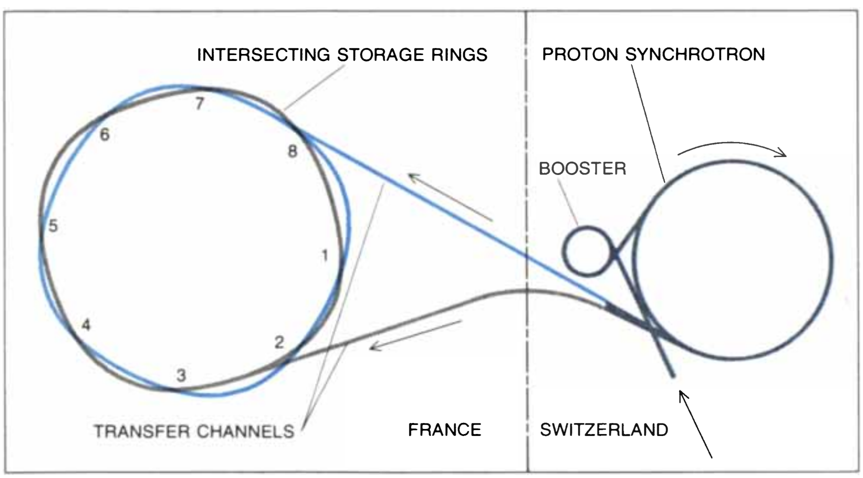
In the second case the incoming plane wave describes a beam of particles whose energy can be deduced from the wavelength by means of the de Broglie relation. The portion of the wave front that strikes the disk is absorbed, and its energy is converted into other forms of radiation emerging from the disk itself. These are the hadrons produced in the inelastic processes. In this case too the inelastic cross section is equal to the area of the disk; owing to the absorption, however, a "shadow" is produced behind the disk. In contrast to the hole produced in the beam of pointlike particles, the shadow extends only to finite distances, because the various portions of the wave front are not independent of one another. For this reason it is impossible to "drill" a hole in the incoming wave, so that the hole remains as the wave runs to greater distances; the sharp edges of the hole in the wave front spread out sideways to fill the hole itself. Thus the very existence of absorption implies the appearance of a wave that has the same wavelength as the incoming wave and propagates from the disk outward. This wave, which is called the diffracted wave, describes particles that are scattered elastically (without losing energy) because it has the same wavelength as the incoming wave. Actually the equality of the wavelengths of the incoming wave and the diffracted wave implies that the energy of the diffracted particles equals the initial energy (as a consequence of the de Broglie relation, which relates wavelength to energy).

How big is the elastic cross section due to this diffraction phenomenon? For a completely absorptive disk it is as large as the inelastic cross section; hence the total cross section is twice the area of the absorbing disk. To accept this perhaps surprising result, it is necessary to understand how the shadow is formed. Two separate waves of the same wavelength propagate in space: the incident plane wave and the diffracted wave (which has a spherical wave front because it has its source in the disk). The two waves interfere; in particular, just behind the disk there must be totally destructive interference, so that the two amplitudes cancel and the shadow is produced. This requires the scattered wave to have the opposite phase but the same intensity as the incident wave. Thus the energy that is diffracted is equal to the energy that impinges on the disk. This fact in turn implies that the elastic cross section equals the inelastic one.

In short, for a completely absorbing disk the elastic and inelastic cross sections are equal and the total cross section is twice the area of the disk; it follows that the ratio of the elastic cross section to half the total cross section is equal to 1. For a disk of uniform but incomplete absorption the elastic cross section becomes smaller than the inelastic one, and the ratio of the elastic cross section to half the total cross section becomes smaller than 1. In this case it can be shown that the ratio in question can be considered a measure of the average absorptiveness that the two hadrons feel as a consequence of the many inelastic processes their interaction produces.

More detailed information on the absorptiveness profile is contained in the form of the diffraction pattern, that is, in the angular distribution of the scattered particles. That distribution presents a large maximum in the forward direction, at a scattering angle equal to zero, and a series of minimums and maximums at larger angles. As in all diffraction phenomena, the angular width of the central peak of the diffraction pattern is proportional to the ratio of the wavelength to the radius of the diffracting object. Hence from a measurement of the angular width of this peak the radius of the object can be obtained. If the absorption is not complete, as in the case of two colliding hadrons, the quantity obtained can be interpreted as being an effective value of the hadron-hadron interaction radius.

An absorptiveness profile can be com-



PLAN VIEW of the CERN site shows how protons accelerated in the laboratory's "conventional" proton synchrotron (*right*) are injected into the intersecting storage rings (*left*) through two transfer channels that cross the border between Switzerland and France. After injection and "stacking," the two opposed proton beams can circulate in the vacuum chambers for more than 1,000 hours; usually, however, the beams are "dumped" after each run, which lasts some 50 hours. The rings, which intersect at eight points, are 300 meters across.

AERIAL PHOTOGRAPH of the CERN laboratory near Geneva must be examined rather closely for evidence of the two major underground proton accelerators. The large circular mound in the left foreground houses the intersecting storage rings. The somewhat smaller circular mound in the right background, partly obscured by buildings, contains the original CERN proton synchrotron. The border between Switzerland and France runs horizontally across the middle of the photograph. Construction of the storage rings, supported by 12 European countries, began in 1966. The experiments reported in this article were carried out in past two years.

puted from a measured angular distribution of elastically scattered hadrons. Since the profile determines both the absorption and the scattering, it is obvious that from the knowledge of the diffraction probability at all angles the total cross section can be derived. The situation is actually even simpler than this, however, because a theorem exists that enables one to compute the value of the total cross section solely from the measurement of the probability of elastic scattering at zero angle, that is, at the maximum of the diffraction peak.

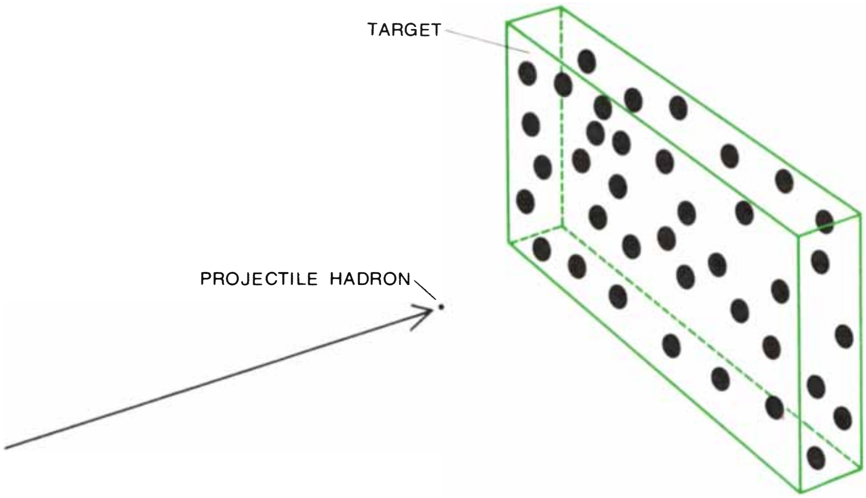
The "optical theorem," which was put in precise mathematical form by Niels Bohr, Rudolf Peierls and George Placzek in 1936, is valid for all wave phenomena and for any absorptiveness profile. Without going into the actual proof of the theorem, it is instructive to indicate why one should expect such a theorem to exist. The wave nature of the interacting particles implies that two quantities determine the entire diffraction phenomenon: the interaction radius and the absorptiveness. From an experimental point of view, on the other hand, there are three quantities that have to be considered independently; they are the total cross section, the probability of elastic scattering at zero angle and the width of the diffraction peak. Clearly some definite relation must hold among these three quantities. As it happens, the proof of the optical theorem shows how the first two of them are related. It is this theorem that enables us to compute total cross sections simply from the measurement of forward elastic scattering.

The information obtained through the measurement of proton-proton elastic scattering and the total cross section can be displayed graphically by plotting on the horizontal axis the energy in the center-of-mass reference system and on the vertical axis the values of the interaction radius (deduced from the measured width of the central diffraction peak) and the average absorptiveness (computed as the ratio of the elastic cross section to half the total cross section). In the illustrations on page 43, for example, all the data available at present are plotted; before the experiments performed at the CERN intersecting storage rings only the points below about 10 GeV were known. The curves clearly show that in the energy range where the total cross section remains essentially constant (that is, for energies lower than about 10 GeV) the proton-proton interaction radius increases and the average absorptiveness decreases.

These results and similar ones obtained in the study of other hadron-hadron collisions had led most high-energy physicists to accept the following hypothesis: With increasing energy, hadrons behave toward each other as absorbing objects of increasing radius but decreasing absorptiveness in such a manner that the overall probability of interaction (measured by the total cross section) remains constant. It is this picture, sometimes called the Regge pole exchange model, that has to be revised in view of the recent experiments performed in the energy range opened up by the CERN intersecting storage rings.

The protons injected at low energy in proton synchrotrons are made to move in circular orbits of fixed radius by means of strong magnetic fields and at the same time are accelerated by radiofrequency electric fields. At the end of the acceleration cycle, which lasts a few seconds, a burst of about a trillion protons is extracted from the machine and directed toward a hydrogen target, where proton-proton interactions can be studied.

Two quantities are important in every high-energy physics experiment: the energy of the collision in the center-of-mass system and the number of useful events produced in the target per second. The energy to which the proton beam has to be accelerated in order to produce, in a collision with a stationary proton, a given center-of-mass energy is quite high. For instance, to obtain a center-of-mass energy of 10 GeV a 60-GeV proton has to be sent against a proton at rest. Clearly the collision of an energetic proton with a stationary one is not a very efficient way of producing large center-of-mass energies, and the inefficiency increases with energy. For instance, passing from a projectile energy of 25 GeV (the energy of the protons accelerated at the Brookhaven and CERN proton synchrotrons) to the sixteenfold larger energy of



EXTREMELY SIMPLIFIED PICTURE of the interaction of two passing hadrons is introduced here solely for the purpose of helping to visualize the concept of the interaction's cross section, that is, the probability that the two hadrons will interact. (Hadrons are particles, such as the proton, that interact with the strong nuclear force.) In this view a hadron impinging on a target is thought of as pointlike, whereas the protons contained in, say, a hydrogen target are each considered as offering to this projectile an equal area such that an interaction takes place whenever the pointlike moving hadron passes through one of these areas. The area associated with a single target proton, measured in square centimeters, is thus merely an "equivalent" target area that depends on the relative energies, or speeds, of the interacting hadrons and is only indirectly related to their actual dimensions. This area, referred to as the total cross section, measures the overall interaction probability of the two hadrons. It in turn is the sum of two partial cross sections, termed the elastic and inelastic cross sections, that measure the probability of elastic and inelastic encounters respectively. (For most other purposes the wave description of a moving hadron is more appropriate.)

400 GeV (the energy of the protons accelerated at the National Accelerator Laboratory in the U.S.) the center-of-mass energy increases by a factor of only four, passing from 7 to 27 GeV.

On the other hand, the interaction rate that can be obtained by having a beam of a trillion protons pass through a target of liquid hydrogen is very large. This property is best described by introducing the concept of luminosity, which is defined as the number of interactions one would get in a second if the cross section were equal to a square centimeter. This apparently complicated definition boils down to a very simple prescription: In order to determine the number of interactions per second one must multiply the luminosity (measured in number of interactions per square centimeter per second) by the total cross section (measured in square centimeters). In a hypothetical situation in which a trillion protons per second pass through a target consisting of a one-meter-thick mass of liquid hydrogen, the luminosity is 5×10^{36} interactions per square centimeter per second. Assuming a total cross section of 4×10^{-26} square centimeter, one obtains 2×10^{11} interactions per second; in other words, 20 percent of the trillion (10^{12}) protons impinging on the target interact.

In accelerators built on the storage-ring principle, in contrast, two intense beams of stable particles are kept moving in opposite directions on almost circular trajectories that cross at a few points. The particles of the two beams collide head on in these intersection regions, so that very large center-of-mass energies result. Thus, for example, when two protons with an energy of 25 GeV each collide head on, the resulting center-of-mass energy is 50 GeV.

The main drawback of colliding-beam accelerators is their low luminosity. Large currents have to be accelerated and stored for long periods of time in order to give a significant interaction rate. In colliding-beam facilities built and operated in the U.S., Italy, France and the U.S.S.R. beams of electrons and positrons have been stored and made to interact [see "Electron-Positron Collisions," by Alan M. Litke and Richard Wilson; SCIENTIFIC AMERICAN, October]. The CERN intersecting storage rings represent the first and the only colliding-beam machine for protons. They were built during the period from 1965 to 1971 by a team under the direction of Kjell Johnsen at the original CERN site near Geneva in Switzerland. Because of the size of this new facility the CERN laboratory was extended just across the border into France [*see illustration on page 38*].

To fill the intersecting storage rings several hundred bursts of a trillion protons each are accelerated in the CERN proton synchroton and "stacked" into the two rings through two transfer channels. This results in currents on the order of 10 amperes circulating in the two evacuated chambers. The vacuum is so good that the losses caused by the interaction of the protons with the residual gas are negligible and data can be collected continuously for periods of up to 50 hours.

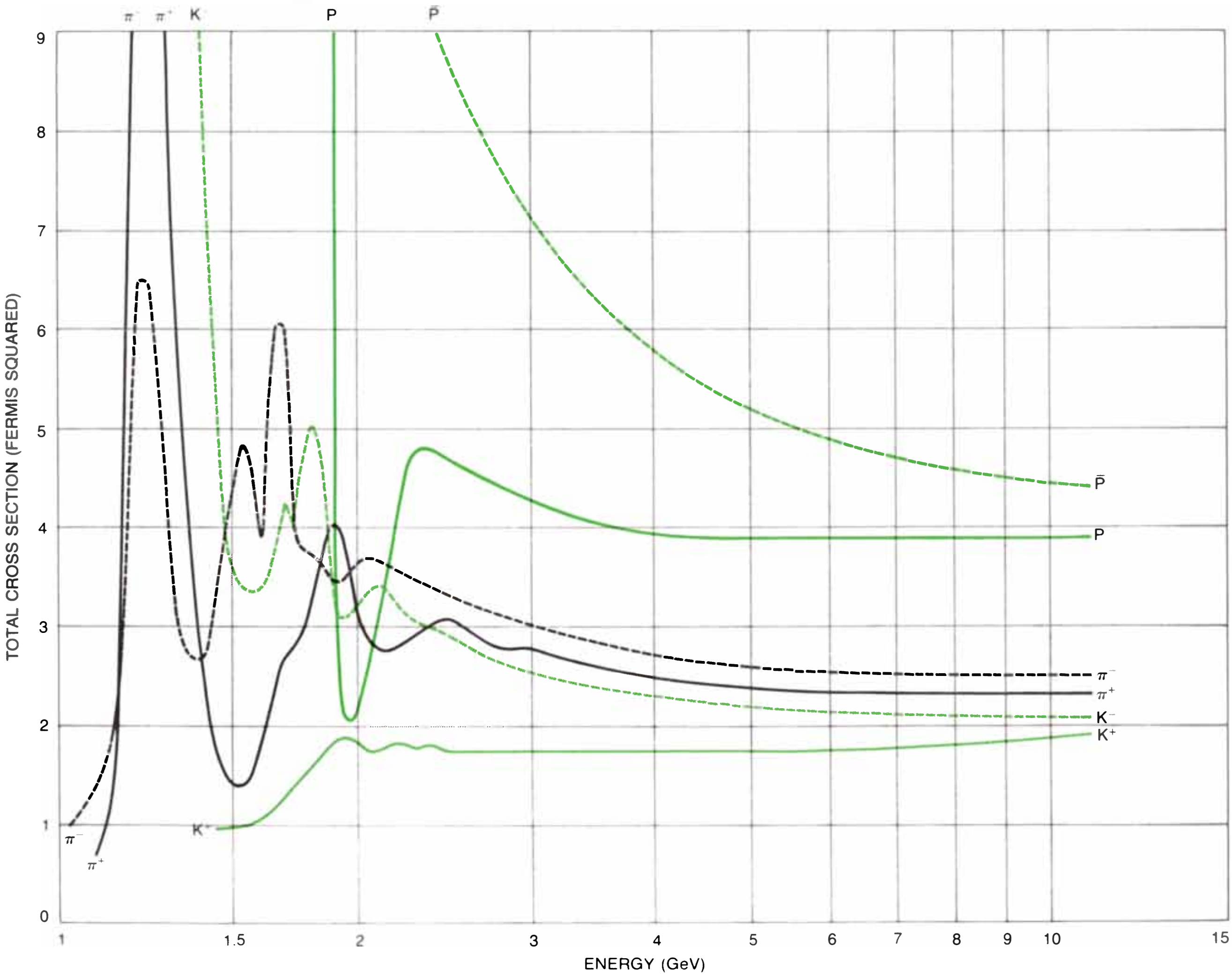
In each of the eight intersection regions the beams, which cross at an angle of 15 degrees, have a horizontal width of between four and seven centimeters and a height of between two and three millimeters. The attainable luminosity depends on the stored currents and on the dimensions of the region of space in which the beams cross. In other words, the luminosity depends on the density of each beam regarded as a target for the other. In order to make the recent measurements of the total proton-proton cross section the luminosity of the machine was determined to a high accuracy by applying a method suggested in 1968 by Simon Van der Meer of the CERN staff. Indeed, it was the accuracy with which this method could be applied that enabled us to detect the quite small increase that was observed in the total proton-proton cross section. Without describing the Van der Meer method in detail, suffice it to say that it requires the vertical displacement of the two proton beams in small and very accurately known steps. These displacements are obtained by means of magnets that produce (some 20 meters from the crossing point) accurately known horizontal magnetic fields. The CERN group has been able to displace the two beams in half-millimeter steps with errors smaller than a hundredth of a millimeter, making possible luminosity measurements with errors as small as ±2 percent.

The original design aim for the machine's luminosity (4×10^{30} interactions per square centimeter per second) was reached at the end of 1972. Compared with what can be obtained in conventional proton accelerators when a proton beam impinges on a hydrogen target (5×10^{36} interactions per square centimeter per second), the luminosity of the CERN intersecting storage rings is about a million times smaller. Even so, the interaction rate at each of the eight intersection regions is more than 100,000 events per second. That is comparable to the rate in most experiments in conventional accelerators, which do not use the accelerated primary proton beam directly but rather employ the secondary beams of pions and kaons produced by the primary beam in a target. Moreover, the advantage of the intersecting storage rings with respect to the center-of-mass energy is best conveyed by the fact that the energies of the beams stored are in the range from 11 to 31 GeV, corresponding to center-of-mass energies that can be varied at will between 22 and 62 GeV. To span the same energy range a proton synchrotron would need to accelerate protons from a minimum of 300 GeV to a maximum of 2,000 GeV.

Of the many experiments performed at the CERN facility during the first two years of operation, I shall concentrate on the two that led to precise measurements of the total proton-proton cross section, thereby providing the evidence that this quantity increases with energy. The results have been obtained by two groups using two different techniques, so that the agreement on the conclusion is a good check of its reliability.

One of the experiments was performed by a collaboration between a group at CERN (J. V. Allaby, W. Bartel, G. Cocconi, A. N. Diddens, R. W. Dobinson and A. M. Wetherell) and a group from the physics laboratory of the Istituto Superiore di Sanità in Rome (C. Bosio, R. Biancastelli, G. Matthiae and me). We have devoted our efforts to the measurements of elastic proton-proton scattering at very small angles with respect to the beams. These small angles were attained by introducing a system of small counters in special movable sections of the vacuum chamber. In this way we could identify a few scattered protons per second in our counters at a distance of about two centimeters from the main flux of 10^{20} protons per second forming the circulating beam.

In an early experiment we had extended the measurements of the width of the diffraction pattern at high energies in order to determine the energy-dependence of the proton-proton interaction radius. The results, shown in the top illustration on page 43, clearly show that the interaction radius continues to increase with energy up to about 60 GeV. In a later experiment we accurately measured the rate of elastic events at four center-of-mass energies (23.5, 30.5, 45 and 53 GeV) and at very small angles (about a third of a degree). By measuring the luminosity with the Van der Meer method we were able to obtain the elastic cross section and the probability of



EARLIER DATA, represented here by six curves summarizing total-cross-section measurements made before 1972 for six different projectile-target combinations, suggested that all such cross sections tend toward constant values with increasing energy. The target particles in every case were protons. The projectile particles were various stable and metastable hadrons produced in proton synchrotrons. The six projectile hadrons used were protons (P), antiprotons (P̄), positive kaons (K^+), negative kaons (K^-), positive pions (π^+) and negative pions (π^-). The horizontal scale gives the energy of the colliding hadrons in GeV (gigavolts, or billions of electron volts) in the center-of-mass reference system. At energies below about 4 GeV the cross sections present a complex landscape of peaks and valleys; above that value, however, they become smooth functions of the energy and tend toward constant values.

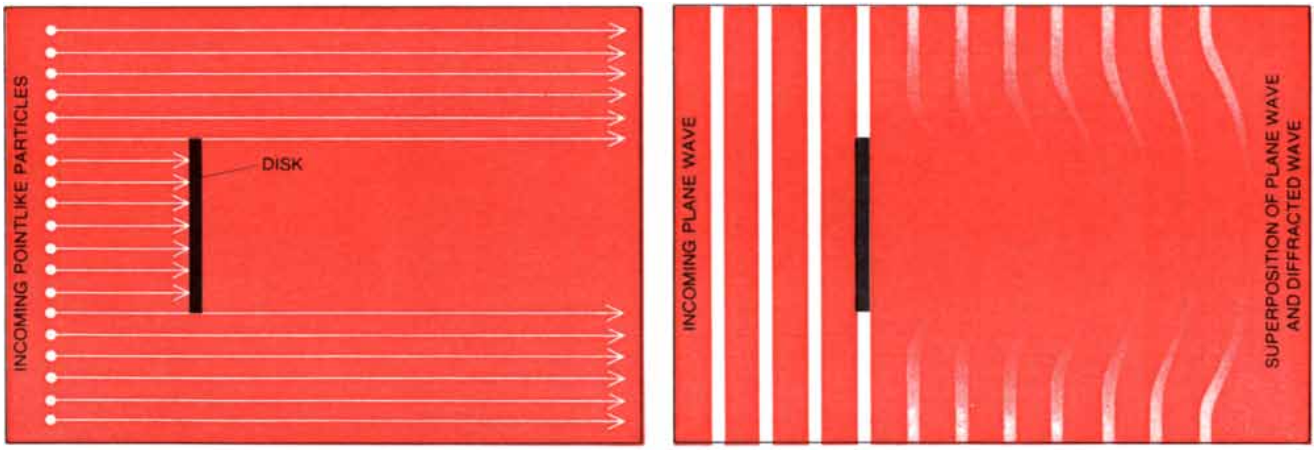
elastic scattering in the forward direction and hence, by virtue of the optical theorem, the total cross sections at the four energies mentioned. From the measured values of the elastic cross section and the total cross section the average proton-absorptiveness can be computed. The four new values, plotted in the bottom illustration on the opposite page, show that in this new energy range the absorptiveness, instead of decreasing continuously, remains practically constant. The new data points for the total proton-proton cross section are presented in the illustration on page 44 together with the lower-energy proton-proton data.

Our method for obtaining total cross sections needs a delicate but compact detection system; it is, however, indirect, in the sense that it utilizes the optical theorem to go from the measured elastic probability to the total cross section. A collaboration between a group from the University of Pisa (S. R. Amendolia, G. Bellettini, P. L. Braccini, C. Bradaschia, R. Castaldi, V. Cavasinni, C. Cerri, T. Del Prete, L. Foà, P. Giromini, P. Laurelli, A. Menzione, L. Ristori, G. Sanquinetti and M. Valdata) and another from the State University of New York at Stony Brook (G. Finocchiaro, P. Grannis, D. Green, R. Mustard and R. Thun), working at a different intersection region, detected the rate of all the events in which two protons interact. Their investigation used a very large system of some 500 counters. This method has the advantage of being direct, but it calls for a very large apparatus and is more sensitive to errors in the measurement of the luminosity. The results of the Pisa–Stony Brook experiment are also plotted in the illustration on page 44.

The results of the two experiments are in excellent agreement and show that the proton-proton cross section increases by about 10 percent when the center-of-mass energy goes from 23 to 62 GeV. In the same energy range the interaction radius also increases and the average absorptiveness remains practically constant. We are thus led to modify the picture that had emerged from lower-energy experiments: In the energy range of the CERN experiments two colliding protons do not behave as objects whose absorptiveness decreases as the radius increases, so that the total cross section remains constant. Indeed, the new results show that as the average interaction radius increases and the average absorptiveness remains constant, the total cross section increases. This behavior definitely departs from what is observed at lower energies.

As often happens in performing experiments in a new energy range, a landscape that was predicted to be flat and uninteresting has revealed unexpected features. We now know that the proton-proton total cross section increases in the energy range opened up by the CERN intersecting storage rings, and that this phenomenon is due to an increase in the interaction radius accompanied by an average absorptiveness that is almost independent of energy. What is the cause of this behavior? Since the absorption is due to inelastic processes, the explanation has to be found in the detailed dynamics of the proton-proton collisions in which other hadrons are produced. Lacking a satisfactory theory of strong interactions, the first questions to be answered are necessarily very simple ones. For instance, it would be interesting to know whether the inelastic processes, whose probability increases with energy and causes the increase in absorptiveness, are characterized on the average by a large number of produced hadrons or a small number.

It is too early to answer even this simple question, but various theoretical models have been proposed to deal with it. At least four fairly simple conceptions are worth mentioning. In the first the increasing cross section is attributed to glancing collisions where one of the two protons remains intact while the other "breaks" into a small number of very energetic hadrons. In the second the inelastic processes, in which many hadrons of comparatively low energy are produced, become more probable when the energy is increased and are responsible for the observed effects. In the third the increasing interaction radius and the almost constant absorptiveness are tentatively attributed to the observed increasing probability of producing proton-antiproton pairs in the interaction of two high-energy protons.



QUANTUM-MECHANICAL PICTURE of a hadron-hadron interaction is based on the assumption that each moving hadron behaves more like a wave than like a little ball. In the diagram at left a large number of particlelike hadrons are shown moving along parallel paths and being absorbed by a hypothetical hadron-absorbing disk. In the diagram at right a plane wave is shown being absorbed by the same disk. In the first case all the particles are absorbed; as a result a "hole" appears in the incoming beam and extends to an infinite distance. This does not happen in the case of the incoming plane wave, because the absorption of a part of the wave front modifies the wave as a whole. As a result a "shadow" extends for a finite distance behind the disk, and the part of the wave that propagates beyond the disk has a complicated structure. This complication is mainly caused by the fact that the full wave is a superposition of two simpler waves: the incident plane wave, which continues to propagate in the initial direction, and a diffracted spherical wave, which moves outward from the disk. The absorbed wave corresponds to particles that are scattered inelastically by the target, that is, particles whose energy is converted into other forms of radiation (such as new hadrons) that emerge from the disk. Diffracted wave corresponds to particles scattered elastically, that is, particles that emerge with same energy as that of incoming particles.

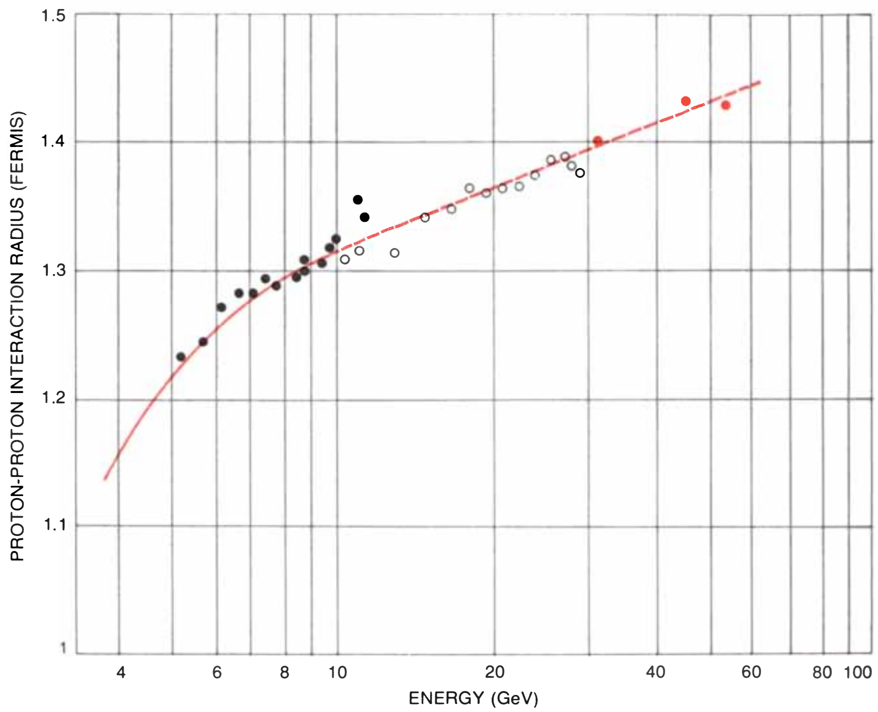
In the fourth, and even more hypothetical, conception the increasing cross section is assumed to be associated with the relatively abundant production of high-energy hadrons emitted at right angles to the direction of the colliding beams. This effect was first observed at CERN by a team of physicists from CERN, Columbia University and Rockefeller University and could be connected to the presence of substructures in the proton, named partons.

Rising cross sections and varying absorptiveness profiles are merely simple manifestations of the nature of inelastic processes at high energies. Only future studies of these processes will reveal the mechanism responsible for the observed increase of the proton-proton total cross section. As we await the results of such experiments at least two further questions arise quite naturally: Do other hadron-hadron cross sections increase with energy? Does the proton-proton cross section increase at energies even greater than the ones reached in the CERN intersecting storage rings?

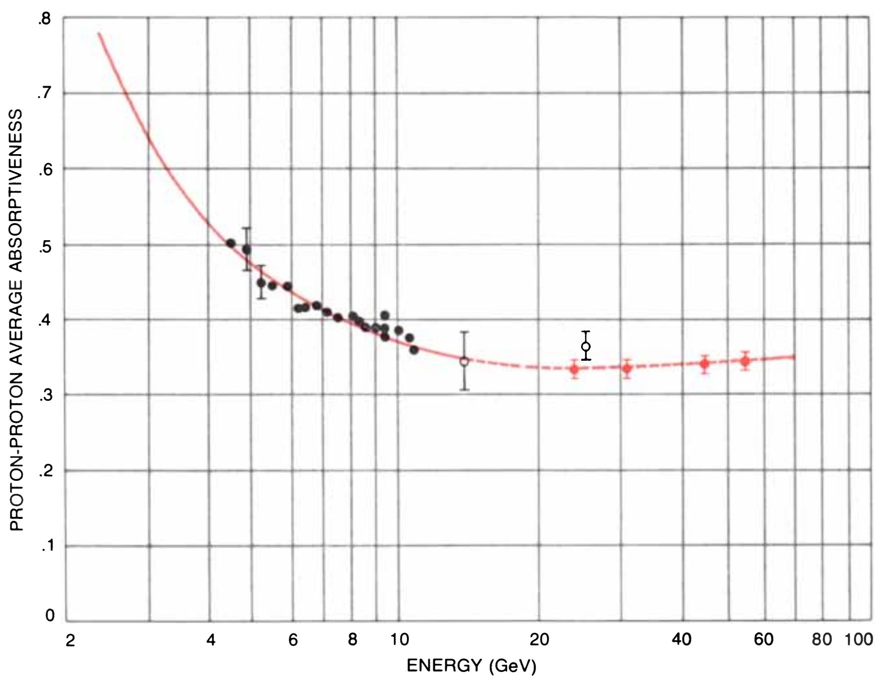
New data bearing on the first questions will soon become available from experiments being set up at the National Accelerator Laboratory in the U.S. There will be much interest in the results obtained with beams of positive kaons, because it is already known that the total cross section of these particles increases slightly at energies higher than 5 GeV.

For the proton-proton cross section, however, the only reliable source of information at energies higher than the ones attainable in the CERN intersecting storage rings comes from cosmic ray experiments. Unfortunately these data are very difficult to interpret. Nonetheless, it is worth mentioning that in 1972, before the new CERN results, an analysis of the absorption of cosmic ray protons in air was published by Gaurang B. Yodh of the University of Maryland, Yash Pal of the Tata Institute of Fundamental Research in Bombay and James S. Trefil of the University of Virginia. They concluded that the proton-proton cross section rises in the center-of-mass energy range from 20 to 200 GeV. At the time it was generally believed this conclusion was not compelling, owing to uncertainties in the data. The recent CERN results have shown that the total cross section increases in the energy range from 20 to 60 GeV and thus support the analysis of Yodh, Pal and Trefil.

Could a hadron-hadron cross section, in particular the proton-proton one, increase indefinitely with increasing energy? The question will never be an-



PROTON-PROTON INTERACTION RADIUS, obtained by detecting elastic-scattering events and measuring the width of the central diffraction peak, is plotted on this graph as a function of energy in the center-of-mass system. At low energies only the trend of the data is shown (*solid colored curve*). The black dots and the open black circles represent data obtained with conventional proton accelerators at Serpukhov in the U.S.S.R. and at the National Accelerator Laboratory in the U.S. respectively. These data points indicate that below 10 GeV interaction radius increases rapidly with energy. Colored dots are averages of the values obtained by two groups with storage rings at CERN. The broken colored line clearly shows that the interaction radius continues to increase uniformly beyond 10 GeV.



PROTON-PROTON AVERAGE ABSORPTIVENESS, computed from the measured values of the elastic cross section and the total cross section, is plotted in this graph. Again the black dots represent the Russian data, the open black circles the U.S. data and the colored dots the CERN data. Until about a year ago only the data below 10 GeV were available. They show that at these relatively low center-of-mass energies (where the total cross section is practically constant) the average proton-proton absorptiveness decreases quite rapidly. The four new data points, obtained by a group working at the CERN storage rings, show that at about 20 GeV the trend changes and the absorptiveness remains practically constant.

swered experimentally, because collisions at infinite energies will never be measured. Nonetheless, the attention of various theoretical physicists had already been attracted by the problem. The recent results have greatly revived interest in their conclusions. The main conclusion can be stated in simple form: A hadron-hadron cross section can increase indefinitely with increasing energy, but the rate of increase is severely limited. That limit is reached when the inelastic processes are such that the absorption is complete up to a certain radius; at the same time the radius must increase at the maximum rate allowed by very general properties of strong interactions, such as the causality principle applied to the microscopic world.

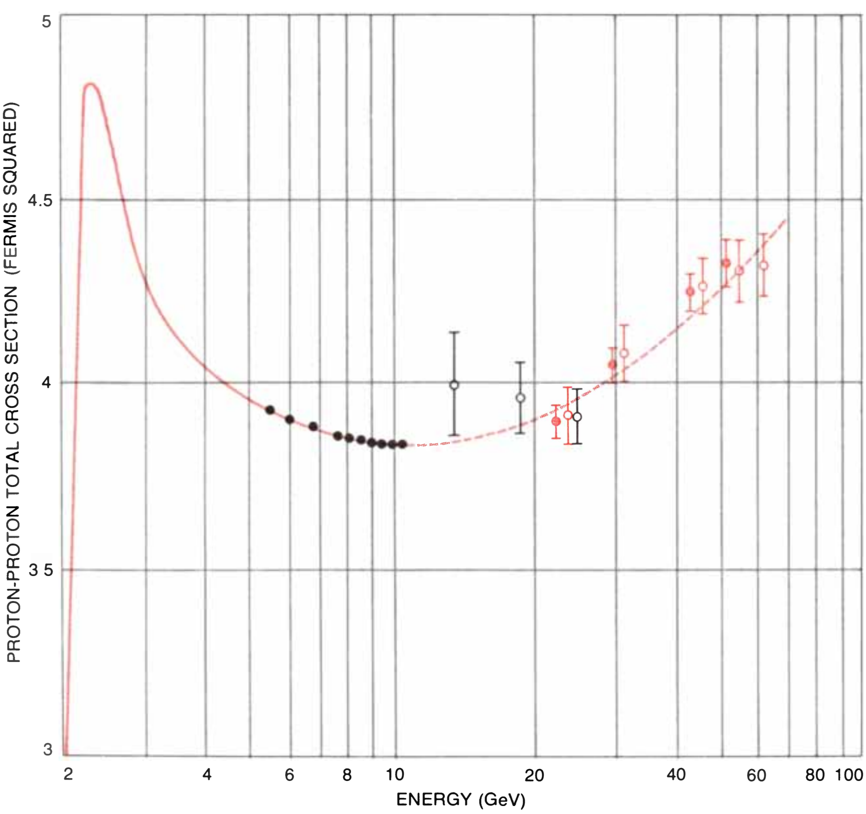
From quite general assumptions the French theorist Marcel Froissart in 1961 derived the conclusion that if in a hadron-hadron collision the absorption is complete, then the interaction radius cannot increase faster than the logarithm of the energy. The experimental proof of a violation of this "Froissart bound" would shake the foundations of our description of strong interactions; so far the CERN experiments show that the measured rise of the proton-proton cross section has nothing to do with such a saturation because the absorption is still far from complete. The experiments do show, however, that models of inelastic processes that "saturate" the Froissart bound by predicting complete absorption at very high energies are not at all excluded, as was widely believed to be the case before.

Actually a few theorists were always convinced that the saturation of the Froissart bound was a satisfactory solution to high-energy dynamics. It is worth noting that as early as 1957 Werner Heisenberg put forward a theory that predicted an increase of the total proton-proton cross section essentially at the limit of the Froissart bound. In 1970 Hung Cheng of the Massachusetts Institute of Technology and Tai Tsun Wu of Harvard University predicted the saturation of the bound on the basis of a theory in which hadrons are assumed to interact with forces similar to (but much stronger than) the electric forces acting between charged particles. In this model all hadron-hadron cross sections increase indefinitely, so that at sufficiently high energy all hadrons appear to one another as completely absorbing objects of increasing radius.

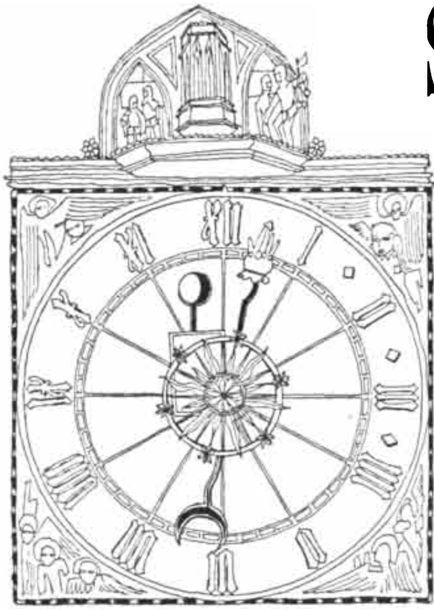
Our results do not prove that this happens; they do show once again, however, that scientists must always keep an open mind. Now we know that a proton does not view another proton as a tiny cloud of matter that expands as the energy increases and becomes less and less absorbing. In a large, although still limited, energy range the interaction radius increases as within that radius the average absorptiveness remains practically constant. What happens at even higher energies we do not know, but certainly complete absorption within an ever expanding radius has to be seriously considered. That solution would have the merit of being aesthetically pleasing, because it would imply that strong interactions are such that at high energies the cross sections increase as fast as they are allowed to.

What are the practical consequences of the rising proton-proton cross sections? The answer is connected with the motivations of high-energy physics. One motivation is the desire to understand the behavior of matter at very high densities. The density of ordinary matter is on the order of one gram per cubic centimeter. At that density electromagnetic interactions dominate the phenomena. The density of nuclei is 10^{14} times greater; such densities are reached on a macroscopic scale in neutron stars. The density that is produced when two hadrons collide head on in the intersecting storage rings is some 500 times greater than the density of the nucleus. The research I have discussed here is a first glimpse into the behavior of matter at those densities. It must be noted, however, that in our experiments such densities have a very short lifetime (about 10^{-14} second), because the interacting protons move with a velocity close to the velocity of light and the distances at which their interaction is strong are on the order of one fermi.

Are there any situations in nature where densities of this kind are achieved for a longer time? If the "big bang" theory of cosmology is correct, such densities were reached during the first milliseconds of the life of our universe. Moreover, those densities must exist even today if, as many astrophysicists believe, stars can collapse to form "black holes." In such a process the phenomenon of the rising hadron-hadron cross sections could certainly play a key role.



UNEXPECTED INCREASE in the total proton-proton cross section, discovered recently by two groups using two different techniques at the CERN intersecting storage rings, is represented here by the broken colored curve. As before, the solid colored curve shows the trend of the earlier data; black dots and open black circles denote the data obtained at Serpukhov and the National Accelerator Laboratory respectively. In this case the colored dots summarize the results of the experiments performed with the intersecting storage rings by a team of investigators from CERN and physics laboratory of Istituto Superiore di Sanità in Rome. Open colored circles denote data obtained by Pisa–Stony Brook collaboration.



SCIENCE AND

The State of Science

How does one measure the health of a nation's science and technology, particularly at a time when U.S. science is believed to be suffering from a drastic loss of public support? The National Science Board of the National Science Foundation is attempting to make such a measurement by developing a body of data that "would reveal the strengths and weaknesses of U.S. science and technology, in terms of the capacity and performance of the enterprise in contributing to national objectives." The first fruits of the effort—a set of indicators dealing mainly with resources and therefore concerned more with capacity than with performance—have been published by the board under the title *Science Indicators 1972.*

Although the board avoids drawing conclusions, the picture that emerges from its survey of resources is a descending curve. Examining the international position of U.S. science and technology, the board found that the proportion of the gross national product spent for research and development declined in the U.S. (from almost 3 percent in 1967 to 2.6 percent in 1971), France and the United Kingdom while rising in the U.S.S.R., Japan and West Germany. The number of scientists and engineers engaged in research and development per 10,000 of population "declined in the United States after 1969 but continued to increase in the U.S.S.R., Japan, West Germany and France, with the result that by 1971 the number... for the U.S.S.R. was 37 as compared with 25 for the United States and Japan, 15 for West Germany and 12 for France."

Indicators dealing with resources for research and development showed that in terms of constant (1958) dollars national expenditures declined 6 percent between 1968 and 1971 and rose slightly in 1972. By the same measure the Federal Government's expenditures for research and development declined 12 percent from 1968 through 1971. Moreover, although the share of Federal expenditures going for research and development in areas others than national defense and space exploration rose from 14 percent in 1963 to 27 percent in 1972, defense and space research still took 73 percent of the money in 1972.

National expenditures for basic research, the board found, rose in terms of current dollars from 1960 to 1972, but in constant dollars "spending in 1972 was approximately equal to the 1967 level and some 6 percent lower than the peak year of 1968." The decline "was least in universities and colleges (3 percent) and largest in industry (14 percent)." In constant dollars the funds provided by the Federal Government for basic research declined 10 percent from 1968 to 1972. Moreover, "Federal support for young investigators (those holding a Ph.D. less than seven years) in universities and colleges declined to a greater extent than support for senior investigators."

Measuring the research and development "intensiveness" of U.S. industry by comparing the ratios of research and development expenditures to net sales and of research and development scientists and engineers to total employment, the board found that the figure rose between 1960 and 1964 "but declined thereafter to a level in 1970 which was lower than in 1960." The largest declines occurred in the industries that are most active in research and development. Another finding was that the proportion of industrial scientists and engineers employed by large companies (5,000 employees or more) rose from 70 percent in 1958 to 85 percent in 1971, whereas the proportion in small companies (1,000 employees or fewer) declined from 20 to 6 percent. Noting "ample historical evidence to suggest that small firms have produced more than a proportionate share of major innovations," the board warned that the changing employment figures "may be a danger signal."

On the Trail of Dioxin

A toxic residue of the U.S. presence in Southeast Asia is a chemical called dioxin, which was present as an impurity in a herbicide, 2,4,5-T, that was sprayed over much of Vietnam. Dioxin's mode of action has not been established; it may inhibit cell division. What is known is that it is cumulative, so that ingestion of even trace amounts for a long time may be dangerous, and it is soluble in fat, so that it may (like DDT) be accumulated and concentrated as it moves up food chains into the human diet. In 1970 the Herbicide Assessment Commission of the American Association for the Advancement of Science sought to determine the presence of dioxin in Vietnamese fish and other materials but found the available analytic methods too insensitive and unreliable. The lethal dose of dioxin in the guinea pig, for example, is .6 microgram per kilogram of body weight, or less than one part per billion; the lowest reported limit of detection for dioxin in whole tissue was 50 parts per billion.

Matthew Meselson and Robert W. Baughman of Harvard University developed an analytic method based on high-resolution mass spectrometry that improved the sensitivity more than a thousandfold. They also added "cleanup" procedures to remove from samples any DDE (a breakdown product of DDT) or polychlorinated biphenyls, both of which might be confused with dioxin in spectrometric results. After applying the improved method to samples from Vietnam they reported last spring that fish and shrimp collected in 1970 near areas heavily exposed to 2,4,5-T contained from 20 to 800 parts per trillion of dioxin, indicating that it "may have accumulated to biologically significant levels in food chains."

Meselson and Baughman have now refined their technique so as to increase the recovery of dioxin from about 30 to about 70 percent. (That is, when a sample of unexposed fish is inoculated with a known amount of dioxin, 70 percent of that dioxin can subsequently be extracted and measured.) They also found ways to make sure that dioxin is not being formed in the course of the analysis itself. At a recent meeting of the American Chemical Society they reported that analysis by the latest method of a carp taken from the Dong Nai River last May (three years after spraying with 2,4,5-T was halted) showed about 400 parts per trillion of dioxin. Many more samples, from different areas and species, would have to be examined to relate dioxin and herbicide spraying and to show whether the chemical is in fact being accumulated in food chains.

Thermonuclear Fission

One of the main attractions of the thermonuclear-fusion approach to the production of power is the promise that the fusion-power plants of the future would be comparatively free of the large burden of radioactive isotopes found in fission-power plants operating at full capacity. What is not generally realized is that most of the fusion-power plants projected so far (specifically those based on the consumption of some form of "conventional," deuterium-containing fuel) would still probably harbor a few percent of the radioisotope inventory characteristic of a fission-power plant of comparable size (albeit in a less hazardous form).

According to a report presented at the recent European Conference on Controlled Fusion and Plasma Physics in Moscow, this situation may be dramatically changed by an "exotic" thermonuclear fuel. In a paper on the U.S. laser-fusion effort A. Thiessen, George Zimmerman, Thomas Weaver, John L. Emmett, John H. Nuckolls and Lowell Wood of the Lawrence Livermore Laboratory described a hypothetical fusion-reactor scheme where the "thermonuclear fission" of the comparatively abundant isotope boron 11, brought about by collisions with protons (hydrogen nuclei) in the thermonuclear plasma, yields energetic alpha particles (helium-4 nuclei). Such a mode of operation, the Livermore group points out, should reduce the plant's total radioisotope inventory to roughly a thousandth of what is estimated for a comparably sized fusion plant burning a deuterium-containing fuel. The resulting level of radioisotopes would be nearly 100,000 times lower than that of a comparable fission plant.

Exotic fuels such as the combination of boron 11 and hydrogen would produce virtually no neutrons or radioactive nuclei during the thermonuclear-burning stage of the reactor's operation. The bulk of such particles would be evolved only in rare subsequent reactions of the high-energy, fusion-born alpha particles with other nuclei in the thermonuclear plasma. According to the calculations of the Livermore group, the theoretical criterion for achieving the "break even" point for net fusion-power output (a product of plasma density and confinement time) is roughly the same for a fuel of boron 11 and hydrogen as it is for a fuel of pure deuterium, although higher ion temperatures would be required for the exotic fuel.

Noting that the performance of an exotic fuel in a fusion-power plant would be limited principally by radiation losses from the plasma, the Livermore group suggested that the comparatively weak coupling between the energy-generating ions and the energy-radiating electrons in the superdense plasmas typical of laser-fusion schemes would make the use of exotic fuels particularly attractive in such schemes. They emphasized, however, that nuclear cross-section experiments currently under way at Livermore and at the California Institute of Technology must be completed before the potential of the exotic fuels for the production of "cheap, abundant, exceedingly clean fusion power" can be fully evaluated.

Black Hole and Blue Star

If black holes exist, they are difficult to detect because they emit no radiation; by definition matter and radiation can only fall into them. One way they might be observed, however, is that if a black hole were in a close orbit around a visible star, matter flowing from the star into the black hole might emit X rays that would be observable just before it disappeared. A leading candidate for this kind of object is the double star HDE 226868 in the constellation Cygnus, which appears to be identified with the celestial X-ray source designated Cyg X-1. The evidence that one of the two stars in HDE 226868 is a black hole was reviewed at a recent International Astronomical Union symposium in Warsaw by Riccardo Giacconi of the Harvard College Observatory and the Smithsonian Astrophysical Observatory.

Cyg X-1 was discovered in 1963. In December, 1971, however, data from the astronomical satellite *Uhuru* revealed that the X-ray emission of the source pulsates. The rate of the pulses varies widely, but the highest rate indicated that the source could not have a diameter greater than 10^9 centimeters (some 6,000 miles, or less than the diameter of the earth). A pulsing source can be no larger than the distance light or other electromagnetic radiation can travel in the interval between pulses.

Subsequently the position of Cyg X-1 was precisely established by observations of its radio emission. The position closely corresponded to the position of HDE 226868 on photographic plates. Spectroscopic observation showed that HDE 226868 was a double star in which each companion rotated around the other once every 5.6 days. One companion was visible: it was a blue supergiant star with a mass on the order of 20 times the mass of the sun. The other companion was invisible, but the dynamical behavior of the system required that it have a mass five or six times the mass of the sun.

The dark companion therefore appears to be a mass five or six times the mass of the sun compressed into a volume smaller than the earth. Such a density is even greater than the density of a neutron star, which is presumably the stage of stellar evolution preceding the collapse of a spent star into a black hole.

Light from the Neutrino Beam

The traditional quest of the physicist is to find a law that will embrace all natural phenomena. In modern times this has meant finding a mathematical theory that would relate the four known types of interaction between material bodies: the electromagnetic interaction, the gravitational interaction, the "strong" interaction and the "weak" interaction. Progress has been slow, but a recent experiment at the European Organization for Nuclear Research (CERN) supports certain theories that unify the electromagnetic and weak interactions.

In the CERN experiment, which was conducted by a group of physicists from five European countries, a beam of neutrinos generated by the 28-billion-electron-volt proton synchrotron was directed into the large bubble chamber named Gargamelle. The interaction between neutrinos and other particles is of the weak type. Since neutrinos have neither mass nor electric charge, interactions between the neutrinos in the CERN beam and the particles in the bubble chamber were extremely rare. Several hundred such interactions were nonetheless observed by the tracks they left.

According to the classic theory of weak interactions, developed by Enrico Fermi, when a neutrino interacts with another particle, it is always transformed into a charged particle; such interactions are therefore called charged-current interactions. The newer theories that relate the electromagnetic and weak interactions, notably those put forward by Steven Weinberg and Abdus Salam, predict that the neutrino can interact with another particle without being transformed into a charged particle; such interactions are called neutral-current interactions.

Of the interactions observed in the CERN experiment 570 had involved the transformation of a neutrino into a charged particle and 160 did not. Similar results have now been obtained in neutrino-beam experiments at the National Accelerator Laboratory in Batavia, Ill. These results lend support to the theories that relate the electromagnetic and weak interactions and predicted the neutral-current events.

"Planned Birth" in China

China has apparently been successful in reducing its rate of population growth through an extensive and highly developed program of "planned birth," according to recent reports reviewed in *Studies in Family Planning*, a publication of the Population Council. The term for planned birth (*chi-hua-sheng-yü*) covers the practice of delayed marriage, modern techniques of contraception (including "the pill," intrauterine devices and sterilization) and induced abortion.

A census taken in 1964–1965 recorded the population of China as being about 700 million with a natural increase rate of just under 2 percent per year. Since then the growth rate in urban areas has been brought down to about 1 percent per year, and in Shanghai, the country's most modern city, the rate of population growth is even less. The crude birth rate in Shanghai is 12 per 1,000, whereas the death rate is five per 1,000, which yields a natural increase rate of seven per 1,000. Peking has a birth rate of 18 per 1,000 and an increase rate of 12.4 per 1,000. The family-planning program has fared less well in the villages, where the traditional preference for many male children is still strong.

Late marriage is encouraged. The ideal age for a man to marry is 30, for a woman 25. The advantages of later marriage are promoted from the standpoint of physiological and emotional maturity, education and occupation, the health and training of future offspring, family economics and living standards. Clinical-survey reports lend support to these advantages. In one such report, dealing with 3,000 women giving birth for the first time, the clinicians concluded that "early marriage and early childbirth have a definite disadvantageous influence on the mental and physical health

of women and that the best age for marriage is after 25 years." Another survey of 4,318 women who had borne one child noted that the incidence of difficult labor increases sufficiently after age 29 as to endanger the well-being of both the mother and the child. On the basis of childbirth complications and maternal mortality, the workers who made the survey suggest that the best time for the birth of the first child is when the mother is between 25 and 29.

Educational and motivational programs concerning planned birth are conducted at the cadre level by fellow workers. Members of the cadre come together to discuss how socialism bears on family planning. ("Nothing should be left to fate" and "The birth of a child would take one away from active participation in production.") It is stressed that it is no longer necessary to have children to guarantee security in old age and that planned birth is not something to be ashamed of. According to the author of one of the reports in *Studies in Family Planning*, Pi-chao Chen of Wayne State University, younger women tend to be more receptive than older ones. In another report Tameyoshi Katagiri of the International Planned Parenthood Federation describes family-planning activities in five areas of China that he visited. In a third report Leo A. Orleans, a specialist on Chinese affairs working in the reference department of the Library of Congress, gives abstracts of virtually all the articles on family planning, abortion, contraception and related subjects that appeared in 18 Chinese medical journals between 1960 and 1966. According to Orleans, "the abstracts... make it clear that considerable activity in family planning took place during the mid-1960's" and that the present accomplishments are the product of this early work.

The Maya Predicted Eclipses

The Maya of Central America were skilled naked-eye astronomers. It now seems that they could even forecast eclipses of the sun. That is the conclusion of a new analysis of certain Maya calculations: the "Venus Table" and the "Lunar Table" contained in the Dresden Codex, one of only three Maya books that survived the Spanish Conquest. Writing in *Journal of the Royal Astronomical Society of Canada*, Charles H. Smiley of Brown University suggests that these first-millennium compilations were efforts to predict when an eclipse of the sun might be visible in the Maya realm.

The Maya were no strangers to eclipses. During the 66 years from A.D. 330 to 396 a total of 20 solar eclipses were visible in their area; six were nearly total and another seven obscured at least half of the sun's disk. Calculating in multiples of their own 260-day "sacred year," Maya astronomers appear to have detected two different kinds of periodicity in the recurrence of eclipses: a "short" interval of 9,360 days (36 sacred years) and a "long" interval of 11,960 days (46 sacred years). For example, of the eclipses visible between A.D. 330 and 396, three comprised a set of the second kind, being spaced 11,960 days apart.

The Venus Table of the Dresden Codex evidently reflects these observations. It consists of 65 columns of dates, written in numerals and glyphs, with three dates in each column. The middle row of dates covers a period of 104 "secular years" of 365 days each, calculated from the 584-day synodic period of the planet Venus. Above the Venus date in each column appears a date that is exactly 11,960 days later than the Venus date and below it appears a date that is exactly 9,360 days earlier. In Smiley's view the "initial" middle-row date is probably that of an eclipse observed near the end of a synodic period of Venus, when the planet is too close to the sun to be seen; perhaps the eclipse recorded is the one of A.D. 344. The other 64 columns, Smiley suggests, represent a forecast of when during the 104-year period other eclipses might be expected.

The Lunar Table of the codex, Smiley notes, contains 70 dates falling within a 11,960-day period beginning in A.D. 477 and ending in 510. He suggests that these are the dates of new moons selected because at the time of the following new moon or the one thereafter, or both, an eclipse was likely. Actually 80 solar eclipses were observable somewhere in the world during the 33-year interval covered by the Lunar Table; each one occurred one or two moons later than the Lunar Table date.

Because the Lunar Table, which Smiley renames the "Solar Eclipse Warning Table," is evidently cyclical in nature, the question arises of whether it would have served to warn of eclipses outside the period covered. As far as solar eclipses visible in Central America are concerned, Smiley finds, the table would have provided satisfactory predictions from A.D. 42 to 886.

The Ecology of Sourdough

Early prospectors in the American West carried with them the ingredients of a highly acidic bread that earned them the name "sourdoughs." The bread has been known for more than 100 years and is now baked commercially in San Francisco, but only recently was the organism responsible for its characteristic sourness isolated and identified as a previously unknown species.

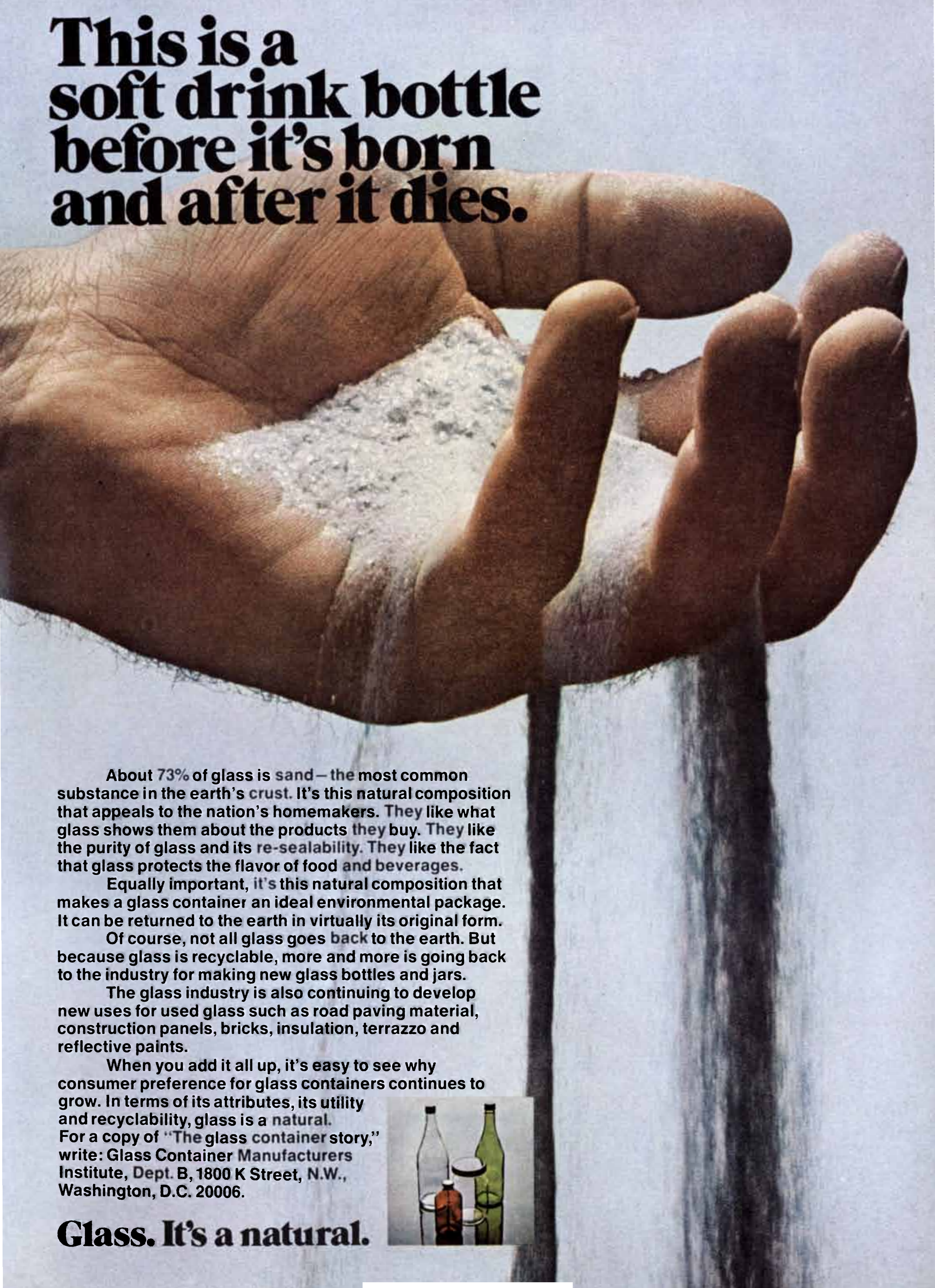
In the making of sourdough, as in the making of other fermented products such as beer and wine, a small portion of each batch is set aside as a "starter" for the next batch. This procedure ensures the preservation of the culture of microorganisms that by fermenting carbohydrate give to such products their essential characteristics: alcoholic content, leavening or, in the case of sourdough bread, acidity.

Leo Kline and T. F. Sugihara of the U.S. Department of Agriculture first isolated the microorganisms of sourdough. They found a fortuitous ecological combination of a yeast, *Saccharomyces exiguus*, and a bacterium, apparently of the genus *Lactobacillus*. For rapid growth the bacteria require the sugar maltose, from which they produce lactic acid and acetic acid, which account for the sour taste of the bread. The yeast is tolerant of this acidic environment, and it ferments carbohydrates other than maltose to produce the carbon dioxide that leavens the bread.

Kline and Sugihara suspected that the bacterium was a previously unreported species; this hypothesis was recently tested at Oregon State University by N. Sriranganathan, Ramon J. Seidler, William E. Sandine and Paul R. Elliker. The uniqueness of the organism was determined by comparing the deoxyribonucleic acid of several species of bacteria with the technique of DNA-DNA hybridization.

The DNA of a culture of sourdough bacteria was labeled with tritium, the radioactive isotope of hydrogen. The DNA was then extracted and its double strands were sheared into fragments and disassociated. It was next allowed to interact with similarly prepared but nonradioactive DNA from species that were believed likely to be related to the sourdough organism. Any recombination of the sourdough-bacterium DNA with the genetic material from other species would indicate a homology, or similarity, between the two. The degree of such recombination, or hybridization, could be measured because of the tritium labeling of one sample.

Only a slight homology was found between the sourdough bacteria and the other species tested, confirming that a new species has been discovered. The name proposed for it is *Lactobacillus sanfrancisco*.


This is a soft drink bottle before it's born and after it dies.
About 73% of glass is sand—the most common substance in the earth's crust. It's this natural composition that appeals to the nation's homemakers. They like what glass shows them about the products they buy. They like the purity of glass and its re-sealability. They like the fact that glass protects the flavor of food and beverages.
Equally important, it's this natural composition that makes a glass container an ideal environmental package. It can be returned to the earth in virtually its original form.
Of course, not all glass goes back to the earth. But because glass is recyclable, more and more is going back to the industry for making new glass bottles and jars.
The glass industry is also continuing to develop new uses for used glass such as road paving material, construction panels, bricks, insulation, terrazzo and reflective paints.
When you add it all up, it's easy to see why consumer preference for glass containers continues to grow. In terms of its attributes, its utility and recyclability, glass is a natural.
For a copy of "The glass container story," write: Glass Container Manufacturers Institute, Dept. B, 1800 K Street, N.W., Washington, D.C. 20006.
Glass. It's a natural.

Raphael · Rubens · Leonardo da Vinci ·

The great masters! Names synonymous with history and great art of yesteryear! Yet their works endure today—priceless, irreplaceable—yours and mine to enjoy—at home and abroad.

Abroad they are collected, preserved and exhibited in the fabulous Uffizi Gallery in Florence, Italy. At home you can view, study and admire these great works of art... the splendid innovative works of the early Renaissance... the fine works of the Florentine period ... the world-famous portraits by the great northern artists.

Yes, you now can view and tour this rare and exciting collection at home.

Through the beautiful pages of *Uffizi/Florence,* Newsweek invites you on a private guided tour of the Uffizi. One of the world's greatest museums, the Uffizi boasts perhaps the richest artistic heritage, including such famous works as Botticelli's *La Primavera* and

Birth of Venus... Leonardo's *Adoration of the Magi*... Rembrandt's *Portrait of an Old Man*... Raphael's *Leo X* ...and many, many other brilliant examples of the incomparable artistry of the great masters.

YOUR TOUR OF THE UFFIZI

You will be met at the door of the Uffizi by the museum's curator, your personal guide on this unique tour. You will see one of the world's richest collections of art—from matchless examples of the Italian Renaissance to the Gallery's German, French, Dutch and Flemish masterpieces. But this is no ordinary museum visit. It is a tour that promises you a lifetime of enjoyment. And it's yours when you examine Newsweek's new book, *Uffizi/Florence.*

Uffizi/Florence is a huge, beautiful volume that reproduces in exquisite full color the finest masterpieces of the Uffizi for you to see right in your own library. You and your family will spend countless pleasurable hours admiring the works of the world's greatest artists—Michelangelo, Uccello, Titian, Van Dyck—a total of 150 beautiful illustràtions, all in full color.

AN UNPRECEDENTED PUBLISHING PROJECT

In the pages of *Uffizi/Florence,* Newsweek has brought together an unsurpassed team of art and publishing experts including specially commissioned color photographers, seasoned art historians, and the peerless printers of Italy's Arnoldo Mondadori Editore. The result is a milestone artbook, first in the distinguished new series GREAT MUSEUMS OF THE WORLD.

Each volume in the series is a personal tour of a different world-renowned museum. *Prado/Madrid*... *Louvre/Paris*... *Vatican Museums/Rome*... *National Gallery/Washington*... *National Museum/Tokyo*... and many more of the world's most distinguished houses of art.

As your introduction to this art series, *Uffizi/Florence* can be yours—free—for ten days without cost or obligation.

The reply card will bring your copy of *Uffizi/Florence* (176 pages, 9¼" x 11½" overall) to examine at your leisure for ten days—without any obligation whatsoever, and without spending a penny. Judge for yourself this volume's beautiful full-color reproduction ...fine art papers... handsome binding and laminated full-color dust jacket—before you decide whether or not to keep the book.

Michelangelo · Botticelli · Rembrandt

FREE PRINT-ON-CANVAS YOURS TO KEEP

PABLO PICASSO
The Lovers

VINCENT VAN GOGH
Sunflowers

PIERRE AUGUSTE RENOIR
By the Seashore*

*The Metropolitan Museum of Art, New York. Bequest of Mrs. H. O. Havemeyer, 1929. The H. O. Havemeyer Collection

Simply for agreeing to examine *Uffizi/Florence,* you may choose—as a free gift, to keep—one of three beautiful 21″ x 27″ masterpieces faithfully reproduced in full color on canvas.

If you're not thoroughly delighted with *Uffizi/Florence,* return it in ten days. Your print-on-canvas is yours to keep. And you owe nothing.

Should you decide to keep *Uffizi/Florence,* just send $6.95 (plus a small postage and handling charge). You will then receive future volumes of Great Museums at regular two-month intervals. Each volume will be offered on the same ten-day free examination basis—at the same low price of only $6.95 each, plus postage and handling.

You are never obligated to buy. There's no minimum number of volumes to order. You're free to cancel whenever you wish.

To order *Uffizi/Florence*—and your free print—mail the reply card today. If the reply card is missing, send your name and address to:

NEWSWEEK/Great Museums of the World
444 Madison Avenue, New York, N.Y. 10022

DELUXE EDITION FOR ADDED ELEGANCE

Because *Uffizi/Florence* is a book that you will keep for a lifetime, you may want to order it—and future volumes in the Great Museums series—in the special Deluxe Edition. This handsome gold-tooled binding has the rich look and texture of expensive leather-bound books—yet you pay only $2.00 more per volume to add this extra touch of elegance to your library.

THE COMPLEMENT SYSTEM

A foreign cell in the body is identified by antibody, but the cell is destroyed by other agents. Among them is "complement," an intricately linked set of enzymes

by Manfred M. Mayer

Immunology is the study of the physiological mechanisms by which men and other animals defend themselves against microscopic invaders such as bacteria, viruses and fungi. The immunological defenses also operate against the malignant transformations of cells that result in cancer and against transplanted foreign tissues or organs. Some immunological phenomena are cellular, in that they involve lymphocytes (cells that mediate immunological reactions) and phagocytes (cells that ingest foreign particles and microorganisms or other cells). Other immunological reactions are humoral, in that they involve substances dissolved in the body fluids, such as antibodies and enzymes. Cellular and humoral phenomena can influence each other. For example, the engulfing of foreign particles by phagocytes is promoted by antibodies.

The role of antibodies is well known [see "The Structure of Antibodies," by R. R. Porter, SCIENTIFIC AMERICAN, October, 1967, and "The Structure and Function of Antibodies," by Gerald M. Edelman, SCIENTIFIC AMERICAN, August, 1970]. Less well known is another important feature of the immune system: attack by complement. The term "complement" refers to a complex group of enzymes in normal blood serum that, working together with antibodies or other factors, plays an important role as a mediator of both immune and allergic reactions. The reactions in which complement participates take place in blood serum or in other body fluids and hence are considered to be humoral reactions.

The discovery of complement came between 1880 and 1890 from studies of the capacity of blood serum to kill certain microorganisms. Antibodies had been discovered a short time earlier, but it was found that their capacity to kill bacteria depended on the collaboration of another constituent of serum: "alexin," or complement. The names were intended to indicate that the agent helps antibody to perform its defensive function. As knowledge of complement has unfolded in the intervening 85 years, it has become evident that the relation between antibody and complement is actually the reverse of what it was originally thought to be. It is now recognized that the invading cells are attacked by complement and that the function of antibody is to identify the invading cell as a foreign organism and activate the complement attack.

Antibody and Complement

The relation between antibody and complement resembles the relation between the ignition key of an automobile and the engine. An antibody molecule has sites that combine with a specific pattern on the surface of a foreign cell or with another molecule: an antigen. The fit between the antigen molecule and the antibody molecule is the fit between the ignition key and the ignition lock. The antibody molecule serves to start the complement system, which, like the automobile engine, does the actual work. The analogy can be carried further: Whereas antibody molecules and keys are relatively simple structures, the complement system and the automobile engine are complex assemblies of many different parts.

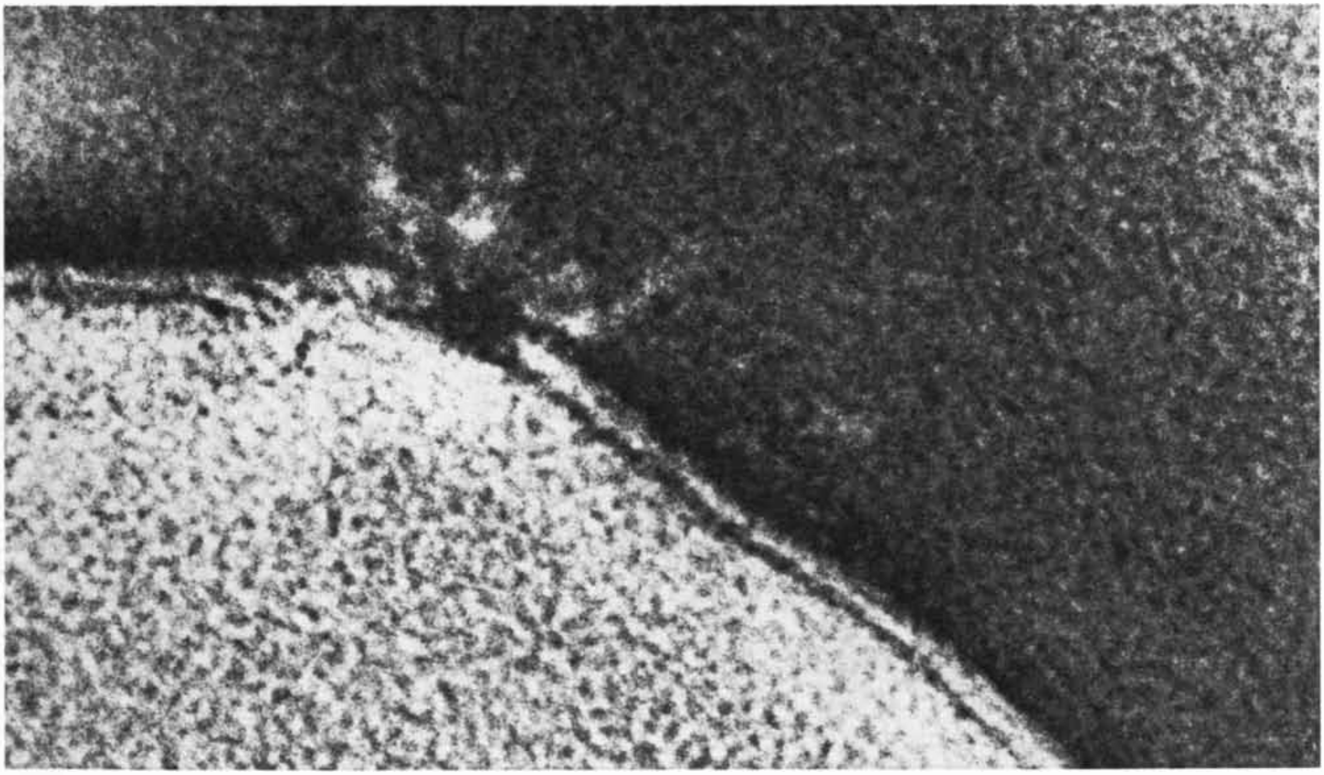
When complement is activated by antibody, it presents a serious threat not only to invading microorganisms but also to the host's own cells. This self-destructive activity is minimized by the fact that antibody fixes complement on the surface of the invading cell. Thus antibody has three specific functions with respect to complement: (1) recognition of the foreign invader, (2) activation of the complement system and (3) fixation of complement on the invading cell's surface.

The complement system also must fulfill three requirements. It must have a recognition unit of its own so that it can respond to the antibody molecules that have detected a foreign invader. It must have receptor sites that will enable it to combine with the surface of the foreign cell when it is activated. And in order to minimize damage to the host's own cells its activity must be limited in time. This limitation is accomplished partly by the spontaneous decay of activated complement and partly by interference from inhibitors and destructive enzymes. The control of complement, however, is not perfect, and there are times when damage is done to the host's cells. Immunity is therefore a double-edged sword. When an immune system acts against a foreign microorganism, the result is protection; when it acts against the host's own cells, the result is the disruption of body systems. Such disruptions are what are known as allergic or hypersensitive reactions.

The Complement Proteins

There are 11 proteins in the complement system. The complement proteins are designated by the letter C and by number: C1, C2, C3 and so on up to C9. The complement protein C1 is actually an assembly of subunits designated C1q, C1r and C1s. From the studies of Irwin H. Lepow of the University of Connecticut Health Center it is known that the subunits are held together by bonds other than the usual covalent ones, and that the calcium ion Ca^{++} is needed to keep the assembly intact. If the calcium ion is removed by chelating agents, the subunits disassemble. The entire assembly is believed to consist of one mole-

LESIONS MADE BY COMPLEMENT in artificial structures called liposomes provide a simple model for studying the mechanism of complement attack. The liposomes (*large circular shapes*) in this electron micrograph were prepared from the lipids sphingomyelin and cholesterol. The lipids form concentric bilayers that enclose water and ions. When the complement proteins C5b, C6, C7, C8 and C9 from human blood serum are added to the liposomes, lesions are formed in the bilayer (*upper right and lower right in micrograph*). There are also two detached lesions (*left of center*), which suggests that once the lesions are formed they are relatively stable. It is believed that the walls of the lesion consist of complement components, possibly in combination with lipids of the bilayer.



FUNNEL-SHAPED COMPLEMENT LESION in a liposome prepared from the lipid lecithin is about 10 or 11 nanometers thick and has a hollow core that must be hydrophilic (water-loving) because it is penetrated by stain. Magnification is some 880,000 diameters. Both electron micrographs are by E. A. Munn and his colleagues at the Institute of Animal Physiology in Cambridge.

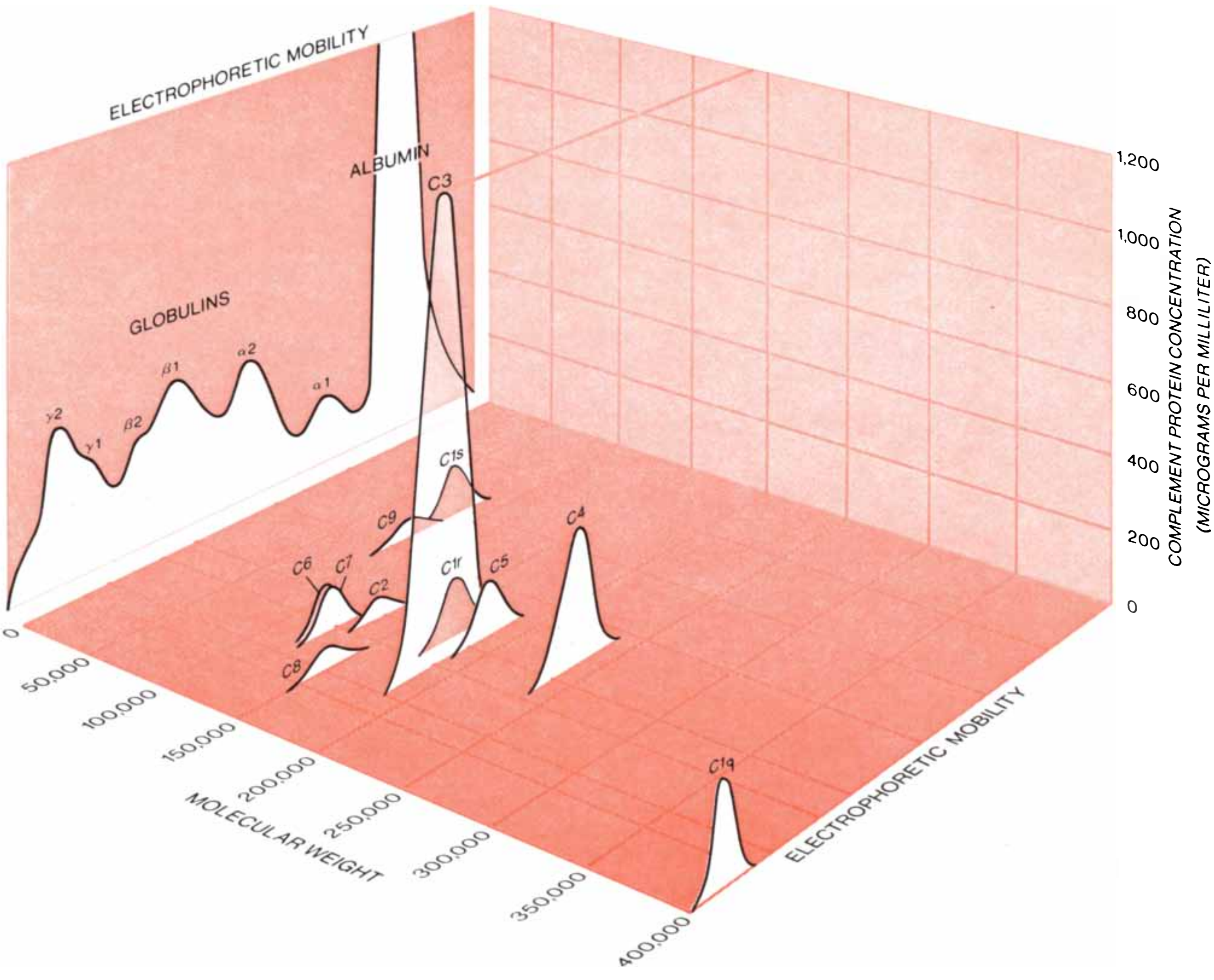
cule of subunit C1q, two molecules of C1r and four molecules of C1s.

The numbers assigned to the complement proteins reflect the sequence in which they become active, with the exception of complement protein C4, which reacts after C1 and before C2. (The numerical assignments were made before the reaction sequence was fully understood.) The recognition of each of the 11 proteins has been achieved through the intensive efforts of many investigators, notably Robert A. Nelson, Jr., of the Lady Davis Institute for Medical Research in Montreal, Kusuya Nishioka of the National Cancer Center Research Institute in Tokyo and William Dean Linscott of the University of California San Francisco Medical Center.

The complement proteins have been isolated in purified form. The concentration of each in human blood serum, along with its molecular weight and electrophoretic mobility, has been determined, largely through the efforts of Hans Müller-Eberhard and his colleagues at the Scripps Clinic and Research Foundation in La Jolla, Calif. [*see illustration below*]. So far little work has been done on the chemical structure of the complement proteins except for the studies of C1q undertaken by Robert M. Stroud of the University of Alabama Medical School and by Müller-Eberhard. These studies reveal that the C1q molecule contains the unusual amino acids hydroxyl lysine and hydroxyl proline. It also contains a large amount of glycine and substantial quantities of carbohydrate (mostly galactose and glucose), which suggests that it is similar in chemical composition to collagen, the principal protein of connective tissue. In explaining the respective functions of the 11 proteins in the complement system, it is convenient to divide the discussion into three parts: recognition, enzymatic activation and the attack by complement factors that results in destruction of the cell.

Recognition

The recognition unit of the complement system is the C1q molecule. It has the capacity to combine with a segment of the immunoglobulin antibody molecules that bind to antigen molecules. This binding of complement to antibody is the basis of the complement-

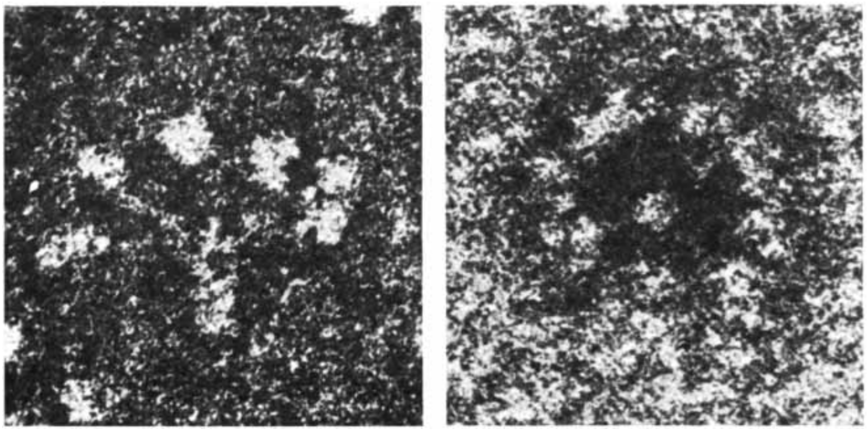


HUMAN COMPLEMENT PROTEINS are characterized by their molecular weight, their electrophoretic mobility (at *p*H 8.6) and their concentration in blood serum. The C3 protein, which has a molecular weight of about 180,000, is the commonest complement protein; its concentration is about 1,200 micrograms per milliliter. The concentrations of the other complement proteins are much lower. For the purpose of comparison the electrophoretic pattern of globulins (gamma, beta and alpha) and albumin in blood plasma are shown. This illustration is adapted from one prepared by Hans Müller-Eberhard of the Scripps Clinic and Research Foundation.

fixation test that has long been used for medical diagnosis (for example in the diagnosis of syphilis).

Only immunoglobulin M (IgM) and several subclasses of immunoglobulin G (IgG) are able to bind the complement factor C1. IgM is a gamma globulin of high molecular weight that appears in the early stages of infection or immunological challenge; IgG has a lower molecular weight but accounts for about 70 percent of the immunoglobulin in normal human blood serum. Other gamma globulins (for example IgA, which is found in saliva and external secretions, and IgE, which is largely responsible for allergic reactions) do not fix C1. Tibor Borsos and Herbert J. Rapp of the National Cancer Institute have shown that in the case of IgM a single molecule of antibody on the surface of a cell is able to bind C1, but that in the case of IgG two adjacent molecules are required for such binding. Since antibodies are scattered more or less at random over the cell surface, the probability of two IgG molecules occupying adjacent sites is quite small, and the frequency with which IgG binds C1 is low. For example, it has been estimated that with the red blood cells of the sheep as many as 800 IgG molecules per cell are needed to create one receptor site for the complement factor C1. On the other hand, only one molecule of IgM per cell is needed.

Antibody molecules change in shape when they combine with antigen, and this event may be responsible for the conversion of the complement factor C1 from an inactive enzyme into an active one. The active site of the enzyme is on the $C\overline{1s}$ subunit (the bar designates activation). It has been suggested that the C1r subunit plays the role of an intermediate agent between C1q and C1s in the activation process. The recognition by Louis Pillemer and Lepow, and by Elmer L. Becker and Lawrence Levine, that $C\overline{1s}$ is an enzyme, and my own subsequent discovery that its enzymatic action cleaves the complement factor C2, opened the way to an understanding of the enzymatic reactions responsible for complement's attack on cells.



MOLECULE OF COMPLEMENT SUBUNIT C1q is enlarged some 1.2 million diameters in these two electron micrographs, which were provided by Emma Shelton of the National Cancer Institute. The molecule of the subunit consists of three distinct parts: a central part, connecting strands and terminal units. The strands (which are possibly single polypeptides, that is, chains of amino acid units) that join the six terminal units to the central stalk can be seen in the "side" view of the molecule (*micrograph at left*). The "top" view (*micrograph at right*) shows the radial arrangement of the terminal units around the central part.

Enzymatic Activation

The second stage involves the complement factors C4, C2 and C3. Their activation is initiated by the enzyme $C\overline{1s}$, and eventually they combine to form another enzyme. The assembly of this enzyme begins with the cleavage of C4 by $C\overline{1s}$ into a large fragment, C4b, and a small fragment, C4a. The C4b fragment has an active site that can combine with a receptor on the surface of the cell's membrane, but this binding site has a short life, and only a small proportion of the C4b fragments that are formed become bound. The unbound C4b fragments quickly become inactive and remain in the blood serum. Many molecules of C4 are cleaved by a single $C\overline{1s}$ enzyme, producing a shower of C4b fragments. Only those fragments that become bound to the cell, however, participate in the subsequent reactions. Little is known about the C4a fragment and its fate or about the receptors on the cell that bind the C4b fragment.

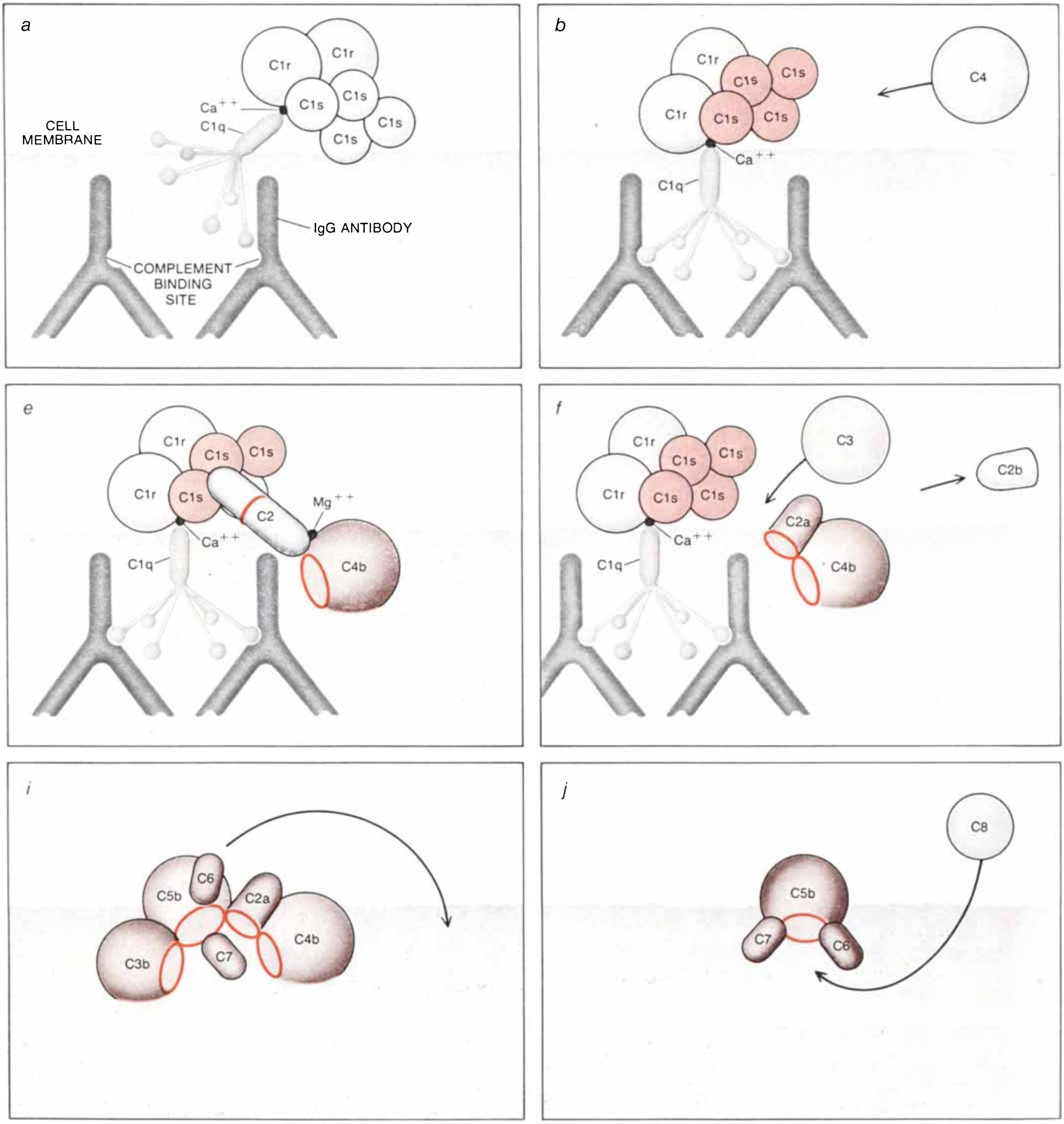
The next step, which has been intensively studied in my laboratory at the Johns Hopkins University School of Medicine, involves the adsorption of the complement protein C2 to the cell-bound C4b. This adsorption is promoted by magnesium salts. Following adsorption the C2 molecule is cleaved by a neighboring $C\overline{1s}$ enzyme into two large fragments, one of which, C2a, becomes bound to C4b. Müller-Eberhard has demonstrated that the $C\overline{4b,2a}$ complex is an enzyme. Nothing is known as yet about the receptor on the C4b fragment that binds C2a.

The complement protein C3 is a natural substrate for the $C\overline{4b,2a}$ enzyme. When they combine, the enzyme splits C3 into two fragments. One fragment, C3a, with a molecular weight of about 10,000, is released into the fluid phase and plays a role as a mediator of inflammation; the other, C3b, with a molecular weight of about 175,000, becomes bound to a receptor on the cell's surface. Since the $C\overline{4b,2a}$ complex combination is an enzyme, it can react more than once and produce a shower of C3b fragments. Only the C3b fragments that become bound adjacent to the $C\overline{4b,2a}$ enzyme, however, are believed to participate in the next reaction, in which the complement protein C5 is cleaved. The C3b fragments that become bound to other sites on the cell surface play an important role as promoters of phagocytosis.

As Hyun S. Shin of Johns Hopkins has shown, the binding of a C3b fragment in the immediate vicinity of the $C\overline{4b,2a}$ enzyme creates a new enzyme that has the capacity to cleave C5. At this point it should be made clear that as the complement proteins are enzymatically cleaved a binding site on the activated complement component is exposed. Because the binding sites have a short life, however, reactivity is soon lost and the complement component disappears as a functional unit. Another matter of importance concerns the instability of the intermediate enzymatic complexes. For example, it has been shown in my laboratory that the $C\overline{4b,2a}$ enzyme is quite stable at zero degrees Celsius, its half-life being about 10 hours. At 37 degrees C. its half-life is only about eight minutes. It loses its activity because the C2a fragment, which contributes the active enzymatic site, migrates into the fluid phase and becomes inactive. The C2a fragment is also released from the $C\overline{4b,2a,3b}$ enzyme. These decay processes may be regarded as one of the factors that limit the ability of the complement system to attack the host's cells.

In the first step of the attack on a cell the complement factor C5 is activated through cleavage by the $\overline{\text{C4b,2a,3b}}$ enzyme. The C5a fragment, with a molecular weight of about 15,000, drifts off into the fluid phase, and eventually plays a role as a mediator of inflammation. The larger fragment, C5b, with a molecular weight of about 170,000, may remain on the $\overline{\text{C4b,2a,3b}}$ enzyme as it combines with the complement proteins C6 and C7, or it may form a complex only with C6 before it dissociates from the enzyme and combines with C7 in the fluid phase. After the $\overline{\text{C5b,6,7}}$ complex is formed, it binds to the cell membrane.

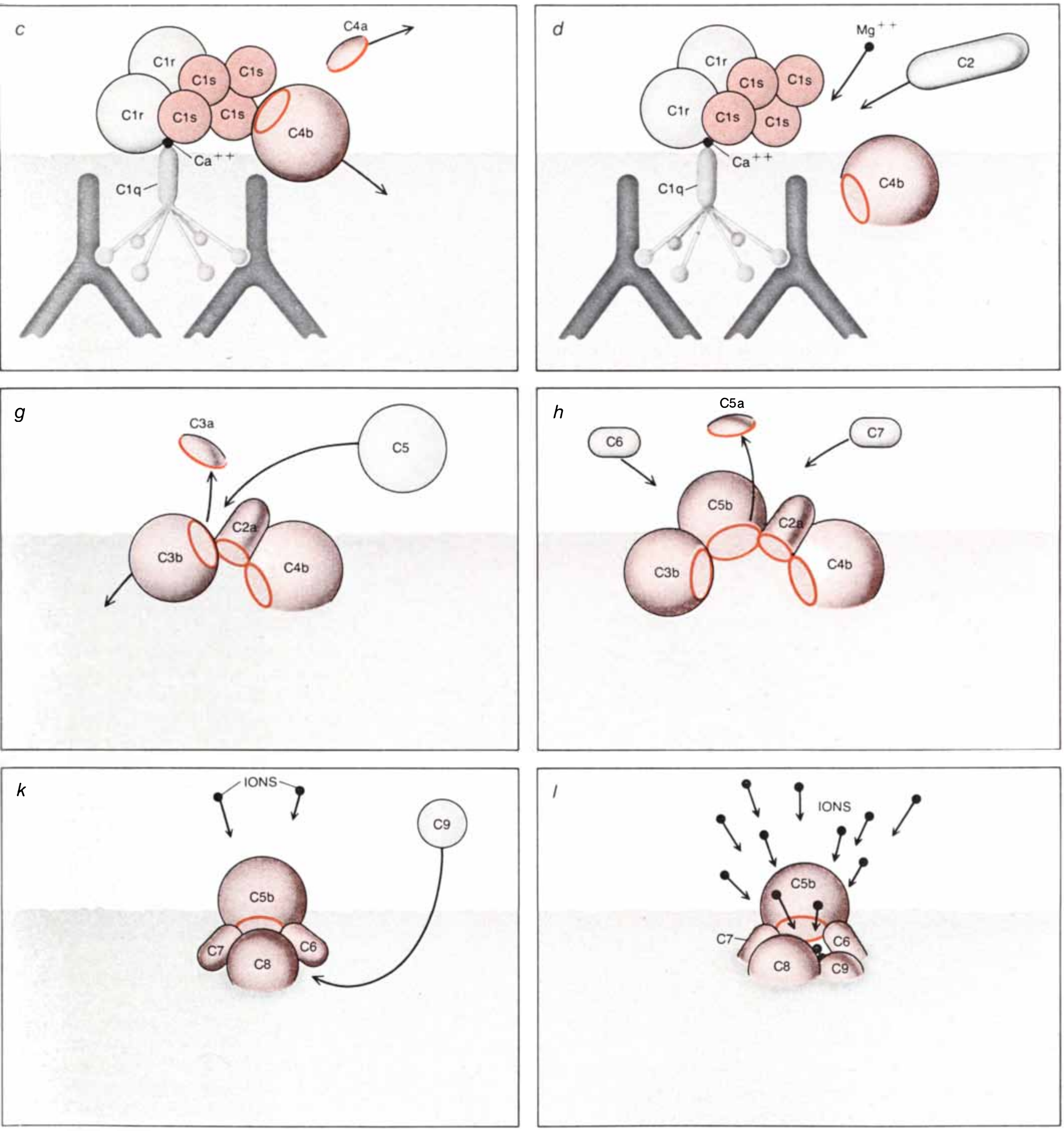
It is believed that the $\overline{\text{C4b,2a,3b}}$ enzyme is freed after the dissociation of the C5b,6,7 complex, and that it may then activate another molecule of C5. Direct experimental evidence demonstrating this capacity, however, has not yet emerged. Moreover, although it is be-



CLASSICAL PATHWAY of complement attack on the cell membrane is depicted. Foreign cells are recognized by antibodies, which bind to antigenic sites on the cell's surface. When two immunoglobulin G (IgG) antibodies are bound to adjacent sites, they can activate complement factor C1, which is inactive until it binds to the antibodies. C1 consists of three subunits, C1q, C1s and C1r, held together by a calcium ion. The C1q subunit is able to bind to the complement binding sites on antibodies (*a*). When it is bound, the C1 complex becomes enzymatically active and will activate complement protein C4 that comes in contact with a C$\overline{1}$s subunit (*b and c*). C4 breaks into two parts, C4a and C4b, and the latter binds to the cell surface nearby (*d*). When C2 comes in lieved the C5b,6,7 complex becomes bound to a site distinctly different from the site on the generating enzyme, there is no direct evidence that this happens. There is some evidence that the C5b,6,7 complex transfers to a different site: the C5b,6,7 complex can be transferred from cells with the C4b,2a,3b enzyme to other cells without the enzyme.

Cell Attack

The cell-attack sequence is initiated with the cleavage of protein C5. What role does the protein C6 play? Increasing the temperature or the ionic strength of the fluid phase causes C5b to break away from the C4b,2a,3b enzyme. Once the C5b fragment leaves the activating enzyme, its ability to combine with C6 and C7 is lost. When C6 is combined



contact with the activated C1s (*e*), it too is split. The C2a fragment combines with C4b to form an enzyme, which splits C3 (*f*). The C3b fragment binds to the surface (*g*). If it is near enough to the C4b,2a enzyme, together they bind C5 (*h*). C6 and C7 bind to C5b (*i*). The C5b,6,7 complex then binds to the cell surface at a new site (*j*). C8 joins the C5b,6,7 complex. The components assemble themselves in such a way that a small hole is formed in the membrane through which a few ions can pass (*k*). The addition of C9 greatly enlarges the hole and speeds up the flow of water and ions into the cell (*l*), causing it to swell and burst. The C3a and C5a fragments produced also play a role in immune and allergic reactions; they cause the release of histamine from cells.

with the C5b fragment, a fair degree of stability is achieved. The C5b,6 complex can be isolated and kept in the fluid phase for extended periods without much loss of its reactivity in subsequent steps. Hence it appears that the C6 complement factor serves as a stabilizer of the activated C5b fragment. It has been shown that C6 itself is not cleaved when it combines with the C5b fragment. It can be recovered in its complete form simply by dissociating the C5b,6 complex.

The complement factor C7 has not been extensively studied. Müller-Eberhard has shown that it binds to the C5b,6 complex. Robert Thompson of the University of Birmingham and Peter Lachmann of the Royal Postgraduate Medical School in London have shown that the combination of C7 with the C5b,6 complex, either in the fluid phase or on the generating enzyme, results in the formation of a new binding site that can attach to the cell's membrane. It is possible that C7 activates a binding site on the C5b,6 complex, or that the C7 molecule itself develops a binding site. The active binding site is short-lived, and the C5b,6,7 complexes that do not couple with the cell's membrane promptly lose their activity and remain in the fluid phase.

The reactions involving the complement factors C8 and C9 are not yet well understood. It is known that both bind by reactions other than covalent ones. C8 combines first with the C5b subunit of the C5b,6,7 complex and then C9 combines with C8. It is believed these reactions take place after the C5b,6,7 complex is bound to the cell's membrane, but interaction in the fluid phase is certainly possible. Cells subjected to the entire complement sequence up to and including C8 dissolve very slowly. The addition of the complement factor C9 greatly accelerates the destructive process.

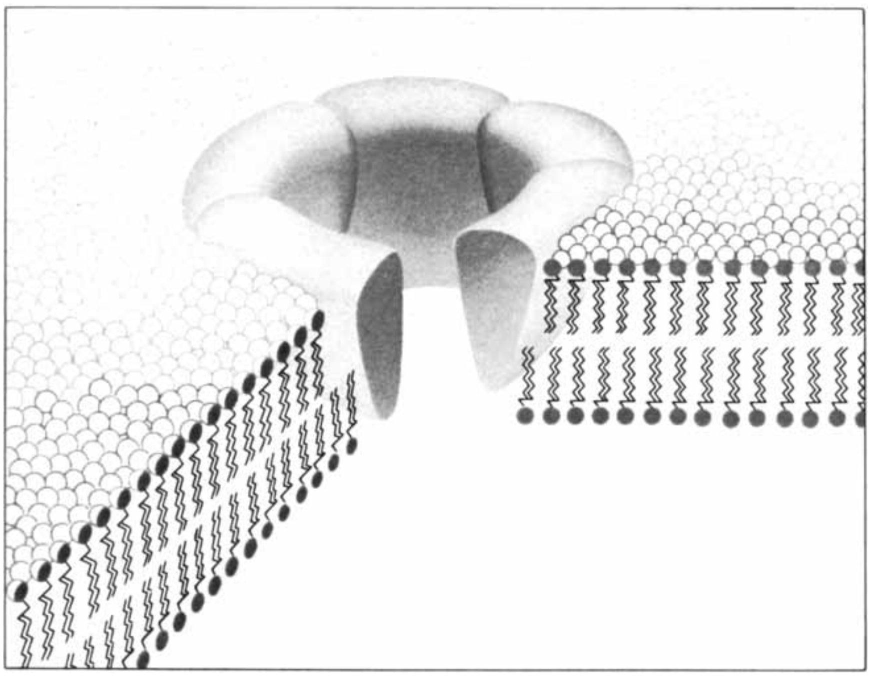
It should be evident from the many qualifications and uncertainties in this description of the complement sequence that far less is known about the late-acting complement components than about the early-acting ones. The late-acting components present fascinating problems in immunology that are being investigated in many laboratories.

The Properdin Pathway

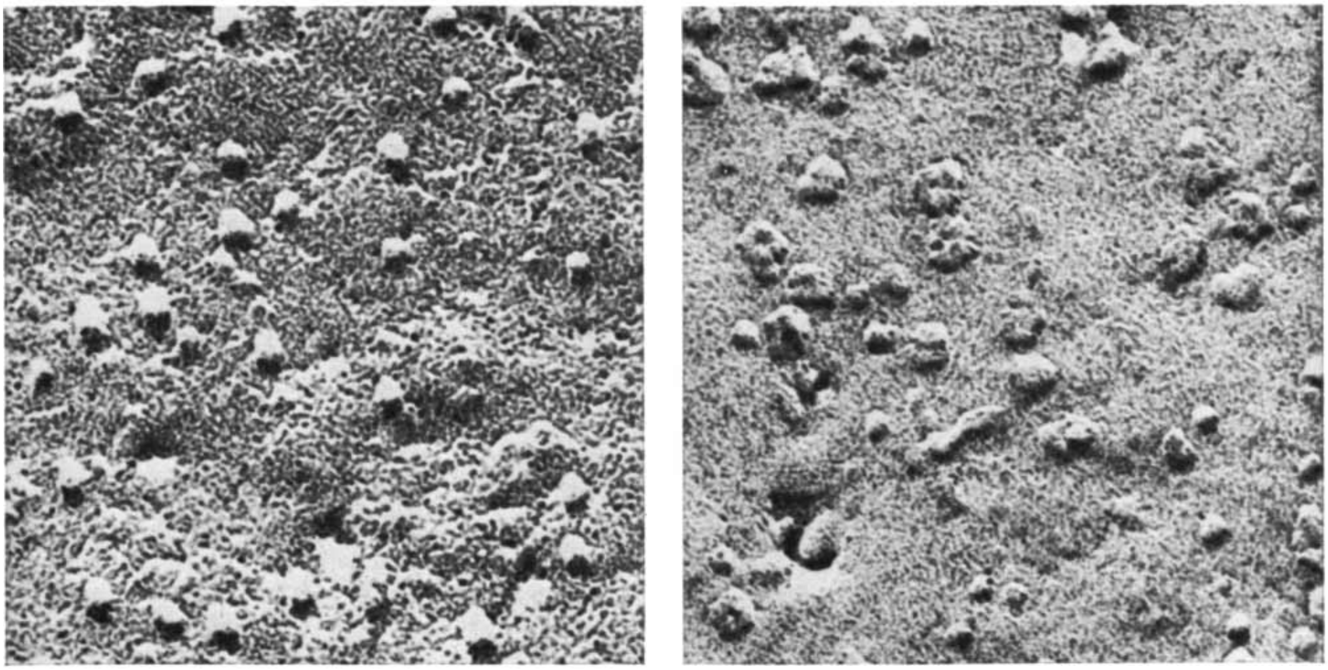
The activation of C3 by way of antibody, C1, C4 and C2 is designated as the classical pathway. An alternate pathway called the properdin pathway was discovered by Pillemer in the 1950's, and it is currently under intensive investigation. Some believe that it does not require antibody to initiate it and that hence it may be a more "natural" or "nonspecific" mechanism of immune defense. Such a mechanism might obviously be of great importance in infections and possibly in other diseases as well.

The term properdin refers to a system in blood serum that, acting together with complement, participates in several immunological processes, notably the promotion of the engulfing of cells and foreign particles by phagocytes and the production of inflammatory reactions. The incubation of normal blood serum with microbial cells or with certain polysaccharides derived from microbial cells (such as zymosan, a carbohydrate of the yeast cell membrane) gives rise to enzymes that activate the complement factors C3 and C5. Gram-positive bacteria (such as the pneumococcus) and toxic lipid polysaccharides from gram-negative bacteria (such as the colon bacillus *Escherichia coli*) also activate the properdin pathway. Once activated, the properdin system enzymes assemble on the surface of the bacterial cell and activate the complement attack sequence, beginning with production of the C3b fragment [*see illustration on page 62*]. Furthermore, the biologically active fragments C3a, C3b and C5a are also produced. These products mediate inflammatory reactions involved in immune defense and in allergic processes.

It is thought that there are two distinct enzymes in the properdin system: one that activates C3 and another that activates C5. Both enzymes are multiunit complexes and correspond in function to the $C\overline{4b,2a}$ and $C\overline{4b,2a,3b}$ complexes of the classical pathway. A key to understanding the mechanism of the properdin system comes from recent studies in my laboratory by Shin and Volker Brade elucidating the differences, as well as the similarities, between the properdin enzymes and their classical counterparts. Three of the subunits have been recognized, and each is found in each of the properdin enzymes. One of the subunits, called Factor B, can be regarded as the counterpart of the C2a unit in the corresponding complement enzymes. Factor B is also referred to as C3PA (for C3 proactivator) or as GBG (for gycine-rich beta glucoprotein). The other, Factor D, serves to activate Factor B. The complement fragment C3b is also present and plays a role in initiating the properdin system, but its precise function is not yet clear. One of the questions currently under study concerns the sources of the C3b fragment. A related



STRUCTURE OF COMPLEMENT LESION postulated by the author is depicted in schematic form. The structure is believed to consist of the five late-acting complement components C5b, C6, C7, C8 and C9. These components are thought to attach themselves to the lipid bilayer of the cell membrane. The bilayer consists of lipid molecules stacked side by side with their polar heads pointing outward and their nonpolar fatty-acid tails pointing inward. The complement components, which are proteins, are believed to assemble themselves into a doughnut or a funnel shape that penetrates the bilayer. The hollow core of the structure could form the lesion through which water and ions flow into cell until it bursts.



FREEZE-FRACTURED MEMBRANE of sheep red blood cells appears in these electron micrographs. The fracture takes place between the two lipid layers of the membrane, so that the membrane's interior is visible. In cell membrane treated only with antibody the protein globules that penetrate the membrane are normal in size, shape and distribution (*micrograph at left*). Membrane treated with antibody and complement, however, contains large doughnut-shaped aggregates of globules that penetrate to the interior of the cell membrane (*micrograph at right*). There are several doughnut-shaped aggregates in this micrograph; for example, one can be seen in the lower left-hand corner and two other doughnut shapes can be seen to the left of center. Both electron micrographs, which enlarged the structures some 420,000 diameters, were prepared by Bernhard Cinader and his colleagues at the University of Toronto.

study deals with the role of KAF, a regulatory enzyme discovered by Lachmann that destroys the C3b fragment.

Even though the properdin pathway and the classical pathway are initiated differently, various complement components exert an influence on the properdin-system enzymes. The complement factor C4 has been found to accelerate the assembly of the properdin enzymes, and it seems likely that C1 and C2 are also involved in the acceleration, although this has not yet been demonstrated directly. The accelerating effect is probably due to the generation of nascent C3b fragments by the complement enzyme $\overline{C4b,2a}$.

One important question about the properdin system is whether or not any antibody is required to initiate it. Since normal blood serum contains at least some antibodies capable of reacting with virtually any bacterial surface, it is possible that in the properdin system the foreign invader is recognized by antibody, just as it is in the classical pathway. Whether or not it turns out that antibody is required, it is already abundantly clear that the classical pathway can be activated only by antibodies of the immunoglobulin-M and immunoglobulin-G classes, and that large quantities of the IgG antibodies are required.

The properdin pathway can be activated by aggregates of immunoglobulins that do not activate complement in the classical way. This has been demonstrated by Abraham G. Osler of the Public Health Research Institute of the City of New York and by Ann L. Sandberg of the National Institute of Dental Research. Although the full implications of these observations are not yet clear, the properdin pathway may represent a mechanism for the activation of immune defenses when sufficient quantities of specific antibody are not available for activation of the classical pathway.

The Donnan Effect

The general nature of complement attack on the cell has been understood for some time. As Burton D. Goldberg of the New York University School of Medicine and Howard Green of the Massachusetts Institute of Technology have shown, when cells are attacked by complement, they swell until the cell membrane is explosively ruptured and the contents of the cell spill out.

The cause of the swelling is a physicochemical phenomenon known as the Donnan effect. It occurs with semipermeable membranes whose pores pass common salts (such as sodium chloride) and water but do not pass large molecules (such as proteins). When such a membrane is set up with a solution of protein, salt and water on one side and a solution of salt and water on the other side, there is a flow of salt and water through the membrane toward the side with the protein.

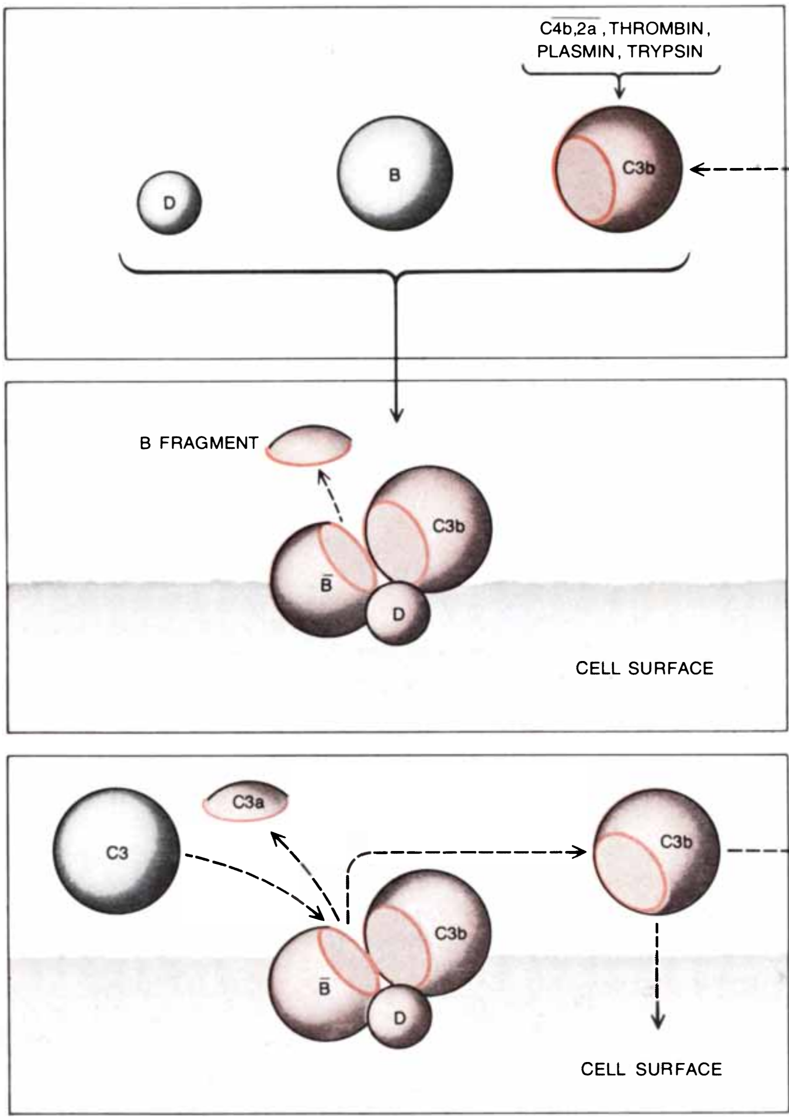
Since living cells contain protein, they would be subject to the Donnan effect if the cell membrane behaved like a semipermeable membrane. The fact is that the cell membrane does not behave in this way. It has transport mechanisms that actively move various substances through it. When a cell membrane is damaged by complement, however, it does behave like a semipermeable membrane. Then salt and water, and other small molecules, readily flow into it.

There are various ways in which complement might possibly damage the cell membrane. The simplest would be that the complement system makes holes in the membrane. If numerous holes are required for the destruction of the cell, the process would exhibit the characteristic pattern of a multi-hit reaction, and it would have a threshold. If a single hole is sufficient, the system would have the characteristics of a one-hit reaction, in which the number of cells destroyed would be directly proportional to the quantity of complement. In order to determine whether the reaction is multi-hit or one-hit, I undertook experiments to determine how the degree of destruction in a cell population varies with the quantity of complement. The results support the one-hit hypothesis [*see illustration on opposite page*]. I have also examined the question in terms of the kinetics of the reaction, a mode of analysis that involves the measurement of reaction velocity and that has played a major role in the efforts of our group. The results again seem to support the one-hit hypothesis. In this approach, however, the speed of the reaction may be so great that it is not possible to differentiate experimentally between a multi-hit and a one-hit reaction.

Since the complement system has multiple components, it is not possible to study whether the reaction is multi-hit or one-hit in terms of the overall process. Instead it is necessary to investigate the individual reaction steps. It has been found that for C1, C4 and C2 the attack is a one-hit process. It would be tempting to jump to the conclusion that this means that a single hole is sufficient for the destruction of a cell. Indeed, this was the view that was widely accepted until it was learned that several of the complement fragments are enzymes that can produce a shower of the next component. Such a process would result in a cluster of active sites. Particularly for the $\overline{C4b,2a,3b}$ enzyme we must consider the likelihood that it produces a cluster of C5b,6,7 complexes on the membrane surface. When these complexes reacted with C8 and C9, numerous holes would be produced, all of which could be attributed to a single $\overline{C4b,2a,3b}$ site. For this reason studies indicating that reactions involving C1, C4 or C5 are one-hit processes do not really tell us whether one hole is sufficient for the destruction of the cell.

The story is different for C9. Since it is the last component in the reaction series, it should be possible to determine if it is involved in a multi-hit or a one-hit reaction. We have found that adding C9 to cells carrying all the other complement components induces the destruction of cells in a manner indicative of a one-hit process. The efficiency of the C9 reaction is so high that fewer than five molecules, probably as few as one or two, are sufficient for the destruction of a single cell. These results indicate that even if there is a cluster of C5b,6,7,8 complexes, the reaction of only one of them with C9 may be sufficient to produce a hole capable of destroying the cell.

The holes made by complement can be seen in electron micrographs. In micrographs of the surface of a membrane with holes in it the lesion is seen as a light ring with a dark central portion [*see illustration on preceding page*]. The ring appears to be a raised surface and the dark center is thought to be a depression in the membrane. Alternatively, it has been suggested that the light ring is a hydrophobic (water-hating) region and the dark center is a hydrophilic (water-loving) one. John H. Humphrey of the National Institute for Medical Research in London and Robert R. Dourmashkin of the Clinical Research Centre at Harrow in England have studied the relation between the number of lesions



PROPERDIN PATHWAY is an alternate way of activating complement attack. The term properdin refers to a system of factors in blood serum that act together with complement in several immunological processes. The properdin pathway is activated by microbial cells or by bacteria. Little is known about the properdin enzymes, but three subunits—Factor B, Factor D and the complement fragment C3b—have been identified (*top panel*). Other subunits have been indicated in studies but have not yet been implicated definitively. Factor D activates Factor B by cleavage (*middle panel*) and the activated fragment is designated $\bar{B}$. The other B fragment goes into the fluid phase. C3b also plays a role, but its precise function is not known. Possible sources of the C3b that helps initiate the properdin system are the complement enzyme $\overline{C4b,2a}$, or plasmin, trypsin or thrombin in blood serum. Factor $\bar{B}$, Factor D and C3b become assembled on the surface of a microbial cell into a properdin system enzyme that corresponds in function to the complement enzyme $\overline{C4b,2a}$, that is, they split C3 into C3a and C3b (*bottom panel*). C5 also is cleaved (*not shown*), but there are indications that the properdin system enzyme that is involved differs slightly from the enzyme that cleaves C3, possibly with respect to an unidentified subunit. The C3b that is generated may bind to the microbial cell surface, where it promotes phagocytosis (the engulfing of the cell by white cells). Or the C3b may join and yield another properdin enzyme complex, thus setting up a positive feedback process. Reactions after cleavage of C3 and C5 may follow the same sequence as found in the classical pathway, but details are not known.

on the cell surface, as seen by electron microscopy, and the number predicted by the one-hit theory. In some instances a one-to-one ratio was found, which supports the one-hit hypothesis; in others the number of lesions exceeded what had been predicted. It may be that in the latter case clusters of lesions were formed as the result of enzymatic showers.

The Doughnut Hypothesis

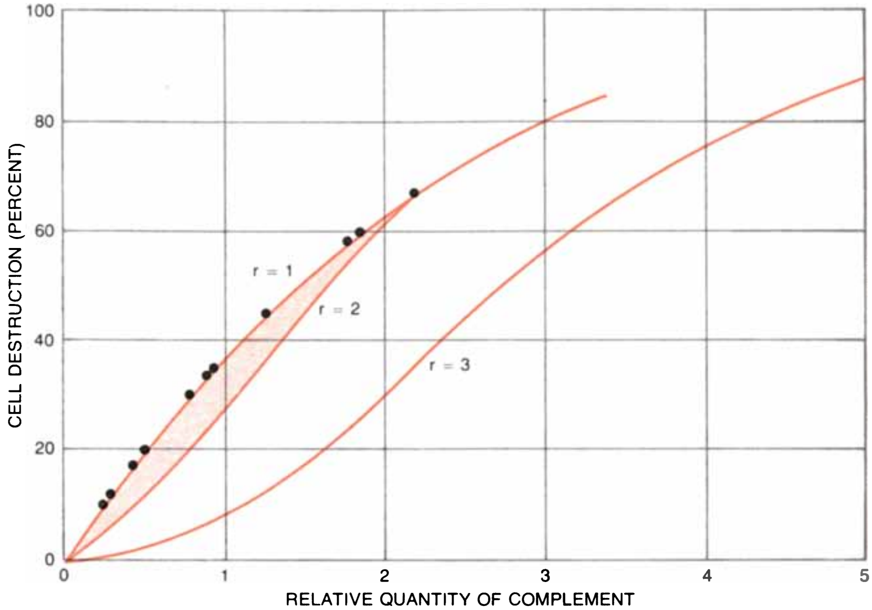
The fundamental structure of a cell membrane is a double layer of phospholipid, in which the hydrophilic heads of the phospholipid molecules point outward and the hydrophobic tails point inward. Embedded in the phospholipid bilayer are various proteins. Compounds soluble in lipid, such as sterols, may also float in the bilayer. A crucial element in this model of membrane is that the basic structure is a viscous fluid. Under these circumstances at least some of the embedded proteins could move laterally through the membrane.

Significant information about the mode of action of complement has come from studies of the artificial lipid bilayers called liposomes. They consist of a series of concentric lipid bilayers with alternating compartments for water. During their formation ions or small molecules may be trapped in the water compartments. When liposomes are attacked by complement, the trapped ions leak out.

Liposomes are not destroyed by complement in the same way as living cells. The Donnan effect does not come into play. Instead the lipid bilayer is somehow opened and ions from the water compartments flow out. The fact that complement can damage liposomes that consist entirely of lipid and glycolipid, that is, that are totally lacking in protein, strongly indicates that the attack of complement is directed against the lipid bilayer of the cell membrane and not against the proteins embedded in it.

How does complement attack the lipid bilayer? One hypothesis, which has been under consideration since the turn of the century, is that complement induces enzymatic action that gives rise to a leaky patch in the membrane. The crucial element in the leaky-patch hypothesis is that there is a small disrupted patch in the lipid bilayer. Since the bilayer is relatively fluid, the leaky patch must be regarded as a hole lacking rigid structure. Such patches would change in size over a period of time.

Electron micrographs show that lesions on cell membranes attacked by complement are quite uniform in size. Furthermore, lesions produced in red



ONE-HIT THEORY states that complement needs only to produce one lesion in the cell membrane in order to destroy the cell. The multi-hit theory states that two or more lesions are required. Theoretical curves were calculated from the binomial probability distribution for threshold values of r = 1, r = 2 and r = 3 (*colored curves*). Measurements were then made in the author's laboratory of the number of cells destroyed by complement reaction in which the amount of complement protein C1, C2 or C4 was varied. The results (*black dots*) fit the one-hit curve. The light colored region is the area in which the experimental measurements make it possible to discriminate between a one-hit model and a two-hit one.

blood cells by guinea-pig complement differ in size from those produced by human complement. The internal diameter of the lesions produced by the guinea-pig complement is between 8.5 and 9.5 nanometers; the internal diameter of the lesions produced by the human complement is between 10 and 11 nanometers. Therefore it appears that the complement from each species can produce lesions of a characteristic size.

The major weakness of the leaky-patch concept is that any lesions that were produced in this way would tend to disappear rather quickly. In my own studies of the stability of membrane lesions I found that they persisted for at least 30 minutes. Now, the leaky patch could be stabilized if it were surrounded by a wall of protein. This concept led me to formulate the doughnut hypothesis: a stable hole is produced by the assembly of a rigid, doughnut-shaped structure in the lipid bilayer of the cell membrane. The hole forms a channel connecting the inside of the cell with the extracellular fluid.

Such a structure could conceivably be assembled from proteins already in the membrane. In view of the fact that liposomes without protein are attacked by complement, however, such a mechanism seems improbable. A second candidate for the structure might be sterols in the lipid, but they too can be rejected because liposomes without sterols are also attacked. This leaves the complement proteins themselves as the source of the doughnut structure. I see no reason why several of the late-acting complement components could not be arranged to form a doughnut [*see illustration on page 60*]. The outside of the doughnut could be composed of nonpolar polypeptides, that is, protein chains that were hydrophobic; the interior would need polar peptides so that it could be hydrophilic. The structure of the doughnut would be similar to that of the antibiotic Valinomycin, the molecule of which has a hollow core. Another analogous structure is that of glutamine synthetase, which has been shown by electron microscopy to have a hollow core.

Soon after I had formulated the doughnut hypothesis Bernhard Cinader and his colleagues at the University of Toronto produced some electron microscope photographs of lesions that lend support to the concept. Cinader used a freeze-fracture technique, in which a frozen membrane is cleaved along the plane of its bilayer, exposing the membrane's interior. Sheep red blood cells treated with antibody and complement showed globular doughnut-shaped aggregates not found in similar cells treated only with antibody [*see illustration on page 61*]. The aggregates seem to be made up of several units. What is of cardinal importance, however, and what was first demonstrated in these micrographs, is that the complement lesions actually penetrate into the interior of the lipid bilayer.

Electron micrographs of lesions on liposomes, also published after the hypothesis had been formulated, show that the structure of the lesions is like a ring, and that a funnel-like lesion penetrates the lipid bilayer. These micrographs, made by Lachmann, D. G. Bowyer, P. Nicol, R. M. C. Dawson and E. A. Munn of the Royal Postgraduate Medical School in London and the Institute of Animal Physiology at Cambridge in England, show the lesions in top view and in side view [*see illustrations on page* 55]. Of particular interest is the fact that some of the rings are observed in a free state, that is, not on the liposome's surface. In the light of the procedures followed in the study, it appears likely that the rings were formed out of or derived from C5, C6, C7, C8 and C9. Direct proof of this interpretation calls for electron-microscopic identification of each of the complement components.

Other Activities

Among the activities of complement apart from cell attack are the release of histamine, chemotaxis and immune adherence. These activities are an important part of the inflammatory response, which plays a central role in immune and allergic reactions.

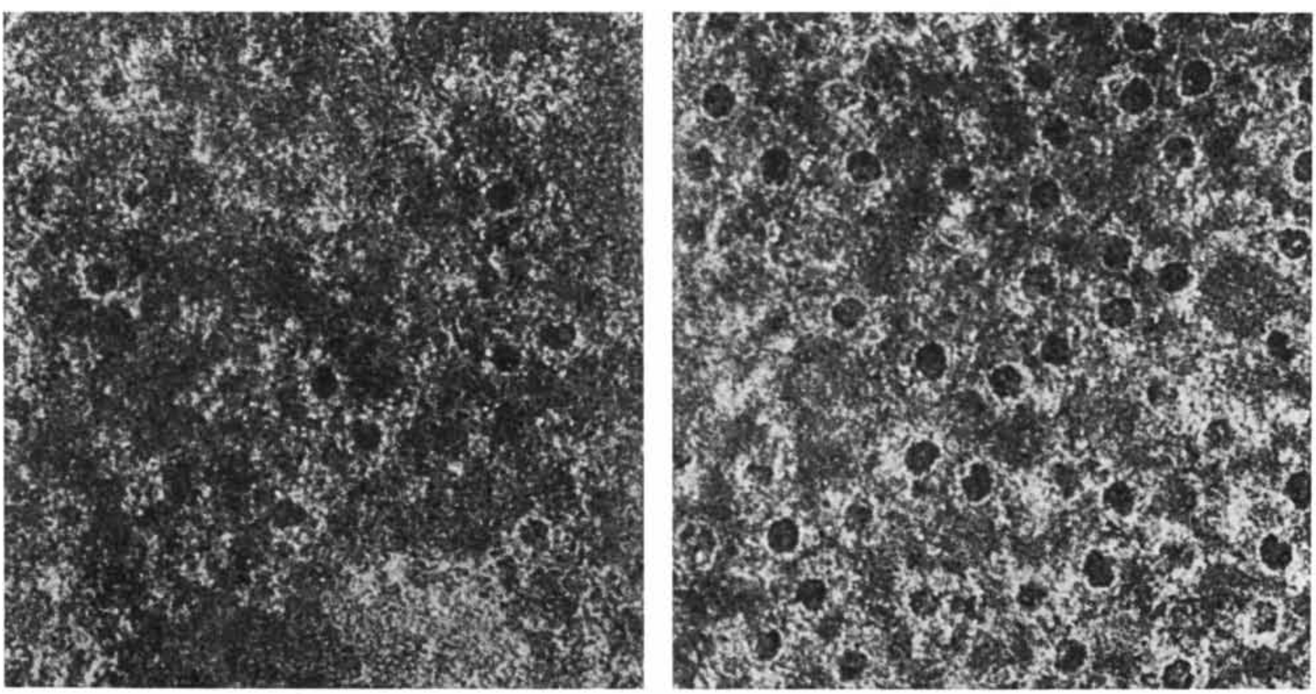
The C3a and C5a fragments cause the release of histamine from the cells that store this substance (leukocytes, mast cells and platelets). The histamine increases the permeability of the blood capillaries, which enables leukocytes (white blood cells) to penetrate into tissues where an infectious or allergic process is under way. Both C5a and C5b,6,7 complex are chemotactic for certain leukocytes, that is, they cause the leukocytes to migrate toward the site from which the chemotactic agents are diffusing. This phenomenon promotes the accumulation of leukocytes at tissue sites where immune reactions are proceeding.

The property of immune adherence is bestowed on cells by the C3b fragment. Cells carrying the C3b tend to adhere to leukocytes and other cells. Immune adherence thus promotes phagocytosis (the engulfing of cells by other cells). It is important to note that blood serum contains enzymes that degrade C3a, C3b and C5a, and that these inflammatory factors have a relatively short lifetime. The action of the degradative enzymes is still another device to limit the action of complement.

Many bacteria, notably some of those classified as Gram-negative, are killed by complement, and some are disintegrated in the process. Because bacteria have a cell wall outside their cell membrane, the bactericidal reaction is more complicated than the attack on other kinds of cells. Blood serum contains the enzyme lysozyme, which attacks the glycopeptide responsible for the rigidity of the bacterial cell wall. The attack dissolves the rigid layer of the wall. Tsunehisa Amano and Kozo Inoue of the University of Osaka have shown that lysozyme and complement act synergistically in the attack on bacterial cells. The bactericidal action of complement, however, is overshadowed by the phagocytosis of bacterial cells.

Phagocytosis is a prime mechanism of defense against bacterial infections. It has been known since the 1930's that antibody and complement render bacteria susceptible to phagocytosis, a process called opsonization. Pathogenic bacteria are often resistant to phagocytosis unless they are treated with antibody or complement.

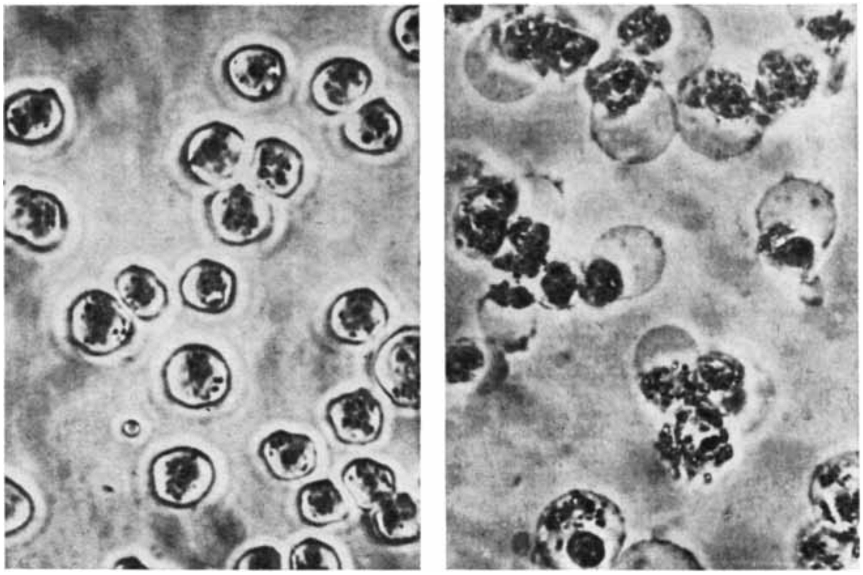
The manner in which complement mediates opsonization was not known until the 1950's, when it was shown by Nelson that the binding of C3b on cell surfaces produced immune adherence. In addition the movement of phagocytic cells toward the infectious agents is facilitated by the histamine-releasing ac-



HOLES PRODUCED BY COMPLEMENT appear to be of a uniform and characteristic size for each species. Sheep red-cell membrane treated with anti-Forssman antibody and complement from guinea pig develops holes with an internal diameter of between 8.5 and 9.5 nanometers (*micrograph at left*). Red-cell membrane from a human patient treated with anti-I antibody and human complement develops uniform holes with a diameter of between 10 and 11 nanometers (*right*). These electron micrographs, in which the enlargement is some 400,000 diameters, were provided by Robert R. Dourmashkin of the Clinical Research Centre at Harrow in England.

tivity of the C3a and C5a fragments. A directional element is provided by the chemotactic action of the C5a fragment and the C5b,6,7 complex.

The role of complement in phagocytosis is most prominent during the first week of a bacterial infection when an adequate antibody response has not yet developed. The late W. Barry Wood, Jr., Jerry A. Winkelstein and Shin of Johns Hopkins showed that during the preantibody phase the invading bacteria activate the properdin system, which in turn activates the complement system. Patients with certain defects related to C3 exhibit a pronounced susceptibility to pus-producing bacterial infections. Chester Alper and Fred Rosen of the Harvard Medical School have studied a number of patients with such a defect. Their investigations have contributed substantially to the understanding of the role of the properdin system and complement in the immune defense against bacterial infections.



TUMOR CELLS before and after treatment with antibody and complement are shown in these phase-contrast light micrographs provided by Burton D. Goldberg of the New York University School of Medicine. The micrograph at left shows untreated Krebs ascites tumor cells, magnified about 700 diameters. After treatment the cells swell and burst (*right*).

Anaphylaxis

The processes that protect the body against foreign cells or toxic substances can also produce undesirable effects, which are collectively termed allergy or hypersensitivity. One such phenomenon is anaphylaxis, an untoward reaction to foreign antigen, which may occur following repeated exposure to the antigenic substance. Anaphylaxis due to the sting of bees or wasps is not uncommon. The injection of certain drugs, particularly penicillin, can also cause anaphylaxis. Some people suffer anaphylactic reactions following the inhalation of allergens such as ragweed pollen or after the ingestion of certain foods. Anaphylaxis, which can be severe enough to cause death, is due to an antigen-antibody reaction that brings about the release of massive amounts of histamine and other substances.

It has been suspected for a long time that complement might be involved in anaphylaxis. This view is derived from the discovery more than 60 years ago that the treatment of blood serum with certain bacterial polysaccharides results in the formation of toxic substances, which were named anaphylatoxins. When serum containing such substances is injected into an experimental animal, the symptoms of anaphylaxis are produced. We now know that these bacterial polysaccharides are similar to the polysaccharides that activate the properdin system. Efforts to elucidate this phenomenon were stimulated by the work of Osler. They came to fruition with the discovery that the properdin system cleaved C3 and C5, and that the resulting C3a and C5a fragments induced the release of histamine.

The role of anaphylatoxins in anaphylactic reactions is quite uncertain. It is known that some reactions such as hay fever are produced by a different mechanism involving antibodies belonging to the IgE class of immunoglobulins. The production of local anaphylaxis with antibody from a different animal species has been extensively investigated, notably by Otto Bier of the Instituto Butantan in São Paulo, Brazil. Although the experiments with heterologous antibodies are not directly relevant to anaphylaxis induced by an animal's own antibody, they nevertheless demonstrate that activation of the complement system by immune complexes, with the consequent production of the complement subunits C3a and C5a, can produce local anaphylactic events.

In general, when immune complexes form in an organ or a tissue, the complement system will be activated at that site, through the classical pathway, the properdin pathway or both. The biologically active complement fragments and complexes can become involved in reactions that damage the host's cells, and these pathogenic reactions can result in the development of immune-complex diseases. For example, in some forms of nephritis complement damages the basal membrane of the kidney, resulting in the escape of protein from the blood into the urine. The disease disseminated lupus erythematosus also belongs in this category; its symptoms include nephritis, visceral lesions and skin eruptions. The treatment of diphtheria or tetanus with the injection of large amounts of antitoxin sometimes results in serum sickness, an immune-complex disease. Rheumatoid arthritis also involves immune complexes. Like disseminated lupus erythematosus, it is an autoimmune disease, in which the disease symptoms are caused by pathological effects of the immune system in the host's tissues.

How is complement involved in such diseases? As we have seen, the histamine released by the C3a and C5a fragments changes the permeability of the blood vessels near the site of complement activity. The chemotactic property of the C5a fragment and the C5a,6,7 complex, which has been studied extensively by Shin, Peter Ward and Ralph B. Snyderman, results in the migration of leukocytes into the area. Furthermore, the C3b fragment binds to immune complexes, and this action promotes the ingestion of the complexes by phagocytes. Phagocytosis, in turn, is followed by release of lysosomal enzymes that damage the surrounding cells and tissues. The cell-destructive capacity of the complement system may contribute to damage of cell membranes in the immediate vicinity. Activation of the complement system also accelerates blood clotting. This action comes about by way of the complement-mediated release of a clotting factor from platelets.

A genetic deficiency of complement

protein C2 has been found and investigated in some humans, but the most extensive studies of genetic complement deficiencies have been done with guinea pigs, mice and rabbits. Guinea pigs lacking C4 appear to be in good health, presumably because of protection provided through the properdin pathway. Michael Frank of the National Institute of Allergy and Infectious Diseases has shown, however, that under certain experimental conditions the complement-deficient guinea pigs exhibit abnormalities. For example, the ingestion of antibody-coated foreign cells by phagocytes does not always proceed properly. The antibody response to protein antigens is also defective.

Mice deficient in C5 also appear to be healthy, although in appropriately designed experiments their resistance to pneumococcal infection is shown to be impaired. The mice are also less resistant to a pathogen related to the bacillus that causes diphtheria in man.

The rejection of skin and organ transplants is caused by poorly understood mechanisms collectively termed cellular immunity. The participation of complement in the rejection process has been suggested by some investigations. For example, when the mechanism of cellular immunity is interfered with by the injection of an antilymphocyte serum, mice deficient in C5 reject skin grafts more slowly than their normal counterparts do. This result suggests that complement is involved in the rejection of grafts by the normal mice. Slower skin-graft rejection has also been noted by Klaus and Ursula Rother of the University of Heidelberg in studies of rabbits lacking C6.

Future Studies

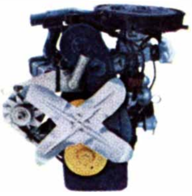
At the present time we are able only to sketch the general outlines of the complement system. Although some steps of the reactions have been worked out, much remains to be done, particularly with the late-acting complements. It should also be stressed that genuinely effective studies of the biological activities of complement (other than its cell-destroying action) have only begun, and that the information now available should be regarded as no more than the result of initial surveys. A fuller elucidation of the role of complement in the body's immune defense system will undoubtedly open the way to new and better means of controlling infections, allergic disorders and autoimmune diseases.

Our basic wagon idea: little cash and lots of carry.

Like the original station wagon Ford pioneered back in 1929, the 1974 Ford Pinto Wagon combines a durable, economical car with a lot of space in back. And what could be more basic than that?

The economy: First and foremost, the Pinto Wagon is an economy car—low in price, easy on gas, low on upkeep.

The space: Put a Pinto Wagon's rear seat down, and you've got over 60 cubic feet of cargo space. To maximize cargo space, the spare tire is carried under the floor.



Other basics: A rugged 2,000cc overhead cam engine is standard. It's been developed for good gas mileage. And for those of you who want even a bit more pep, there's an optional 2,300cc engine.

Front disc brakes are also standard—for efficient and fade-resistant braking and little pedal effort.

Further, the steering is rack-and-pinion, like in sports cars. Who'd have thought a station wagon could handle that easy? They're all good reasons why the closer you look, the better we look.

See the 1974 Pinto Wagon at your Ford Dealer.

When you get back to basics, you get back to Ford.

FORD PINTO

FORD DIVISION Ford

(Shown here: Pinto Wagon with Squire Option, optional whitewall tires, luggage rack, and deluxe bumper group.).

There's nothing like a
But there's a lot more to a

A

B

With a Polaroid portrait identification card, a glance tells you whether the person and the picture are the same. But an ID card should do more than match faces; it should suit the conditions of use.

For that reason we've developed all sorts of ways to make our ID cards more convenient for both the user and the issuer.

In addition to our standard cards (A and B)—carried in pocket, purse, or wallet—we have others with more sophisticated functions.

For instance, there's a card (C) which can be punch-coded. A computer reads it just as accurately as you match a picture to a face.

We have badges (D) with a clip for attachment to outerwear. They have regular or extra-large portrait areas.

We have a card (E) which can be encoded to act like a key. Slipped into an electronic device, it allows access to secure areas, parking lots, etc.

We have a new credit card (F) combining a magnetic stripe with a portrait and an embossed name and number. It's another way of encoding information securely for automatic banking transactions.

Chances are we have an ID card that will exactly fill your needs.

Security starts with the card itself.

The Polaroid ID card is highly secure. It's laminated with a clear, tough plastic which becomes integral with the card.

The card works so well that more than 6000 organizations all over the world use it. You'll find our ID cards in industry, universities, research and computer centers. They're used as credit cards and drivers' licenses.

People like to carry them. They like having their own color portrait, and they like being able tc

Polaroid®

›rtrait for an ID card.
ID card than a portrait.

ove who they are.

They carry them with confidence.

Credit cards, for example. Before the portrait, ıny people were afraid to carry them around. ıt now, with the portrait, a lost or stolen credit rd is no longer a serious problem.

You can't forge a face.

A handy system for making cards.

Any organization can produce its own ID cards sily and quickly with a low-cost Polaroid ID-3 ınd Identification System.

It can be carried around like a suitcase, so you n take it from one branch of your organization another. An operator can be trained to run it in ew minutes.

With an ID-3, you can immediately issue per-anent ID cards to all your employees. If any-›dy blinks, you can make a new card on the spot.

You can use different background colors to color-code the employee's function, work area or security clearance.

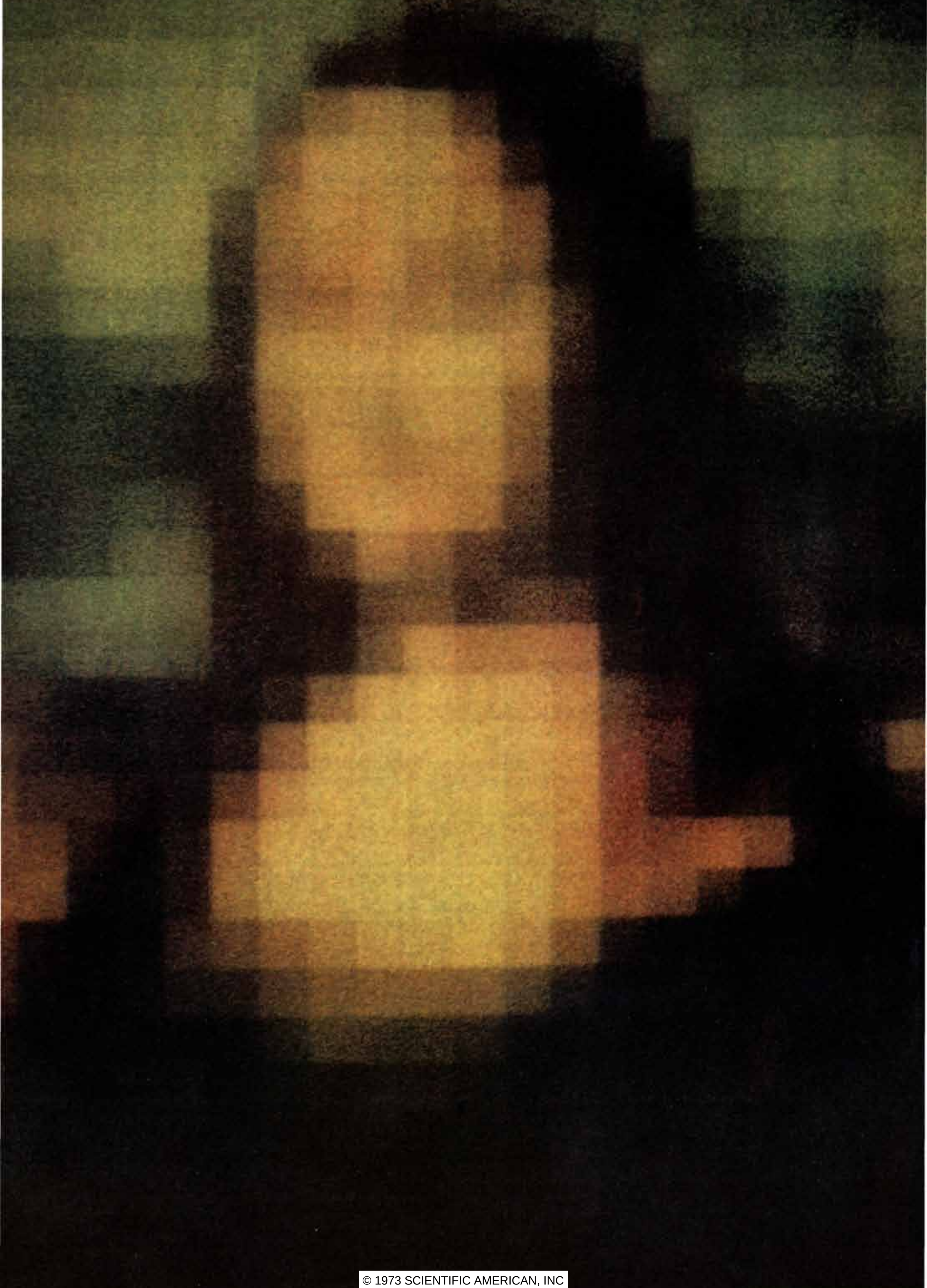
Security starts with the ID-3. It's the basis of entire security systems.

We'd like to demonstrate it in your own office or factory.

We'll put your portrait on a card and let it speak for itself.

For more information or a demonstration, write to Polaroid Corporation, Dept. 57-255, 549 Technology Square, Cambridge, Mass. 02139. Or you can call collect (617) 864-6545.

The Polaroid 2-Minute ID System.



THE RECOGNITION OF FACES

One of the subtler tasks of perception can be investigated experimentally by asking how much information is required for recognition and what information is the most important

by Leon D. Harmon

Faces, like fingerprints and snowflakes, come in virtually infinite variety. There is little chance of encountering two so similar they cannot be distinguished, even on casual inspection. Unlike fingerprints and snowflakes, however, faces can be recognized as well as discriminated. It is possible not only to tell one from another but also to pick one from a large population and absolutely identify it, to perceive it as something previously known, just as in reading one not only can tell that an *A* is different from a *B* but also can identify and name each letter.

Why are faces so readily recognized? In seeking the answer to this question my colleagues and I posed several related but more modest questions that we believed would be more amenable to experimental investigation: How can a face be formally described? Given a verbal description, how well can a particular face be identified? To what extent is recognition impaired when the image of a face is blurred or otherwise degraded? What kinds of image degradation most seriously affect recognition? Can faces be classified and sorted as numerical data?

This inquiry was inspired by yet another question: How can a computer be made to recognize a human face? This question remains unanswered, because pattern recognition by computer is still too crude to achieve automatic identification of objects as complex as faces. Machines can recognize print and script, craters and clouds, fingerprints and pieces of jigsaw puzzles; the recognition of human faces, however, is a much subtler task.

Even though machine recognition of faces has not been attained, the investigation of how it might be done has led to a number of related issues that in themselves are worthwhile (and tractable) areas of research. Several new approaches to problems in the manipulation of visual data have emerged. I shall recount here four series of experiments that were directed to an understanding of recognition. The first is concerned with how artists reconstruct faces from descriptions and how closely the resulting portraits resemble the person described. Next I shall comment on a set of experiments in which faces were identified from pictures that had limited information content. The third approach examines the recognition of faces from formal numerical descriptions. Finally, I shall describe a system in which man and computer interact to identify faces more efficiently than either could alone.

If one could devise an objective formulation of the criteria used by an artist in drawing a portrait, a set of properties useful for automatic recognition might emerge. One kind of art that we thought might provide useful information is the sketches drawn by police artists (called face-reconstruction artists) from descriptions provided by witnesses. (Another promising possibility is the caricature, but we have not yet studied it.)

Verbal descriptions are rarely used in the drawing of police sketches. Few observers, unless they are specially trained, can give satisfactory clues to appearance in words. Most can point to features similar to those they remember, however, and that is how the reconstruction artist usually begins. Our initial experiments were intended to test the effectiveness of this procedure and to gain some preliminary notions of what features are considered important in describing or recognizing a face.

Frontal-view photographs were shown to an experienced artist, who compiled a written description of each face; the description included references to facial features in a catalogue of faces made up of photographs of various head shapes, eye spacings, lip thicknesses and so on, organized by feature type. Thus a large part of the description consisted of "pointing to" similar features on other portraits. The completed description was given to another artist, whose task was to reconstruct the face from the written description [*see illustration on next page*].

The first attempt, although obviously resembling the original photograph, differed from it in the depiction of important features and proportions. When limited feedback was allowed, however, there was rapid improvement. The describing artist, with the initial sketch in hand, provided simple verbal corrections, such as "The hair should be bushier at the temples"; with this information the reconstructing artist was able to draw a much more accurate likeness. Finally, to find the limit of improvement, that is, to discover just how faithful a portrait could be drawn, the reconstructing artist was given the photograph to work from. Under those conditions he was able to produce a strikingly realistic representation. Some sketches, in fact, were judged to look more like the per-

LEONARDO'S "MONA LISA," rendered as a "block portrait," consists of 560 squares, each of which is uniform in color and brightness. The transformation of the familiar painting was accomplished in the same way as that of the portrait of George Washington on the cover of this issue of SCIENTIFIC AMERICAN. Recognition can be enhanced by rapidly moving the page, by squinting at the image or by viewing it from a distance of 10 feet or more.

son than the photograph did. Presumably the artist enhanced recognition by in some way emphasizing significant detail.

All the sketches were shown to test subjects who, as fellow employees, had seen the "suspect" often. Almost half of the sketches drawn from descriptions were correctly identified and about 93 percent of the drawings made directly from photographs were recognized.

Our work with face-reconstruction artists was a pilot experiment we hoped would lead, through informal observation, to a better understanding of the problems confronted in the recognition of faces and to the formulation of further experiments. Some of the incidental information derived from the study was indeed interesting. For example, we found that several of the faces were outstandingly easy to recognize in the sketches. Presumably those subjects were more easily described than the others, or perhaps they possessed certain features that are conspicuous or rare. Several subjects remarked that the nose and eyes in one sketch were important to identification, yet for the same face other subjects observed that although the nose, mouth and hair were well drawn, the eyes were not and did not aid recognition.

Another way to study recognition is to ask how little information, in the informal sense of "bits," or binary digits, is required to pictorially represent a face so that it can be recognized out of a finite ensemble of faces. We explored this "threshold" of recognition with portraits that had been precisely blurred.

The type of blurring commonly encountered in photographs is caused by an improperly focused optical system; it reduces the information content of the picture, but it proved unsuitable as a technique in our investigations because the degree of blurring cannot be precisely specified or controlled. A more measurable method degrades the image in quantifiable steps through a relatively simple computer process.

In our experiments a 35-millimeter transparency of a conventional portrait photograph is scanned by a beam of light moving in a raster pattern of 1,024 lines. The variations in the intensity of the beam caused by the varying transparency of the film are detected by a photomultiplier tube. The analogue signals produced by the photomultiplier are converted into digital form by sampling each line in the raster at 1,024 points and assigning a brightness value to each point, so that the completed image consists of $1{,}024^2$ (or 2^{20}) discrete points, about four times the resolution of the commercial television image. Each of the points may have 1,024 brightness values, or tones of gray. The dissected image is stored in the magnetic-tape memory of a digital computer.

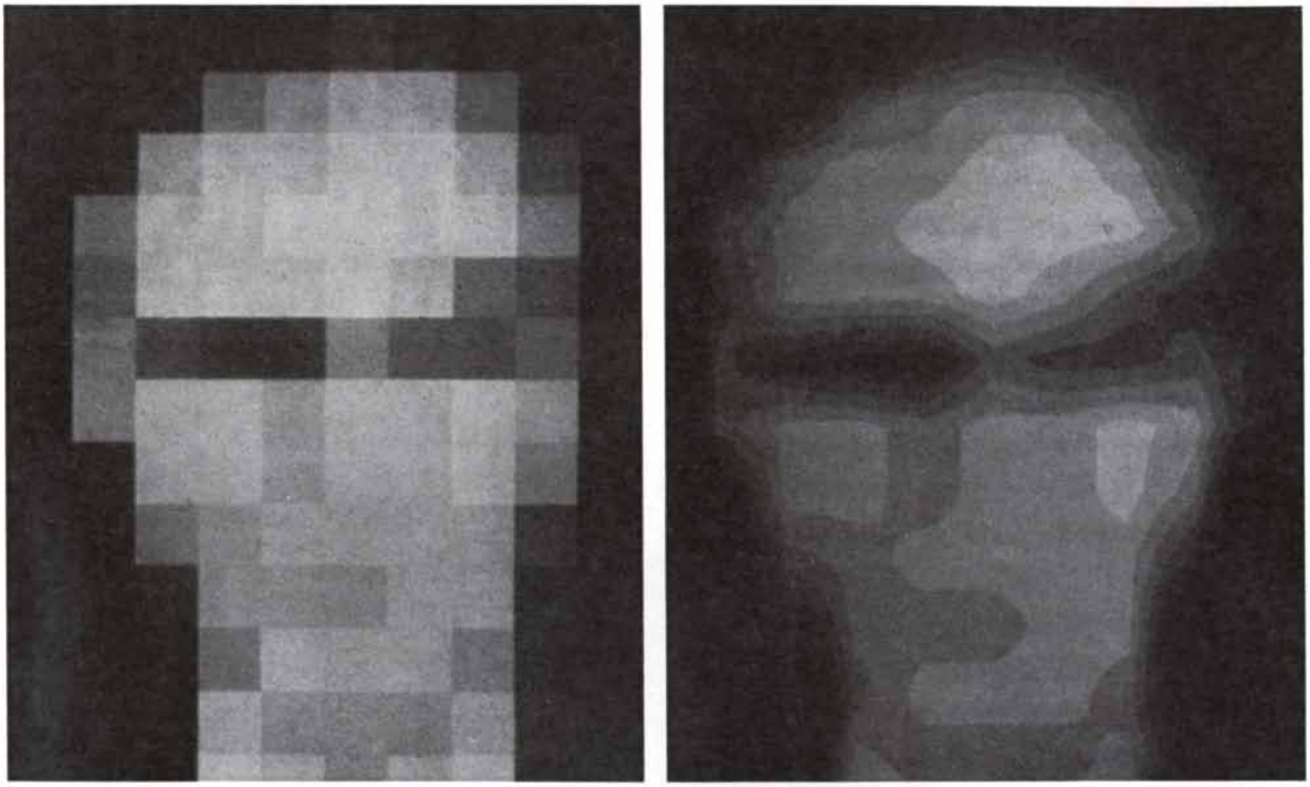
To create the degraded image the computer divides the picture into $n \times n$ squares of uniform size and averages the brightness values of all the points within each square. For example, if a photograph is to be made into an array of 16×16 squares, each square will contain 64×64, or 4,096, points; the brightness to be assigned to the entire square will be found by averaging the values of these points. In a final step the number of brightness values is reduced to eight or 16 by assigning to each square the gray tone closest to its original averaged value.

The computer stores the digital information comprising the picture on magnetic tape and the tape controls a cathode-ray-tube monitor, which then displays the completed portrait. A photograph of this display constitutes the finished product. Alternatively, the magnetic tape can be used to control a facsimile printer that produces a print of the processed image without the intermediary cathode ray tube [*see bottom illustration on opposite page*].

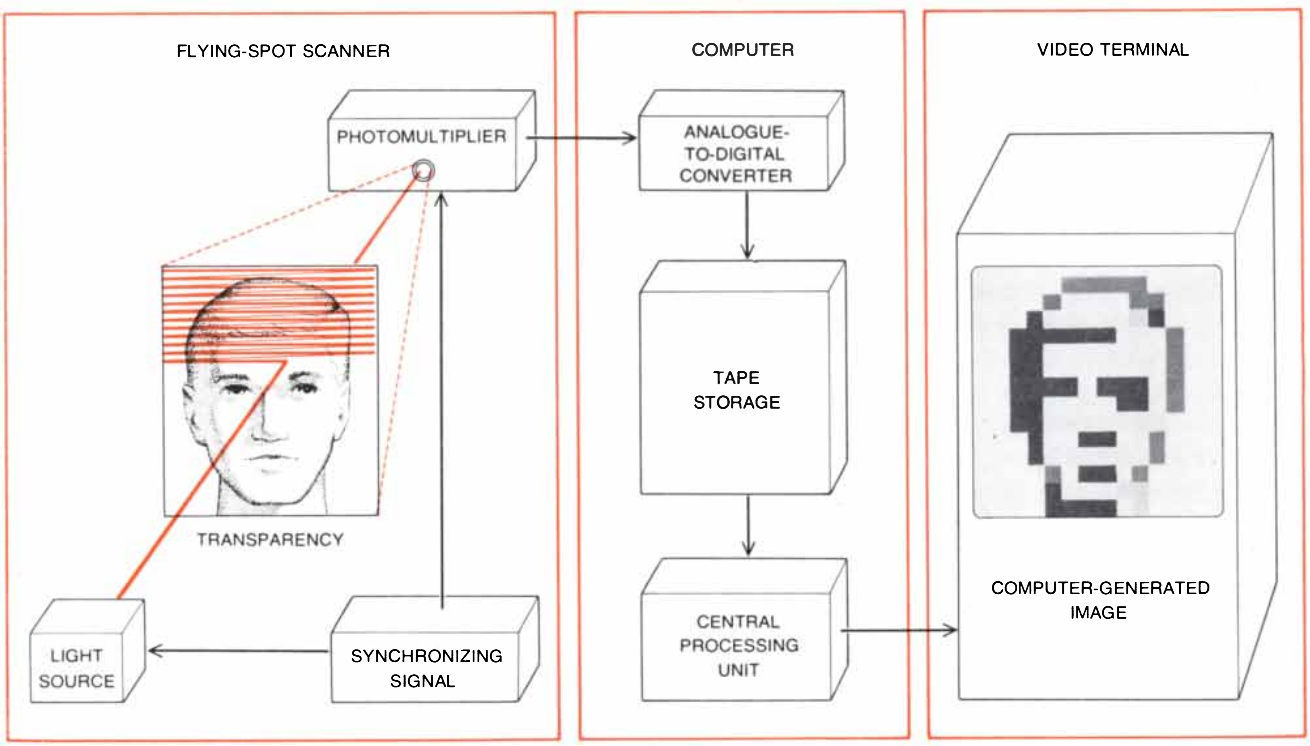
Viewed from close up, these "block portraits" appear to be merely an assemblage of squares. Viewed remotely, from a distance of 30 to 40 picture diameters, faces are perceived and recognized.

Preliminary experiments were made to select the coarsest image that might be expected to yield about 50 percent accuracy of recognition. For some kinds of picture, resolution of only a few thou-

SKETCHES FROM DESCRIPTIONS were made by a "face-reconstruction artist" skilled in drawing portraits from information provided by witnesses. At top left is the photograph from which the three sketches are derived. For the first drawing (*top right*) a written description of the face, including references to illustrations in a catalogue of facial features, was presented to the artist. A better likeness was produced (*bottom left*) when simple verbal corrections were provided. For the final version (*bottom right*) the artist was given the photograph; the resulting portrait represents the limit of accuracy of the process.



REDUCED-INFORMATION-CONTENT PORTRAITS were generated by a computer. The picture at left is a block portrait; it is an array of 16×16 squares, each one of which can assume any one of 16 levels of gray. Not all the 256 squares are required to represent the face. The contoured representation at right was produced by filtering the block portrait to remove high frequencies.



SYSTEM FOR MAKING BLOCK PORTRAITS uses a flying-spot scanner, a device similar to a television camera. The image, usually in the form of a 35-millimeter photographic transparency, is scanned in a raster pattern of 1,024 lines. In the analogue-to-digital converter each line is sampled at 1,024 points and the brightness of each point is assigned one of 1,024 values. Using this information stored on magnetic tape, the central processing unit divides the image into $n \times n$ squares and averages the brightness values of all the points within each square. The number of permissible brightness values is then reduced to eight or 16. The resulting image is displayed on a video terminal (a television screen) and photographed. The computer can also be made to operate a facsimile printer, which produces a finished picture directly. Most of the portraits used in these experiments were made by the latter process.

sand elements provides acceptable quality; the limits of recognition for photographs of faces, however, have not been reported. Our informal investigation revealed that a spatial resolution of 16 × 16 squares was very close to the minimum resolution that allows identification.

Tests were also made to determine the useful limits of gray-scale representation. The relation between gray-scale and spatial resolution is an interesting one: either factor can serve as a limit to recognition. It was not the object of our experiments to document this relation, however, and so only a few gray-scale tests were made once the 16 × 16 spatial pattern was decided on. For 16 × 16 portraits gray scales of either eight or 16 levels yielded eminently recognizable portraits; consequently our experiments used those levels exclusively. (The allowed gray levels can be expressed in terms of bits. A gray scale of eight levels requires three bits of information; a scale of 16 levels calls for four bits.)

Fourteen of the block portraits were shown to 28 subjects. Each subject was given a list of 28 names, including the names of the 14 persons depicted. The experiment was intended to investigate the effects of changing the gray scale from a three-bit to a four-bit one, as well as to test identification performance.

Overall recognition accuracy was found to be 48 percent. (Random guessing would produce such a result only four times in a million trials.) The result was essentially indifferent to the resolution of the gray scale. Thus the number of bits required for approximately 50 percent accuracy of recognition was no more than 16 × 16 squares times three bits, or 768 bits. None of the portraits, however, filled all the squares in the 16 × 16 grid; therefore fewer than 256 squares made up each face. An average of 108 squares was needed.

Recognition of particular faces ranged from 10 percent to 96 percent. In these experiments too some faces were always easy to identify, although, as will be seen, the reasons are peculiar to the conditions of the experiment. Two portraits received outstanding recognition and four were rarely identified correctly.

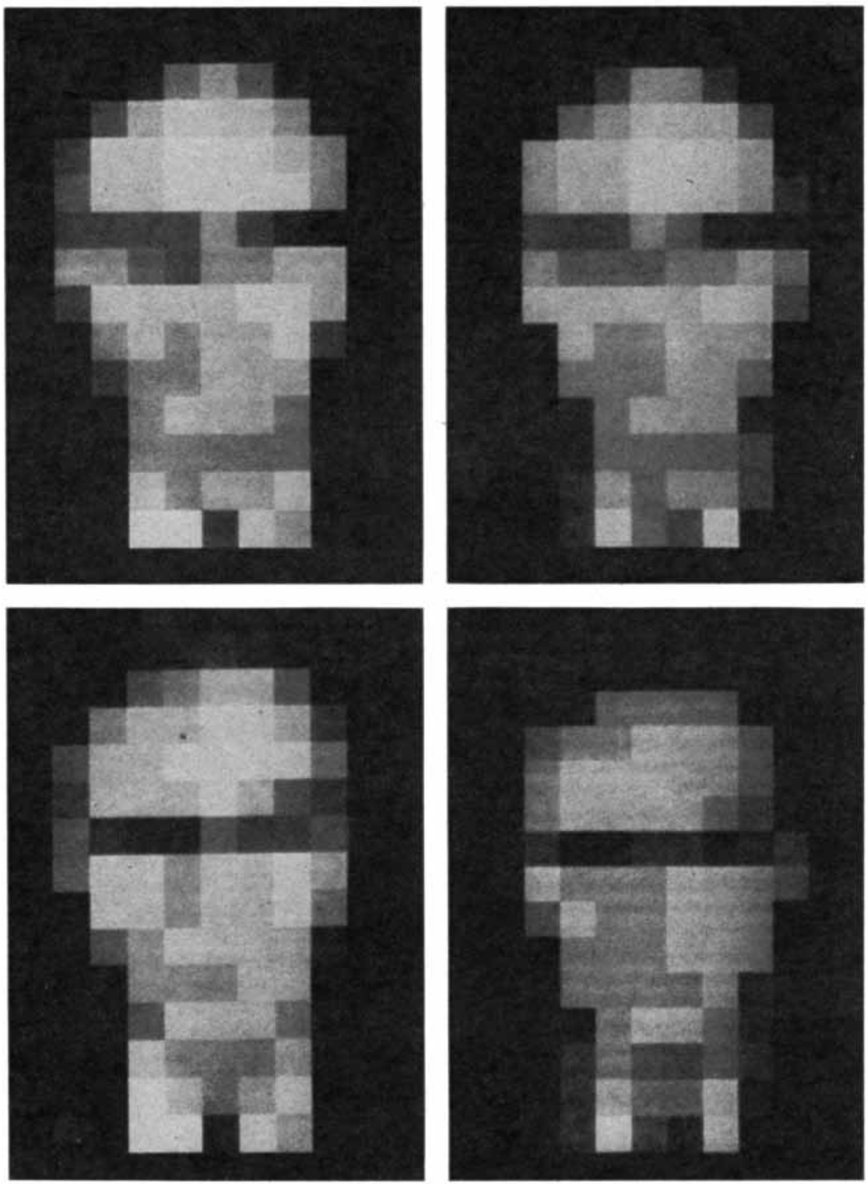
Two possible explanations of these disparities were suggested. First, some faces, because of the peculiar arrangement of their features, respond notably well or particularly poorly to coarse spatial presentation. Second, the grid, arbitrarily positioned over a given face by the scanning process, may land luckily or unluckily for adequate representation. For example, a square might just bracket an eye, or it might land half on and half off. The latter possibility was judged to be the more likely. I hypothesized that those pictures that were recognized well probably had a fortuitously placed grid.

To test the hypothesis each portrait was reprocessed by shifting the 16 × 16 matrix with respect to the original block portrait. Three new pictures were made: one shifted a half-square to the right, one a half-square down and a third a half-square to the right and down [*see illustration at left*].

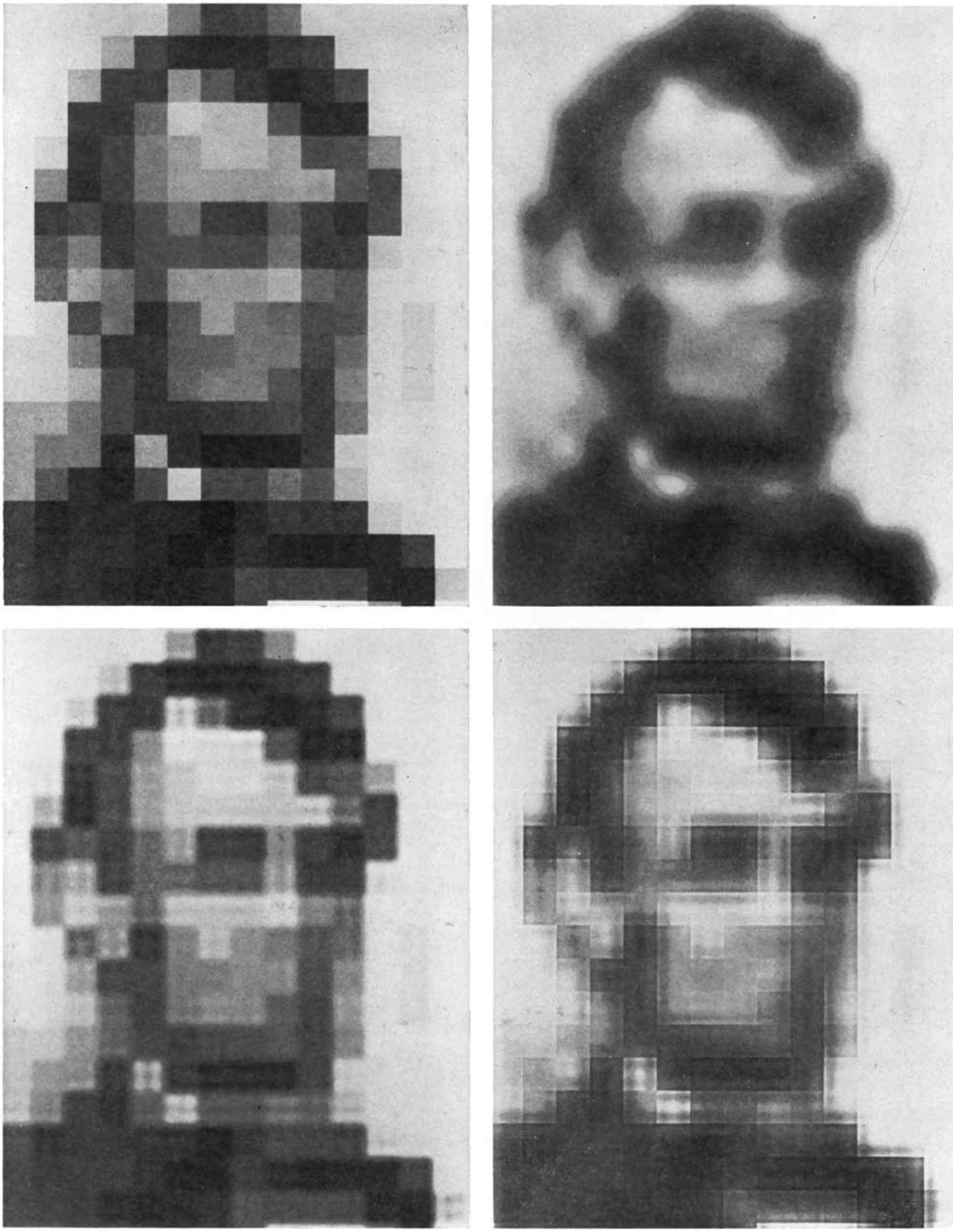
Recognition of the sets of four shifted pictures was tested. The subjects were given the identity of each photograph; their task was to rank the four portraits in each set in order of pictorial accuracy. My hypothesis predicted that in these tests those pictures that were readily identified in the earlier experiment would be ranked first in their set and that those scoring worst initially would be ranked near the bottom. So it turned out; both correlations were confirmed.

This result led us to believe that if the best grid positions had been found and used in the earlier experiments, the average accuracy of recognition might have been closer to 100 percent than to 50 percent. A new experiment confirmed this: performance rose to 95 percent.

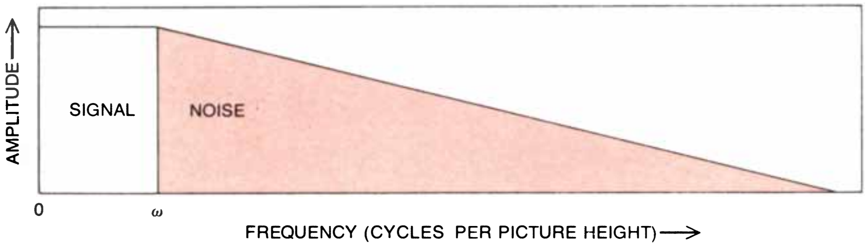
An interesting and provocative characteristic of block portraits is that once recognition is achieved more apparent



EFFECTS OF GRID PLACEMENT on recognition are illustrated by four block portraits of the same face. The original is at top left. Alternative versions were made by shifting the grid placement one half-block to the right (*top right*), one half-block down (*bottom left*) and one half-block right and down (*bottom right*). When portraits made with optimum placement replaced those made with random placement, recognition accuracy doubled.



SELECTIVE FREQUENCY FILTERING influences the ease with which block portraits are recognized. The original block portrait of Abraham Lincoln is at top left. It consists of the photographic "signal," whose highest spatial frequency is 10 cycles per picture height, and noise frequencies extending above 10 cycles. As was anticipated, filtering out all spatial frequencies above 10 cycles (*top right*) greatly enhances recognition. Selective removal of only part of the noise spectrum, however, reveals which frequencies most effectively mask the image. At bottom left all frequencies above 40 cycles have been removed; even though the sharp edges of the squares are eliminated, perception is improved only slightly. When the two-octave band from 10 to 40 cycles is removed (*bottom right*), the face is more readily recognized. The phenomenon apparently responsible for this effect is critical-band masking.



BLOCK-PORTRAIT SPECTRUM consists of a signal that extends to some finite spatial frequency ω, corresponding to the block-sampling frequency and noise of frequencies above ω. The amplitude of the noise typically decreases with increasing spatial frequency.

detail is noticed. It is as though the mind's eye superposes additional detail on the coarse optical image. Moreover, once a face is perceived it becomes difficult not to see it, as if some kind of perceptual hysteresis prevented the image from once again dissolving into an abstract pattern of squares. The observation that is most intriguing, however, is that recognition can be enhanced by viewing the picture from a distance, by squinting at it, by jiggling it or by moving the head while looking at it. The effect of all these actions is to blur the already degraded image.

Why should recognition be improved by blurring? The explanation almost certainly lies in the "noise" that tends to obscure the image.

A picture, like a sound, can be described as the sum of simple component frequencies. In acoustical signals pressure varies with time; in the optical signals discussed here the frequencies are spatial and consist of variations of "density" (or darkness) with distance. Just as a musical note consists of a fundamental frequency and its harmonics, so an optical image consists of combinations of single frequencies, which make up its spatial spectrum. The spectral representation exists in two dimensions. This spectrum refers only to spatial frequencies; the color spectrum describes another aspect of the image.

When pictures are considered combinations of spatial frequencies, they can be manipulated in the same ways as other frequency-dependent signals are. For example, Fourier analysis can be used to determine the component frequencies of an image, or low-pass filtering can be used to remove the high frequencies that represent fine detail. Signal-frequency bands, the signal and noise spectrum and other terms usually associated with discussions of acoustical phenomena can be applied to the processing of visual images.

The description of a two-dimensional image as a signal of various spatial frequencies leads to a possible explanation of the enhancement of block portraits with blurring. Whenever a signal with a spectrum running from zero to some frequency designated ω is reduced by sampling to discrete frequency components, noise artifacts whose spectrum extends above ω are introduced. The noise is a product of the sampling procedure. In two-dimensional signals it appears as patterns not present in the original image.

Because the noise in these pictures is ordinarily of higher frequency than the signal it can be readily eliminated by a low-pass filter, that is, a filter that preserves only the low frequencies, eliminating the high frequencies that represent fine detail. This operation too is performed by the computer; all spectral components above ω are removed while the desired signal is retained [*see top illustration on this page*].

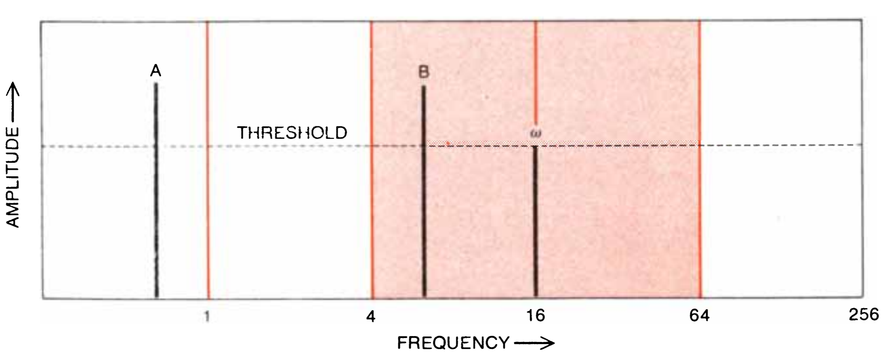
In block portraits the most obvious noise is that introduced by the sharp edges of the squares. Although Fourier analysis shows that the energy content of these high frequencies is relatively small, one might speculate that because the eye is particularly sensitive to straight lines and regular geometric shapes such square-patterned noise masks particularly well. That is, such image-correlated noise might mask more effectively than randomly distributed noise of equal energy. If so, low-pass filtering should enhance perception. This explanation would seem to be confirmed by the fact that recognition is improved by progressive defocusing or distant viewing, since the effect of both of these actions is to filter out high frequencies.

This hypothesis, however, is not the only candidate; another possibility is called critical-band masking. In both hearing and vision the spectral proximity of noise to a signal drastically influences the detection threshold of the signal. For example, the threshold for detecting a single sinusoidal wave anywhere in the spectrum is elevated when a noise signal is introduced if the noise lies within about two octaves of the signal. If the noise lies outside this "critical band," masking does not occur [*see bottom illustration on this page*].

This phenomenon has been tested and confirmed by others for relatively simple visual presentations such as sine-wave and square-wave gratings in a single dimension. My colleague Bela Julesz and I reasoned that similar masking might occur in more complicated two-dimensional patterns.

If critical-band masking is the mechanism that hinders the recognition of block portraits, then those components of the noise that fall within about two octaves of the sampling frequency ω would be primarily responsible. The rest of the noise spectrum, including the high-frequency signals contributing to the sharp edges of the blocks, should cause little or no masking.

To resolve this question we prepared a series of block portraits that were spectrally manipulated by the computer. The original image was transformed to obtain its Fourier spectrum, filtered to specification, then transformed back and



CRITICAL-BAND MASKING is known to occur in presentations of simple visual or audio signals, such as single sinusoidal waves. The test signal ω, at the threshold of perception, would be masked by signal *B*, within the band, but not by signal *A*. The critical band (*colored area*) extends for about two octaves above and below the test frequency. The author's investigations indicate that critical-band masking also affects two-dimensional signals.

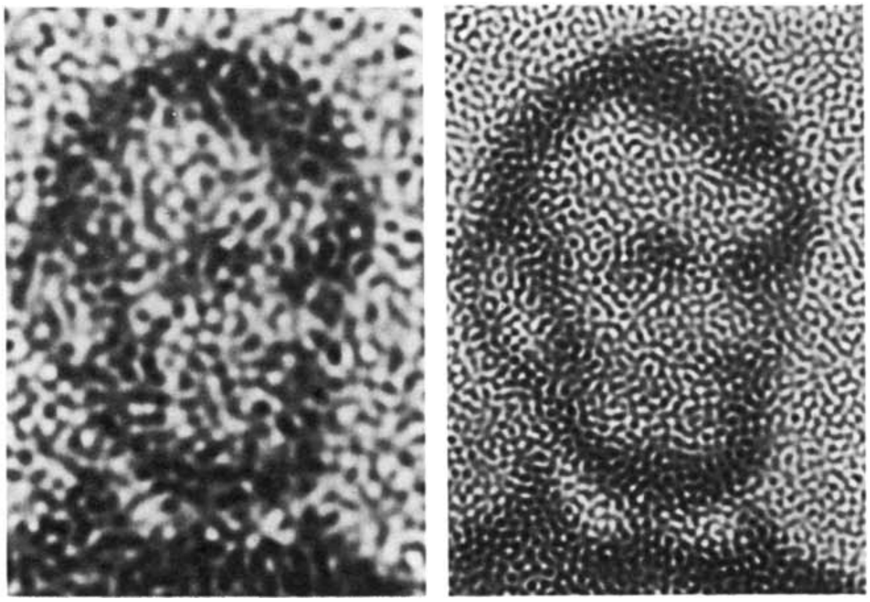
printed out. This technique provides precise control of spatial frequencies. We were able to remove all signals above a specified frequency, or to remove only a band of frequencies adjacent to ω.

In our first attempt to evaluate the relative importance of high-frequency and critical-band noise masking we prepared a series of filtered block portraits [*see illustration on page 75*]. The result looked promising: removal of the very high frequencies did little to change the block aspect, and some effort was still required to perceive the face. Removal of only the frequencies adjacent to the signal produced pictures that were much closer in appearance to the original photographs. That is, the very high frequencies did not seem to be the most important in masking.

Although the results of this experiment suggest that noise spectrally adjacent to the signal is most effective in masking recognition, the point is not proved. There are three reasons why the experiment is not conclusive. First, the noise generated by the block-sampling process is spatially periodic, at the block frequency and at higher harmonics. Second, the noise amplitudes are correlated with picture information: the magnitude of the noise in any block depends on the density of the image in that block. Finally, the energy of the noise spectrum is greatest at the block-sampling frequency, and it decreases with increasing frequency. Hence the adjacent band noise may mask more effectively simply because its amplitude is higher, not because of the critical-band effect.

There is a straightforward way to avoid these difficulties. We can simply add random noise of the proper frequency to a picture that is smoothly blurred rather than block-sampled. We added random noise of constant spectral energy to a portrait that had been low-pass filtered to the same bandwidth as that used in the first recognition studies. When such a picture containing adjacent-frequency noise is compared with one masked by remote-frequency noise, the result is unequivocal: critical-band masking is responsible for the suppression of recognition [*see illustration on this page*].

The discovery that critical-band masking affects complex pictures as well as simple sinusoidal presentations raises additional questions. How effective in masking are noises of equal energy and bandwidth but of various spectral shapes? When noise is added to a signal, is the shape of the noise signal or its location in the spectrum more important? What are the relative effects of spatial disposition and spectral disposition? That is, if equal amounts of noise energy are added to visual scenes, is the placement with respect to position or with respect to frequency more important for masking? These and related questions remain for future investigation. Their answers will provide new insights into the psychophysics of vision.



RANDOMLY DISTRIBUTED NOISE of uniform amplitude is added to smoothly blurred portraits of Lincoln. When the noise is in the band adjacent to the signal frequencies (*left*), it obscures the picture more effectively than when it is at least two octaves removed from the picture frequencies (*right*), confirming that critical-band masking is the most important mechanism limiting the recognition of degraded or blurred images such as block portraits.

A more conventional means of blurring pictures is continuous smearing. In optical systems one can simply project the image out of focus; as I have noted, however, this operation cannot be precisely controlled. The analogous operation performed by a digital computer is intrinsically discrete, but by using sufficiently numerous sample points to represent an image, blurring can be made arbitrarily smooth and extremely precise.

Pictures made up of 256×256 elements (about one-fourth the resolution of television) produce fairly sharp portraits. Such pictures can be blurred by selecting for each point a brightness value computed by averaging the brightness of the points surrounding it. An "averaging window" of $n \times n$ points is used to compute a new value for each point in the 256×256 array; after a new point is written the window is moved over one element and a new average is made.

Through this process the computer can rapidly and accurately blur a picture to a specified degree. The size and shape of the averaging window and the relative weight given to each element in the array can be selected at will. For example, the average could be uniformly weighted or computed on a Gaussian, or bell-shaped, curve. In our experiments we used a square window of varying size with uniform weighting; each element contributed equally to the average value assigned the new point.

We have used portraits made in this way to study the limits of face recognition. Fourteen portraits were shown to subjects who were given a list of 28 names, including the names of the 14 "target" individuals. All 28 persons were known to the test subjects. Several degrees of blurring were tested [*see illustration on next page*]. Some subjects were shown the most blurred pictures first, some the least blurred, and so on, in order to simultaneously test for the effects of learning. The experiments were conducted by Ann B. Lesk, John Levinson and me.

Contrary to what we had expected, recognition scores were quite good. For photographs blurred by a 27×27-point averaging window the recognition was 84 percent. As the degree of blurring increased, the scores declined to about 65 percent for those portraits made with a 43×43-point window, which represents severe blurring. (The expected score for random guessing is 3.5 percent.)

Even more surprising were the results

of trials with photographs blurred with a 51 × 51-point window. Here the width of the window is 20 percent of the picture's width, and blurring is so extreme that facial features are entirely washed out. Nevertheless, the accuracy was almost 60 percent. (This level of recognition cannot continue with much more extensive blurring. When the averaging window includes all the picture elements, the field will be smeared to a uniform gray and pictures will differ only in the level of that gray.)

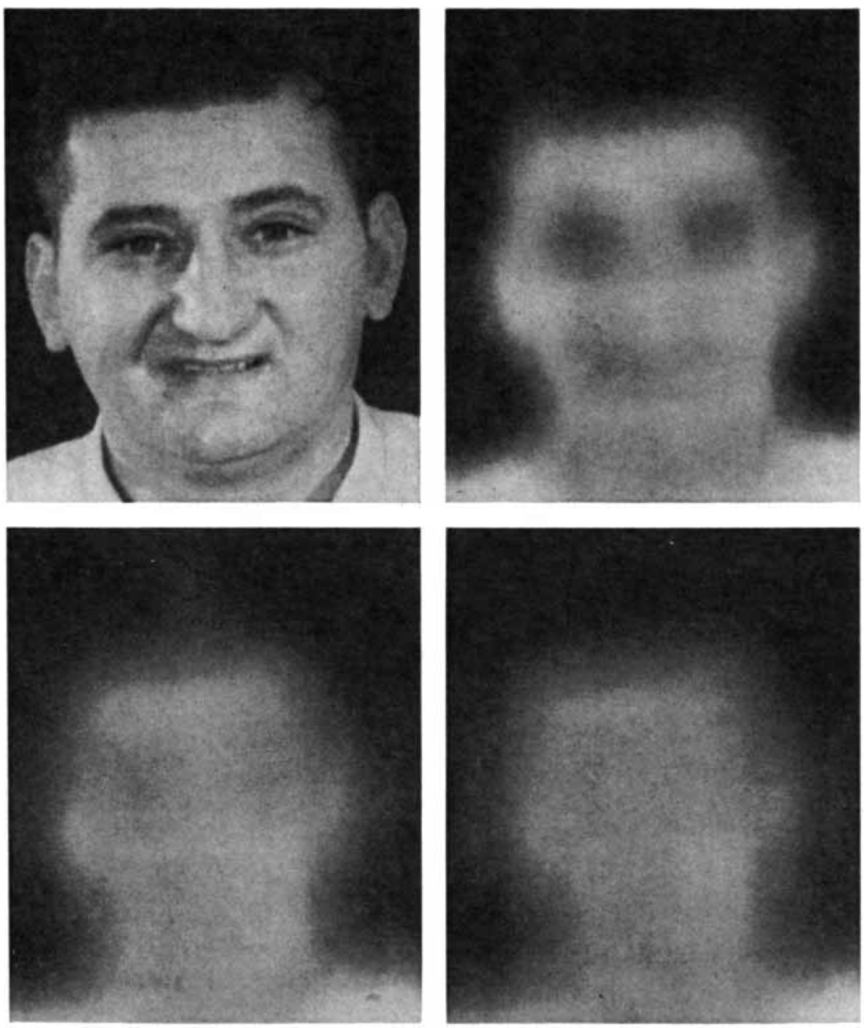
Recognition of the most strongly blurred of these portraits cannot depend on the identification of features. The high-frequency information required to represent the eyes, the ears and the mouth is lost. Although some intermediate frequencies remain, their representation of the chin, the cheeks and the hair is not clear. The low-frequency information that relates to head shape, neck-and-shoulder geometry and gross hairline is all that remains unimpaired, yet this alone seems to be adequate for rather good recognition among individuals in a restricted population.

Again, some faces were consistently well recognized. This time the responsible cues were easy to see. One portrait, for example, was distinguished by a round, bald head, and the picture was consistently recognized, even when it was badly blurred.

Some learning apparently took place in these experiments; it would appear that practice at struggling with the task improved performance.

Determining exactly how one recognizes a face is probably an intractable problem for the present. It is possible, however, to determine how well and with what cues identification can be achieved. Similarly, although machine recognition is not yet possible, search for and retrieval of faces by machine is a problem suitable for research. My colleagues and I approached these matters by investigating how effectively one can identify an individual face from a group of faces by using verbal descriptions.

It should be noted that successful identification of faces by feature descriptions does not suggest that the normal processes of recognition regularly detect and assess such features. All we can determine from experiments of this kind is how effectively people can perform a certain recognition task on the basis of certain assigned measures.



PRECISELY BLURRED PORTRAITS were constructed by a computer using an "averaging window." At top left is the original picture; it is not a continuous-tone photograph but an array of 256 × 256 dots. The averaging window determines a new value for each of the dots by averaging the values of those that surround it in some $n \times n$ field. When the window is set at 27 × 27 points (*top right*), basic facial features are still discernible. A 43 × 43-point window (*bottom left*) produces severe blurring and a 51 × 51-point window (*bottom right*) eliminates almost all information except gross forms. Accuracy of identification declined as blurring increased but even with the worst pictures approached 60 percent.

The problems of the automatic analysis of faces have received little attention. The work begun by W. W. Bledsoe and his colleagues is one of the few attempts I know of to automate the recognition of faces; the method uses a hybrid man-machine system in which a computer sorts and classifies faces on the basis of fiducial marks entered manually on photographs. The technique is called the Bertillon method, after Alphonse Bertillon, a French criminologist, and is better known for its application to fingerprint classification. A similar method has been developed by Makoto Nagao and his colleagues in Japan in an attempt to devise an automated system that would produce simple numerical descriptions of faces.

I was led to this line of inquiry by wondering if one could play a "20 questions" game with faces. (In games of this kind one player thinks of a person and the other asks him up to 20 questions, which must be answered yes or no, until the subject is guessed.) An informal, preliminary experiment began with 22 portraits; they were shown to subjects who were asked to list features they thought striking or extreme in order of decreasing extremeness. If a face displayed very wide-set eyes, for example, that statement was put at the top of the list. Or if the chin jutted extremely, that fact was listed first. A consensus list was compiled for each of the 22 faces, then new experimental subjects were selected.

Each subject was given the pile of pictures and a list of features, derived from the earlier work, describing one of the faces. He was asked to do a binary sorting, one feature at a time, starting with the most extreme and working down the list. The first sorting, therefore, produced a pile of pictures that the subject believed satisfied the first statement on the list and a pile of rejected pictures. The "accept" pile was then sorted for the second feature and for additional features until the pile was reduced to one portrait.

Two interesting questions arose: How often is the remaining portrait the correct one? How many sortings are required to reduce the population to a single member?

In this preliminary study the remaining portrait was always the correct one, and the average number of sortings required for the isolation was 4.5. We were led to wonder how the accuracy would decline as the population size increased and how rapidly the number of sortings required would grow. If the number of sortings increased linearly with population size, the process would soon become too cumbersome to be effective. (It is difficult to enumerate more than a few tens of features.) If the number of feature sortings grew, say, logarithmically, however, the process could remain useful for quite large populations.

A theoretical model devised by A. Jay Goldstein, Ann Lesk and me indicated that the feature set could indeed be expected to grow logarithmically. Under the experimental conditions we planned to employ, the feature set would grow to 5.4 for a population of 256 faces, to about 6.5 for 1,000 faces and to about a dozen features for a population of a million faces. We decided to test this model in a series of experiments using a population of 256 faces.

To make these studies it was necessary to find a pool of features that could be judged quantitatively and reliably. It was also necessary that these features be independent of one another, so that each one carried useful information.

Portraits were made of 256 faces. Each consisted of three views: frontal, three-quarter and profile. The population was deliberately made homogeneous in order to make the subsequent tasks more difficult. All the subjects were white males between 20 and 50 years

		1	2	3	4	5
HAIR	COVERAGE	FULL		RECEDING		BALD
	LENGTH	SHORT		AVERAGE		LONG
	TEXTURE	STRAIGHT		WAVY		CURLY
	SHADE	DARK	MEDIUM	LIGHT	GRAY	WHITE
FOREHEAD		RECEDING		VERTICAL		BULGING
EYEBROWS	WEIGHT	THIN		MEDIUM		BUSHY
	SEPARATION	SEPARATED		MEETING		
EYES	OPENING	NARROW		MEDIUM		WIDE
	SEPARATION	CLOSE		MEDIUM		WIDE
	SHADE	LIGHT		MEDIUM		DARK
EARS	LENGTH	SHORT		MEDIUM		LONG
	PROTRUSION	SLIGHT		MEDIUM		LARGE
CHEEKS		SUNKEN		AVERAGE		FULL
NOSE	LENGTH	SHORT		MEDIUM		LONG
	TIP	UPWARD		HORIZONTAL		DOWNWARD
	PROFILE	CONCAVE		STRAIGHT		HOOKED
MOUTH	LIP THICKNESS (UPPER)	THIN		MEDIUM		THICK
	LIP THICKNESS (LOWER)	THIN		MEDIUM		THICK
	LIP OVERLAP	UPPER	NEITHER	LOWER		
	WIDTH	SMALL		MEDIUM		LARGE
CHIN	PROFILE	RECEDING		STRAIGHT		JUTTING
		1	2	3	4	5

FACES WERE CLASSIFIED in the author's system by numerical judgments of 21 selected features. The population of 256 portraits was examined by a panel of 10 observers, who rated each face according to the 21 criteria shown in the chart at right. The judgments of the panel became the "official" values used as standards in later experiments.

old, wearing no glasses, having no beards and displaying no unusual facial marks or scars.

Starting with a tentative set of 35 features, a panel of 10 trained observers filled out questionnaires describing the 256 faces. Each feature was assigned a numerical measure, usually on a scale of from 1 to 5. After a week of tedious labor the resulting data were analyzed statistically for reliability and independence. Twenty-one features were found to be the most useful and were preserved for all further experiments [*see illustration on preceding page*]. The "official" value of each feature was taken as the average of the values assigned by the 10 observers.

With these measures of features it is possible to program a computer to sort a population of faces. If the value of each feature is considered a coordinate in 21 dimensional space, each face will represent a point in that space, the collection of features providing its 21 coordinates. The distinction between any pair of faces can be calculated simply as the Euclidean distance between the points [*see illustration below*].

With this technique one can produce listings of, say, the 100 most similar pairs of faces or the 100 least similar pairs. When the computer was instructed to name the face whose feature values were closest to the average values of those in the population, "Mr. Average" was identified through a procedure that required the nontrivial comparison of 32,640 pairs of 21-dimensional items.

While seeking the most similar faces, we discovered that the Euclidean distance separating one pair was extremely small, much smaller than the distance between the faces chosen as the authentic closest pair. When we checked the photographs, we found that the same person was shown in both. One of our colleagues had visited the studio twice, and none of the subjects examining the photographs had discovered the duplication until the computer analysis revealed it. For all subsequent study the population was reduced to 255.

Once the feature judgments were classified in the computer memory a number of simulation experiments were conducted in which the computer modeled human performance in sorting the portraits. Given a description, the computer sorted through the population, starting with the most extreme feature. It decided for each feature whether to accept or reject a given portrait on the basis of several different judgment thresholds. For the criteria and conditions that seem most reasonably to replicate human judgments the computer required about six sortings to isolate a photograph.

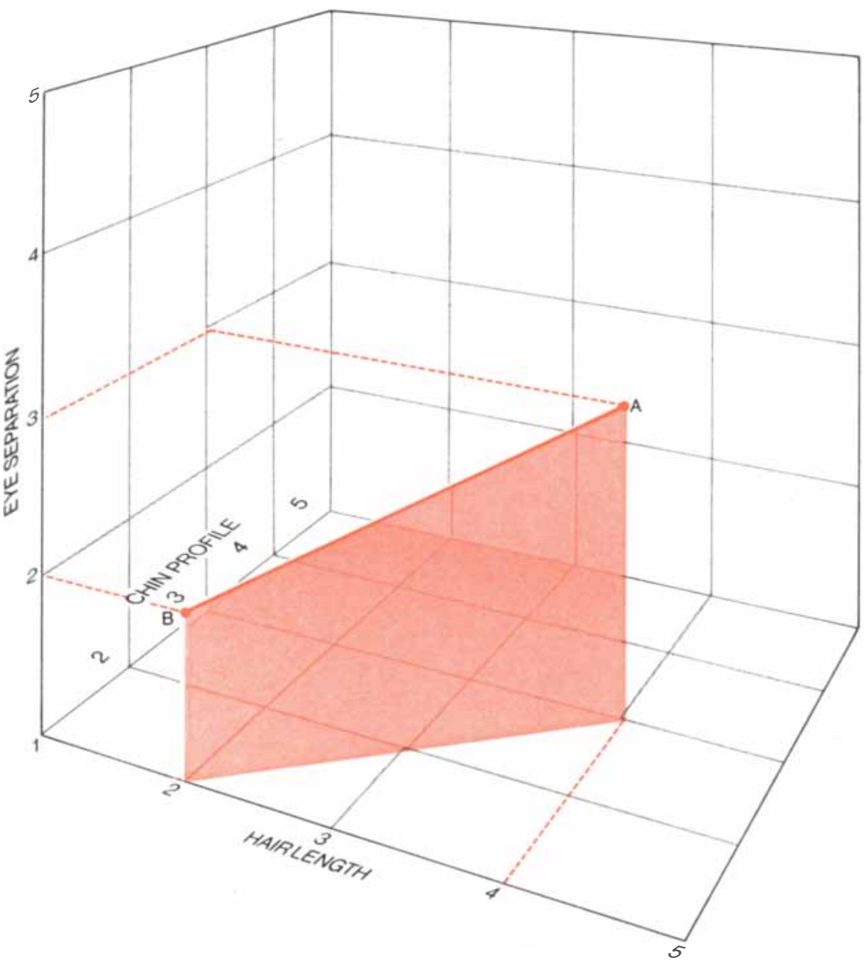
When human subjects were given the same task, 7.3 sortings were needed; the remaining portrait was the correct one in 53 percent of the trials. If one does not insist on absolute identification but asks only for a reduction of the population, the performance was fairly good. The population was reduced to no more than 5 percent in three-fourths of the trials. That is, 75 percent of the time the "target" face was included in a reduced group of no more than 13 faces.

One of the factors that limits the reliability and accuracy of this procedure is a characteristic of the binary sorting process itself: a mistake made in any decision can lead to the irretrievable loss of the target photograph. Once a portrait has been rejected it can no longer even be considered in later decisions.

A more forgiving process is rank ordering. If at each step the photograph is ranked according to how well it fits the description but is never discarded entirely, then any reasonably accurate description can be expected to place the correct portrait high in the resulting rank-ordered list. Again, even if the accuracy of the individual judgments is not extremely high, and even if a few judgments are clearly wrong, we can expect the procedure to focus attention on a small subset of the population that has a high probability of containing the target. (Population-reduction techniques of this kind are useful in many sorting tasks, such as handwriting recognition and document retrieval.)

The rank-ordered sorting leads to the fourth study to be discussed here: a system in which man and machine interact to identify faces by feature descriptions.

A subject at a computer terminal was given frontal, three-quarter and profile



THREE-DIMENSIONAL ANALOGUE of the 21-dimensional face-classification system makes each face a point inside or on the surface of a cube. For this simplified illustration three features are judged so that the assigned values become the coordinates of the point representing the face. Face *A*, for example, has a hair length of 4, eye separation of 3 and chin profile of 3. The distinction between any two faces can be measured simply as the Euclidean distance between the points. Thus the distinction between face *A* and face *B* in the drawing is $(2^2 + 2^2 + 1^2)^{1/2}$, or 3. In the 21-dimensional model each point is described by 21 coordinates and the equation for the distance between two points has 21 terms.

photographs of one member of the population. He was instructed to describe this target face to the computer, using the numerical feature values. After each description is entered on the keyboard the computer assigns a goodness-of-fit measure, called a "weight," to each member of the population. The weight represents the similarity of the subject's description to the official description. The population is ranked by weight and the list is revised each time a new feature is described. No portraits are rejected, but the target face is expected to climb through the ranks and eventually, if the process is effective, to be listed in first place.

The questions of interest are the same as those in the manual studies: How many feature steps are required and after a specified number of steps how often is the top-ranking face the correct one?

In addition to rank ordering we introduced another procedure to improve performance. In the earlier experiments the subject had chosen features to describe in descending order of extremeness. This technique takes advantage of the human ability to detect and describe conspicuous features, a process beyond the capabilities of machines. Eventually, however, and usually after finding only three or four extreme features, the observer is unable to identify more; few faces have more than four features that could be described as extreme. At this point the machine can contribute to identification in a way that would be difficult for a man.

Rather than the subject's being asked to choose features at random after he has exhausted his judgments of extreme features, he is instructed to invoke "automatic feature selection." The subject possesses exhaustive knowledge of the face he is describing, yet he knows very little about the characteristics of the population stored in the machine. The computer, on the other hand, does not know who the target is but does have the official descriptions of all the faces and their goodness-of-fit to the description that has been given so far. Automatic feature selection enables the computer to ask for a description of the feature that will be most discriminating at any stage of the identification process.

For example, if all members of the population have close-set ears, a description of that feature would not discriminate between faces. The most discriminating feature is the one that has the most uniform distribution of judged values over the permitted range. The computer, having knowledge of these statistics, can select the sequence of features that will most efficiently separate the members of the population.

After a few feature-description steps, directed by the human subject, the probability is high that the target face has risen in the rank-ordered list. The computer can therefore confine its search for discriminating features to some subset of the population that the portrait thus far describes well. This has the effect of enhancing discrimination of the target.

In our experiments subjects were told to describe conspicuous features until no more were apparent and then to invoke automatic feature selection. The procedure was terminated after 10 steps, since theory and previous experience predicted that this should be sufficient for good accuracy.

Performance was excellent. The subjects' votes and the official values were in good agreement. In spite of the vaga-

MOST SIMILAR PAIR of faces was found by comparing the 21 judged values of the features of the 32,640 possible combinations. The operation was performed by a computer.

LEAST SIMILAR PAIR of faces was determined by the same procedure. The population was deliberately made homogeneous: all members were white males from 20 to 50 years old.

ries of subjective judgment the difference between the experimental and the official values, in a scale typically ranging from 1 to 5, was less than one in more than 95 percent of the trials. Accuracy of identification was also impressive. The population was reduced to less than 4 percent 99 percent of the time, that is, the correct face was in 10th place or better in 99 of every 100 trials. By the 10th sorting the target was in first place in 70 percent of all trials.

In control experiments features were selected by the human operator only and by the computer only; in both cases performance was poorer than when both man and machine participated.

This last exercise has more general applications than the identification of faces. It is a technique for the retrieval of any multidimensional vectors by information obtained from imprecise descriptions. Such probabilistic file searches are important in answering telephone-directory assistance inquiries, in medical diagnosis and in law-enforcement information retrieval.

Our studies have touched on a host of questions about human perception, automatic pattern recognition and procedures for information retrieval. Although the ultimate question of how a face is recognized remains unanswered, a few promising lines of inquiry have emerged. It has once again been clearly shown that the human viewer is a fantastically competent information processor. In some recognition tasks a synergy of man and machine is effective, but in further explorations of the identification of complex images both by men and by machines there is much to learn.

```
DESCRIBE NEXT PICTURE.

FEATURE
      EYEBRØW WT.
THIN        MEDIUM       BUSHY
    1    2    3    4    5
=1
  93  244  183  223  159
1.00 1.00 1.00 1.00 0.82

FEATURE
      EAR LENGTH
SHØRT       MEDIUM       LØNG
    1    2    3    4    5
=1
  72  244  175   93   43
1.00 1.00 0.82 0.67 0.66

FEATURE
      LIP ØVERLAP
UPPER       NEITHER      LØWER
    1          2            3
=1
  72  226  114  122   76
1.00 0.73 0.66 0.61 0.60

FEATURE
      HAIR TEXTURE
STRAIGHT    WAVY         CURLY
    1    2    3    4    5
=4
  76  122   32  244   52
1.00 0.74 0.56 0.55 0.50

FEATURE

        AUTØMATIC FEATURE SELECTIØN
******EYE SHADE
LIGHT       MEDIUM       DARK
    1    2    3    4    5
=3
  76   52   72  221  191
1.00 0.56 0.45 0.38 0.36

******EYEBRØW SEP.
SEPARATE    MEDIUM       MEETING
    1          2            3
=2
  76  147   52   84   72
1.00 0.50 0.42 0.37 0.34

******EYE ØPENING
NARRØW      MEDIUM       WIDE
    1    2    3    4    5
=2
  76   72  226   26  191
1.00 0.51 0.40 0.38 0.36

******UPPER LIP
THIN        MEDIUM       THICK
    1    2    3    4    5
=3
  76  191   72  221   52
1.00 0.33 0.28 0.23 0.21

******HAIR SHADE
DARK   MED.   LT.   GRAY   WHT.
    1    2    3    4    5
=2
  76  221   72  226  191
1.00 0.34 0.34 0.33 0.25

******LØWER LIP
THIN        MEDIUM       THICK
    1    2    3    4    5
=1
  76   72  221   84  191
1.00 0.19 0.13 0.12 0.11

PLEASE TYPE TARGET NUMBER.
=76
```

ØRDER	FEATURE	DESCRIPTIØN YØU	AVG.	RANK NØ.	%
1	EYEBRØW WT.	1	2.2	27	10.2
2	EAR LENGTH	1	2.3	8	2.7
3	LIP ØVERLAP	1	1.2	5	1.6
4	HAIR TEXTURE	4	3.0	1	0.
5	EYE SHADE	3	2.7	1	0.
6	EYEBRØW SEP.	2	1.3	1	0.
7	EYE ØPENING	2	2.6	1	0.
8	UPPER LIP	3	2.9	1	0.
9	HAIR SHADE	2	1.5	1	0.
10	LØWER LIP	1	2.3	1	0.

DIALOGUE WITH A COMPUTER records a search for a "target" face. The computer "speaks" first and requests a description; the subject replies by announcing that he will describe eyebrow weight. The computer then prints the range of allowable values. The subject selects "1" for "thin" and the computer ranks each member of the population according to how well it fits this value. The five members that best fit the description are printed in the next line, followed by their relative "weights." The first four faces here are tied with weights of 1.00. The target face in this trial was No. 76; by the third step it was in fifth place and by the fourth step in first place. After the fourth feature description the subject called for "automatic feature selection," which enables the computer to request descriptions of those features that would be most discriminating. After the 10th step No. 76 had a weight of 1.00 and its nearest neighbor a weight of .19. The correct face was clearly identified even though the first two descriptions were in error. Following the dialogue is a summary comparing the subject's judgments with the "official" values (*AVG.*) and showing the rank of the target at each step and the percent of the population with a higher rank. The procedure was stopped after 10 steps; 21 steps were possible.


The Silver Martini.
For people who want a silver lining
without the cloud.
SMIRNOFF SILVER
VODKA
Smirnoff Silver
Ninety point four proof. Smirnoff leaves you breathless

HILBERT'S 10TH PROBLEM

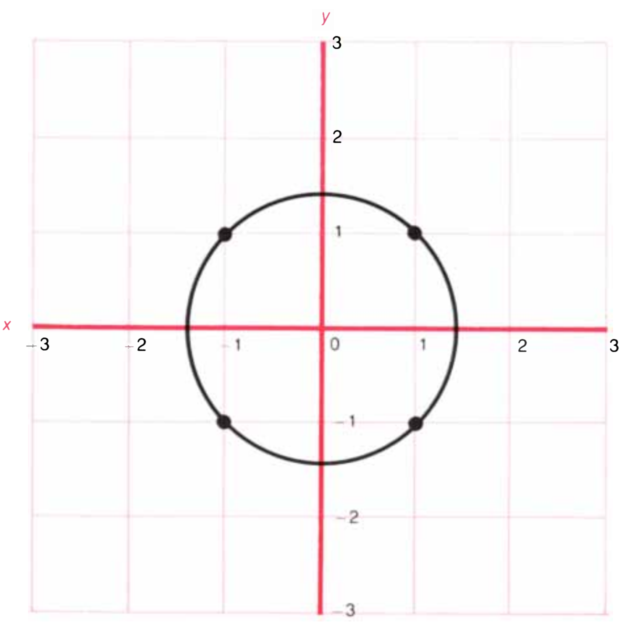
Can a procedure be devised that will indicate if there are solutions to a Diophantine equation (an equation where whole-number solutions are sought)? This question on a famous list has now been answered

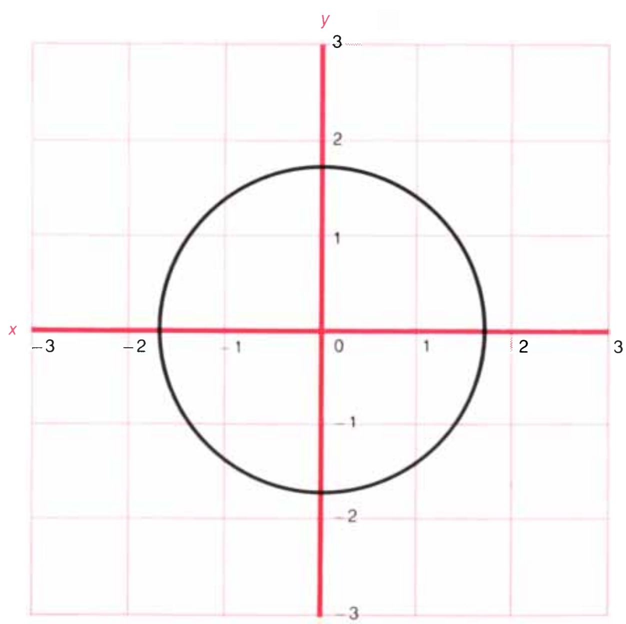
by Martin Davis and Reuben Hersh

"We hear within us the perpetual call: there is the problem. Seek its solution. You can find it by pure reason, for in mathematics there is no *ignorabimus* [We shall not know]." So did David Hilbert address the Second International Congress of Mathematicians in Paris on August 8, 1900, greeting the new century by presenting a list of 23 major problems to challenge future mathematicians. Some of Hilbert's problems are still unsolved. Others have inspired generations of mathematical investigators and have led to major new mathematical theories. The most recently conquered of Hilbert's problems is the 10th, which was solved in 1970 by the 22-year-old Russian mathematician Yuri Matyasevich.

David Hilbert was born in Königsberg in 1862 and was professor at the University of Göttingen from 1895 until his death in 1943. After the death of Henri Poincaré in 1912 he was generally regarded as being the foremost mathematician of his time. He made fundamental contributions in several fields, but he is perhaps best remembered for his development of the abstract method as a powerful tool in mathematics.

Hilbert's 10th problem is easily described. It has to do with the simplest and most basic mathematical activity: solving equations. The equations to be solved are polynomial equations, that is, equations such as $x^2 - 3xy = 5$, which are formed by adding and multiplying constants and variables and by using whole-number exponents. Moreover, Hilbert specified that the equations must use only integers, that is, positive or negative whole numbers. No irrational





GRAPHS OF TWO EQUATIONS illustrate the difference between an ordinary equation and a Diophantine equation, for which one is interested only in whole-number solutions; this difference is central to Hilbert's 10th problem. The equations in point are $x^2 + y^2 - 2 = 0$ (*left*) and $x^2 + y^2 - 3 = 0$ (*right*); both are represented by circles with their center at the origin, that is, at the point with coordinates $x = 0$, $y = 0$. In the case of $x^2 + y^2 - 2 = 0$ the circle has a radius of $\sqrt{2}$. If the equation is treated as an ordinary equation, there are infinitely many solutions. If, however, it is treated as a Diophantine equation, there are only four solutions: (1) $x = 1$, $y = 1$, (2) $x = -1$, $y = 1$, (3) $x = 1$, $y = -1$, and (4) $x = -1$, $y = -1$. These solutions are represented by dots where the graph crosses the four points with those coordinates on the Cartesian grid. In the case of $x^2 + y^2 - 3 = 0$, the circle has a radius of $\sqrt{3}$. As an ordinary equation it has an infinite number of solutions; as a Diophantine equation, however, it has none at all.

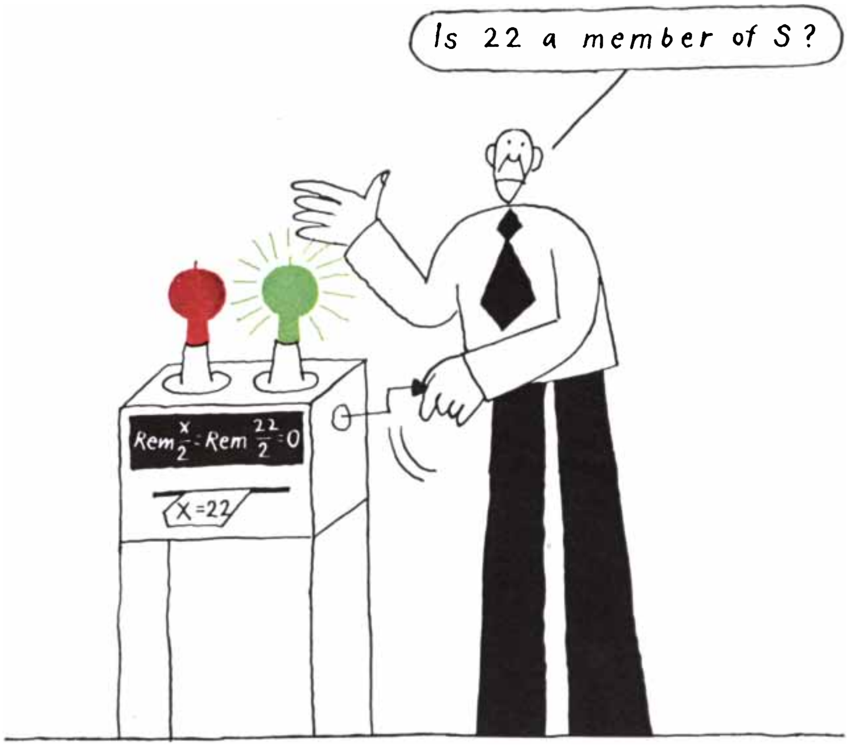
or imaginary numbers or even fractions are allowed in either the equations or their solutions. Problems of this type are called Diophantine equations after Diophantus of Alexandria, who wrote a book on the subject in the third century.

Hilbert's 10th problem is: Give a mechanical procedure by which any Diophantine equation can be tested to see if solutions exist. In Hilbert's words: "Given a Diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: to devise a process according to which it can be determined by a finite number of operations whether the equation is solvable in rational integers." Hilbert does not ask for a process to find the solutions but merely for a process to determine if the equation has solutions. The process should be a clear-cut formal procedure that could be programmed for a computing machine and that would be guaranteed to work in all cases. Such a process is known as an algorithm.

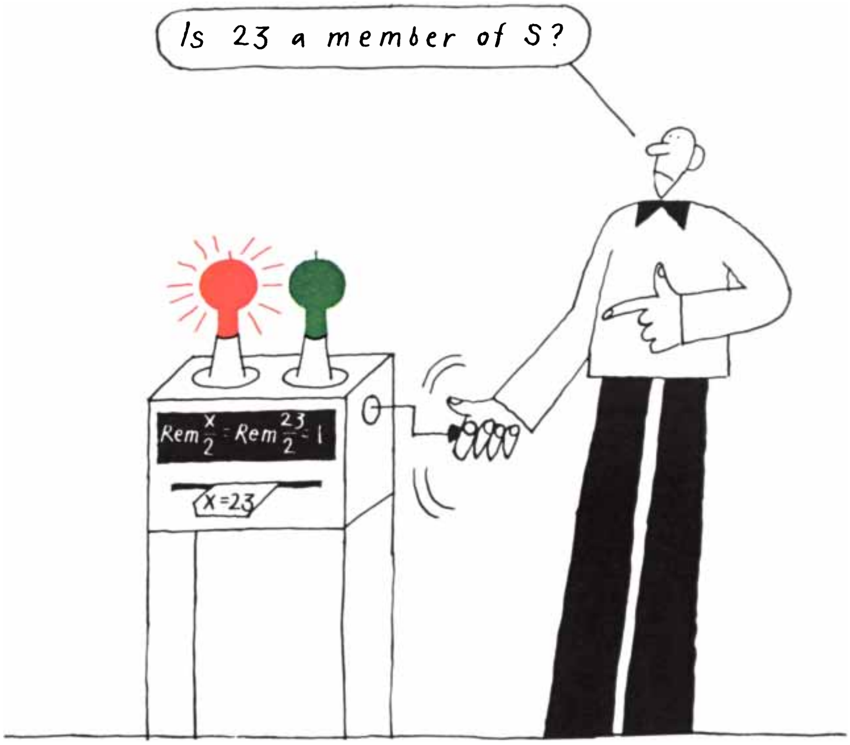
If Hilbert's problem is simply stated, Matyasevich' solution is even more simply stated: No such process can ever be devised; such an algorithm does not exist. Worded in this way, the answer sounds disappointingly negative. Matyasevich' result, however, constitutes an important and useful addition to the understanding of properties of numbers.

Matyasevich' work extended a series of researches by three Americans: one of us (Davis), Julia Robinson and Hilary Putnam. Their work in turn was based on earlier investigations by several founders of modern logic and computability theory: Alan Turing, Emil Post, Alonzo Church, Stephen Kleene and the same Kurt Gödel who is famous for his work on the consistency of axiomatic systems (Hilbert's second problem) and on the continuum hypothesis of Cantor (Hilbert's first problem).

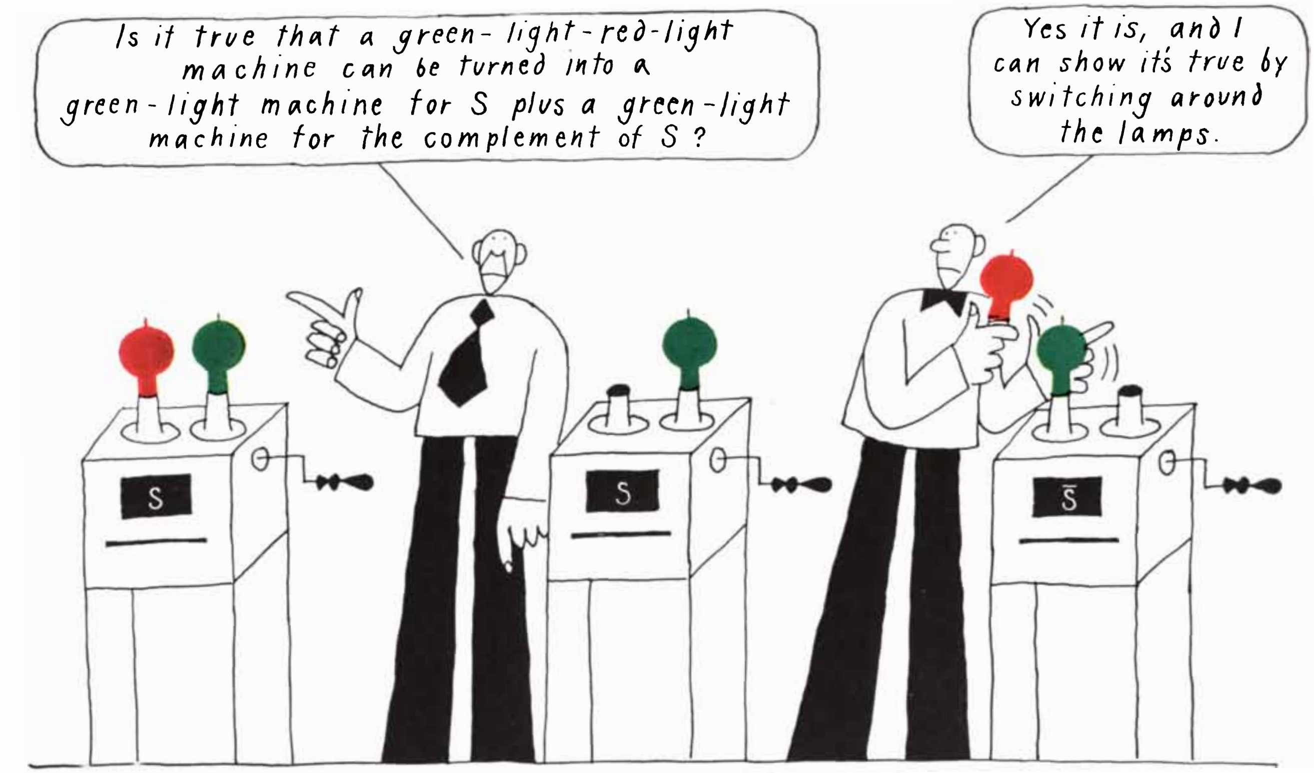
Let us start on Hilbert's 10th problem by looking at a few Diophantine equations. The term "Diophantine equation" is slightly misleading, because it is not so much the nature of the equation that is crucial as the nature of the admissible solutions. For example, the equation $x^2 + y^2 - 2 = 0$ has infinitely many solutions if one does not think of it as a Diophantine equation. The solutions are represented by the graph of the equation, which is a circle in the plane formed by the x axis and the y axis. The center of the circle is at the coordinates $x = 0$, $y = 0$. That point is called the origin; it is abbreviated (0,0). The radius of the circle is $\sqrt{2}$ [*see illustration on opposite page*]. The coordinates of any



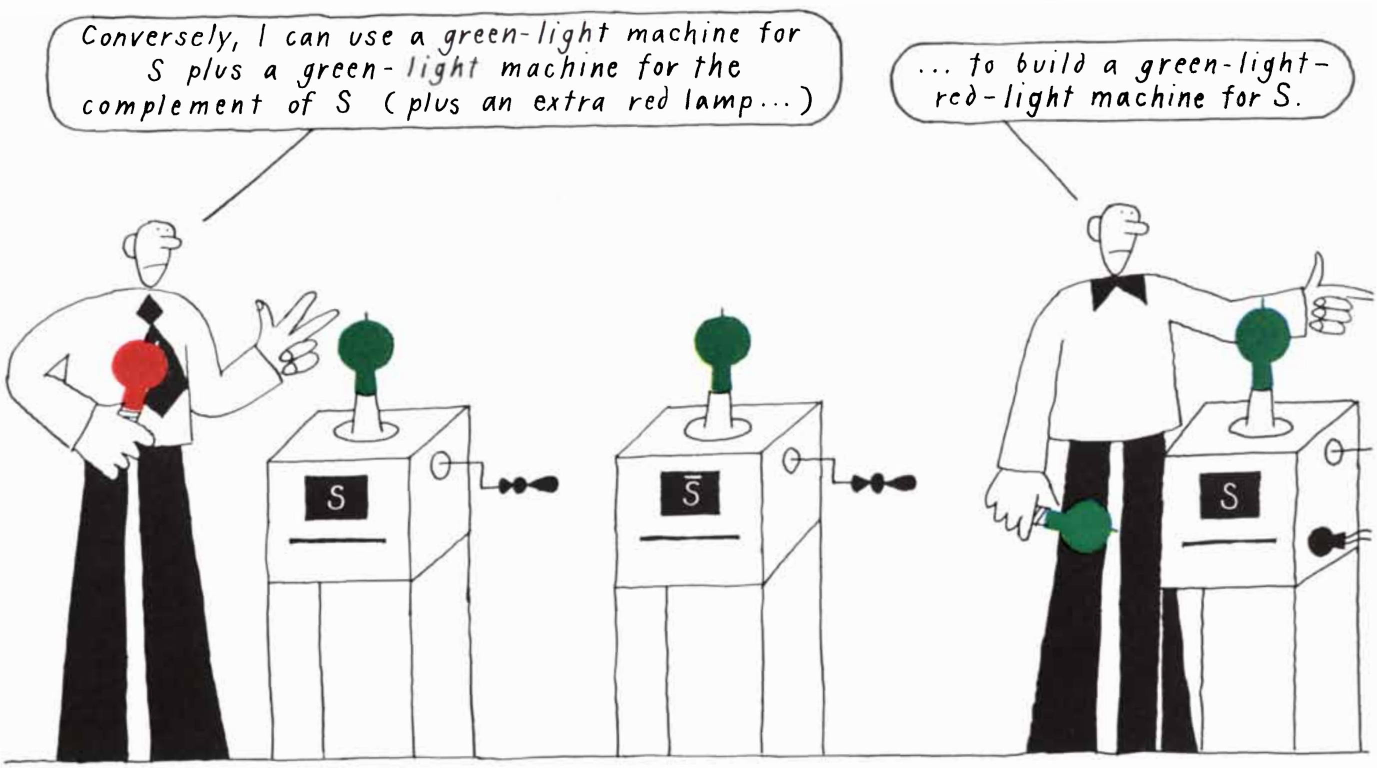
GREEN-LIGHT-RED-LIGHT MACHINE is an imaginary device that tests numbers to determine if they are members of a given set. Hilbert's 10th problem asks if a green-light-red-light "Hilbert machine" can be built to test Diophantine equations to see whether or not they have solutions. In the case of testing numbers for membership in a set, green light goes on if the machine can determine in a finite number of steps that a given input is a member of the set. Say that S is the set of all even numbers. To test inputs one can devise an algorithm for dividing each input x by 2. If the remainder of the division is 0 (written Rem $x/2 = 0$), machine would turn on its green light, signifying that x is a member of S.



RED LIGHT GOES ON on the green-light-red-light machine if the machine can determine that the input is not a member of the set. Suppose the input x is the whole number 23; 2 goes into 23 with a remainder of 1, signifying that 23 is not a member of S. Complement of set S is $\overline{S}$, the set of odd numbers; 23 is a member of $\overline{S}$. Since a green-light-red-light machine can be built to sort members of S from members of $\overline{S}$, the set S is called computable.



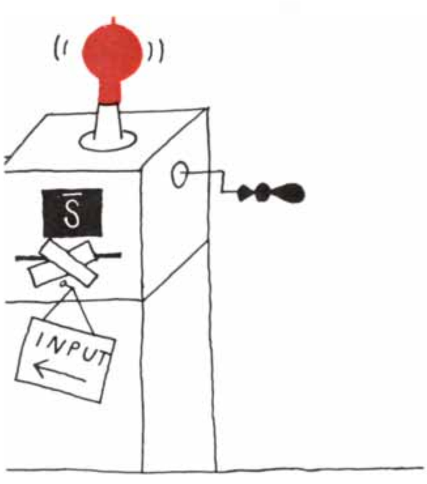
GREEN-LIGHT-RED-LIGHT MACHINE FOR THE SET *S* can be transformed into a green-light machine for *S* (that is, a machine that simply lights up when the input is a member of *S*) plus a green-light machine for $\overline{S}$, the complement of *S*. The proof is simple. To build a green-light machine for *S*, unscrew the red lamp of the green-light-red-light machine. To build a green-light machine for $\overline{S}$, unscrew the green lamp of the green-light-red-light machine and put it into the socket that held the red lamp. This fact can be stated in another way. If a set (such as *S*) is computable, then both the set and its complement (such as $\overline{S}$) are listable, that is, the members of *S* (in this case the set of even numbers) can be listed separately and sorted from the members of $\overline{S}$ (the set of odd numbers).



GREEN-LIGHT MACHINE FOR EACH OF *S* AND $\overline{S}$ can be used to construct a green-light-red-light machine for the set *S*. This statement is the converse of the one for the top illustration on this page. In the green-light machine for $\overline{S}$ replace the green lamp with a red lamp. Then hook the machines in parallel so that the input goes into both simultaneously. The result is clearly a green-light-red-point on the circle satisfy the equation, and there are an infinite number of such points. If we consider the problem as a Diophantine equation, however, there are only four solutions: (1) $x = 1, y = 1$; (2) $x = -1, y = 1$; (3) $x = 1, y = -1$, and (4) $x = -1, y = -1$.

Suppose the equation is changed to $x^2 + y^2 - 3 = 0$. There are still an infinite number of solutions if it is treated as an ordinary equation but no solutions at all if it is treated as a Diophantine equation. The reason is that now the graph is a circle with radius equal to $\sqrt{3}$, and no points on this curve have both coordinates simultaneously equal to whole numbers.

A famous family of Diophantine equations has the form $x^n + y^n = z^n$, where n may equal 2, 3, 4 or any larger integer. If n is equal to 2, the equation is satisfied by the lengths of the sides of any right triangle and is called the Pythagorean theorem. One such solution is the set of numbers $x = 3$, $y = 4$, $z = 5$. If n is equal to or greater than 3, the equation is what is known as Fermat's equation. The 17th-century French mathematician Pierre de Fermat thought he had proved that these equations have no positive whole-number solutions. In the margin of his copy of Diophantus' book he wrote that he had found a "marvelous proof"



light machine. This assertion can be stated differently: If both a set and its complement are listable, then the set is computable.

that was unfortunately too long to be written down in that space. The proof (if indeed Fermat had one) has never been found. Known as Fermat's last theorem, it is probably the oldest and most famous unsolved problem in mathematics. These examples show that Diophantine equations are easy to write down but hard to solve. They are hard to solve because we are so exclusive about the kind of numbers we accept as solutions.

For first-degree equations, that is, equations in which unknowns are not multiplied together and all exponents are equal to 1, such as $7x + 4y - 3z - 99t + 13u - 10 = 0$, the existence of solutions can be determined by a technique of division known since ancient times as Euclid's algorithm. For second-degree equations with two unknowns, such as $3x^2 - 5y^2 + 7 = 0$ or $x^2 - xy - y^2 = 1$, a theory developed early in the 19th century by the great Karl Friedrich Gauss enables one to determine whether there are any solutions. Recent work by the young British mathematician Alan Baker has shed considerable light on equations greater than the second degree that have two unknowns. For equations greater than the first degree that have more than two unknowns, there exist only some special cases that can be handled by special tricks, and a vast sea of ignorance.

Why is it so difficult to find a process such as the one Hilbert called for? The most direct approach would be to simply test all possible sets of values of the unknowns, one after another, until a solution is found. For example, if the equation has two unknowns, one could make a list of all pairs of integers. Then one would simply go through the list trying one pair after another to see if it satisfies the equation. This is certainly a clear-cut, mechanical procedure that a machine could carry out. What will be the result?

If the equation is the first one we mentioned, $x^2 + y^2 - 2 = 0$, one would test (0,0), (0,1), (1,0), (0,−1), (−1,0) and reject them all. The next candidate, (1,1), is a solution. We were lucky: only six pairs had to be considered. If, on the other hand, the equation were $x^2 + y^2 = 20{,}000$, one would have to test thousands of pairs of numbers before a solution was found. Still, it is clear that if a solution exists, it will be found in a finite number of steps.

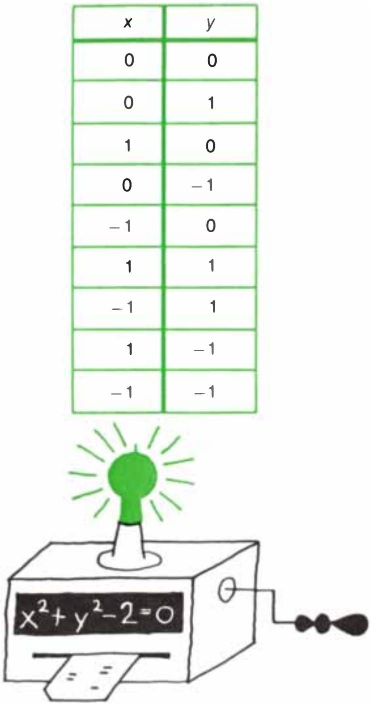
On the other hand, what about the second equation: $x^2 + y^2 - 3 = 0$? One can try pairs of integers from now till eternity, and all that will ever be known is that a solution has not been found yet. One would never know whether or not the next pair tried would be a solution. For this particular example it is possible to prove there are no solutions. But the proof requires a new idea; it cannot be obtained merely by successively substituting integers into the equation.

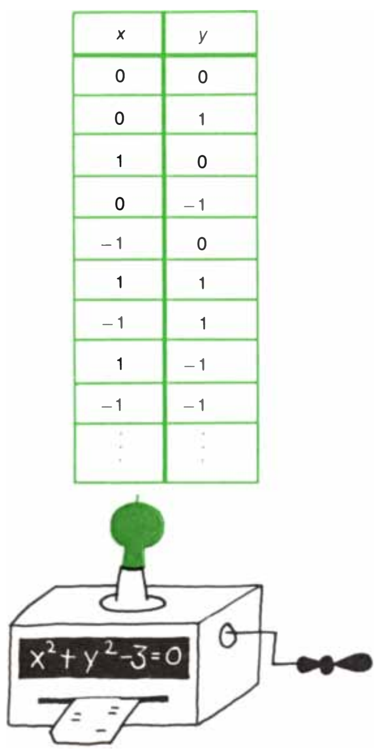
A device that carries out a process of the kind suggested by Hilbert should accept as an input the coefficients of an arbitrary Diophantine equation. As an output it should turn on a green light if the equation has a solution and a red light if it has none. Such a machine might be called a Hilbert machine. By way of contrast a device that simply searches for solutions by successive trials *ad infinitum* could be described as a green-light machine. If the equation has a solution, the green light goes on after a finite number of steps. If the equation has no solution, the computation simply goes on forever; unlike the Hilbert machine, the green-light machine has no way of knowing when to give up.

It is easy to build a green-light machine for Diophantine equations. The question is, can we do better and build a Hilbert machine, that is, a green-light-red-light machine that will always stop after a finite number of steps and give a definite yes or no answer? What Matyasevich proved is that this can never be done. Even if we allow the machine unlimited memory storage and unlimited computing time, no program can ever be written and no machine can ever be built that will do what Hilbert wanted. A Hilbert machine does not exist.

Hilbert continued in his address of 1900: "Occasionally it happens that we seek the solution under insufficient hypotheses or in an incorrect sense, and for this reason do not succeed. The problem then arises: to show the impossibility of the solution under the given hypotheses, or in the sense contemplated." That is exactly what has happened with the 10th problem.

In order to explain how we know that no Hilbert machine exists, we have to discuss some simple ideas about computability. Suppose S stands for a set of integers. S is "listable" if a green-light machine can be built that will do the following job: accept any integer as an input, and as an output turn on a green light after a finite number of steps if and only if the input (the integer) belongs to S. For example, the set of even numbers is listable. In this case the machine would divide the input by 2 and turn on a green light if the remainder is 0. In mathematical literature such sets are called recursively enumerable; the word





PAIRS OF INTEGERS can be individually tested by green-light machines to see if they are solutions to Diophantine equations. Trial and error comes up with a solution for the equation $x^2 + y^2 - 2 = 0$ on the sixth try (*top*). Green-light machine testing equation $x^2 + y^2 - 3 = 0$ has no way of knowing when to give up, however, because there are no whole-number solutions (*bottom*). All it knows is that it has found no solutions yet.

"listable" is our informal equivalent.

The set S is "computable" if a green-light-red-light machine (similar to the Hilbert machine for Diophantine equations) can be built to do a more difficult job: accept any integer as input and, after a finite number of steps, turn on a green light if the integer is in S and a red light if the integer is not in S. For example, the set of even numbers is computable. The machine would divide the input by 2; if the remainder is 0, it turns on a green light, and if the remainder is 1, it turns on a red light [*see illustrations on page* 85].

There is a close connection between these two definitions. For the purposes of explanation, let $\bar{S}$ denote the complement of S, that is, the set of all integers that do not belong to S. If in the two examples S is the set of even integers, then $\bar{S}$ is the set of odd integers. We can prove that if S is computable, S and $\bar{S}$ are both listable. To put that statement another way: If a green-light-red-light machine exists for S, then there exists a green-light machine for S and a green-light machine for $\bar{S}$. The proof is simple. To build a green-light machine for S, just unscrew the red bulb of the green-light-red-light machine. To build a green-light machine for $\bar{S}$, unscrew the green bulb of the Hilbert machine and put it into the socket that held the red bulb.

The converse is also true: If S and $\bar{S}$ are listable, then S is computable. The equivalent of this statement is: If a green-light machine exists for each of S and $\bar{S}$, then a green-light-red-light machine can be built for S. This is easily done. In the green-light machine for $\bar{S}$, replace the green bulb with a red bulb. Then hook up the two machines in parallel, so that the input goes into both simultaneously. The result is clearly a green-light-red-light machine.

Knowing all of this, we can now state one of the crucial facts in computability theory, one that plays a central role in the solution of Hilbert's 10th problem: There is a set K that is listable but not computable! That is, there exists a green-light machine for K, but it is impossible to build a green-light machine for $\bar{K}$, the complement of K.

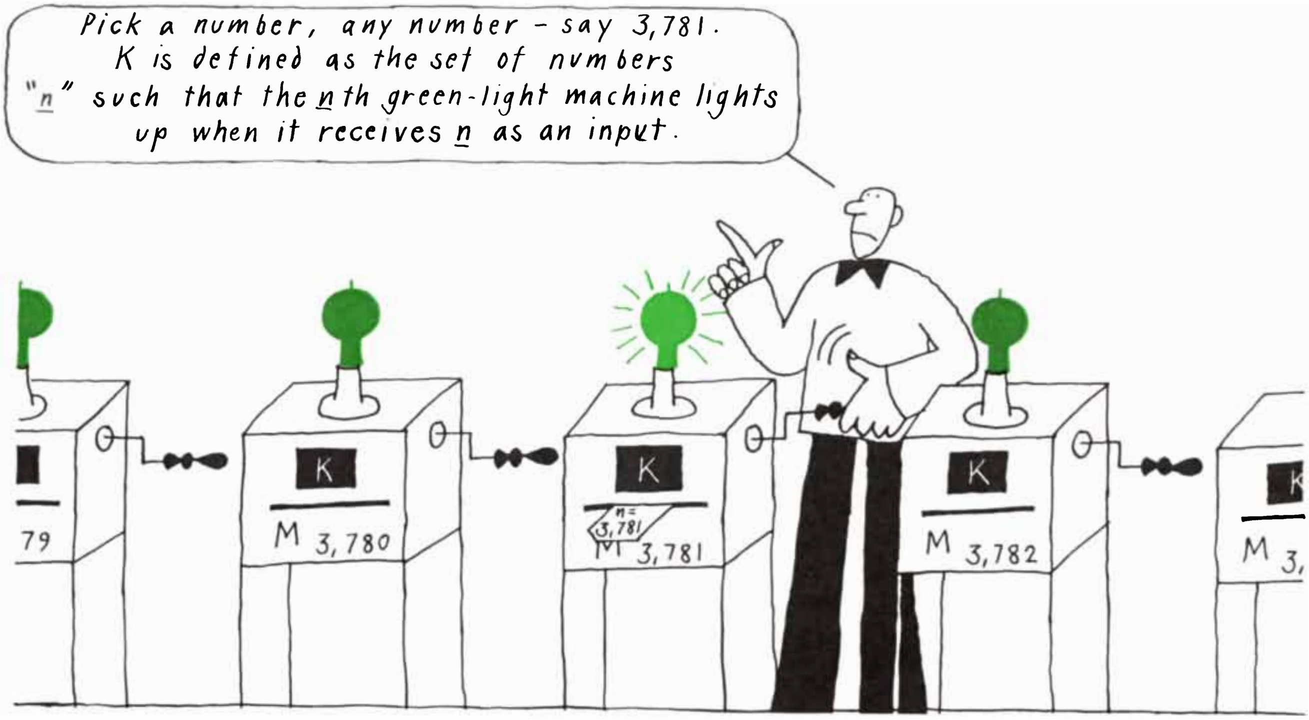
To prove this seemingly strange fact, let each green-light machine be specified by a detailed "customer's manual" in the English language. The customer's manual describes exactly how the machine is constructed. The customer's manuals can be set in order and numbered sequentially 1, 2, 3 and so on. In that way all green-light machines are numbered; M_1 is the first machine, M_2 is the second and so on. There is a subtle point hidden here. Such an ordered list of customer's manuals would not be possible for green-light-red-light machines. The difficulty is that one cannot tell from the manual whether the red light or the green light will turn on for any input to the corresponding machine.

The set K is defined as the set of numbers n such that the nth machine lights up when it receives n itself as an input. In other words, the number 1 belongs to K if and only if M_1 turns on its green light when "1" is entered into its input. The number 2 belongs to K if and only if M_2 eventually lights up when "2" is entered into its input, and so on [*see top illustration on opposite page*].

In order to build a green-light machine for K we need, along with the library of customer's manuals, a little man who can read them and carry out their instructions. He should perhaps be a wise old man, but he must be an obedient man who does exactly what he is told. We give the little man a number, say 3,781. The little man looks into customer's manual No. 3,781. Reading the manual, he is able to build the green-light machine $M_{3,781}$. Once this is done, he inserts the integer 3,781 as input into green-light machine $M_{3,781}$. If the green light goes on, the number 3,781 belongs to K. Thus we have a green-light machine for K.

What about $\bar{K}$? How can we be sure there is no green-light machine for it? Well, suppose there were such a machine. Then since $\bar{K}$ is the complement of K, this machine should light up for any input, say for 297, if and only if M_{297} does *not* light up for 297. (If M_{297} lit up, it would mean that the integer 297 belongs to K and not to $\bar{K}$.) Thus the machine for $\bar{K}$ certainly is not the same as M_{297} [*see bottom illustration on opposite page*]. By the same token, however, it is not the same as M_n for any other value of n. The same argument would apply to any other number just as well as to 297, and it shows that no green-light machine for $\bar{K}$ appears anywhere in the library of customer's manuals. Since every possible green-light machine eventually turns up in our list, it follows that no green-light machine for $\bar{K}$ can possibly exist. That is to say, $\bar{K}$ is not listable.

The result is certainly remarkable. It deserves contemplation and appreciation. We know perfectly well what the set K is; in principle we can produce as much of it as we wish with a computer printout. Nevertheless, there can never be a formal procedure (an algorithm or a machine program) for sorting K from



THE SET K IS LISTABLE, that is, a green-light machine for K exists. Let all conceivable green-light machines be numbered: M_1 is the first machine, M_2 is the second machine, M_3 is the third machine and so forth up to the nth machine. K is defined as the set of numbers n such that the nth machine lights up when it receives n itself as an input. In the illustration a little man has entered the number 3,781 as an input to $M_{3,781}$ and the green light has turned on, indicating that the whole number 3,781 is a member of set K.

THE SET K IS NOT COMPUTABLE, that is, no green-light machine exists for $\overline{K}$, the complement of K. Suppose there was such a green-light machine for $\overline{K}$. Since $\overline{K}$ is the complement of K, this machine should light up for any input, say for 297, if and only if M_{297} does not light up for 297. Thus the machine for $\overline{K}$ is certainly not the same as M_{297}. By the same token, it is not the same as M_n for any other value of n. Thus no green-light machine exists for $\overline{K}$, meaning that $\overline{K}$ is not listable. A listable set whose complement is not listable is not computable; no green-light-red-light machine can be built for it. Thus there is no algorithm for sorting K from $\overline{K}$.

1. 1
2. 1
3. 1 + 1 = 2
4. 1 + 2 = 3
5. 2 + 3 = 5
6. 3 + 5 = 8
7. 5 + 8 = 13
8. 8 + 13 = 21
9. 13 + 21 = 34
10. 21 + 34 = 55
11. 34 + 55 = 89
12. 55 + 89 = 144
13. 89 + 144 = 233

⋮

n. $\approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n}$

FIBONACCI NUMBERS were discovered in A.D. 1202 by Leonardo of Pisa, known as Fibonacci. The sequence is obtained by starting with 1 and 1 and successively adding the last two numbers to get the next one. The sequence grows exponentially: the *n*th number in the sequence is approximately proportional to the *n*th power of the real number $[(1+\sqrt{5})/2]^n$.

PROBLEM: To find the smallest number n that has the remainders of 4, 2, 3 and 1 when it is divided by 10, 3, 7 and 11.

SOLUTION: Let x be the number sought. "Rem" will be the abbreviation for "The remainder of . . . ." The problem can then be rewritten:

$$\text{Rem}\left(\frac{x}{10}\right) = 4 \qquad \text{Rem}\left(\frac{x}{7}\right) = 3$$

$$\text{Rem}\left(\frac{x}{3}\right) = 2 \qquad \text{Rem}\left(\frac{x}{11}\right) = 1$$

In order to find x four auxiliary problems for new unknowns y_1, y_2, y_3 and y_4 must be solved. In each case the numerator is obtained by multiplying three of the divisors together and using the fourth as the denominator. For example, in the first equation with y_1 the numerator 231 is equal to $3 \times 7 \times 11$, and 10 is put in the denominator:

$$\text{Rem}\left(\frac{231y_1}{10}\right) = 4,\ y_1 < 10 \qquad \text{Rem}\left(\frac{330y_3}{7}\right) = 3,\ y_3 < 7$$

$$\text{Rem}\left(\frac{770y_2}{3}\right) = 2,\ y_2 < 3 \qquad \text{Rem}\left(\frac{210y_4}{11}\right) = 1,\ y_4 < 11$$

The set of smallest integers that are solutions to these auxiliary equations is $y_1 = 4$, $y_2 = 1$, $y_3 = 3$ and $y_4 = 1$.

To get x (the original number sought) the numerators of the four auxiliary equations are added together:

$$x = (231y_1) + (770y_2) + (330y_3) + (210y_4)$$

$$= (231 \times 4) + (770 \times 1) + (330 \times 3) + (210 \times 1)$$

$$= 924 + 770 + 990 + 210$$

$$= 2{,}894$$

Thus 2,894 is one value of x. A smaller number can be obtained if the product of all four divisors is subtracted from this solution:

$$2{,}894 - (10 \times 3 \times 7 \times 11) = 2{,}894 - 2{,}310 = 584.$$

Therefore 584 is the smallest solution to the problem.

CHINESE REMAINDER THEOREM is used in the solution to Hilbert's 10th problem. In this case the theorem is employed to find a number whose remainders, when divided by the numbers 10, 3, 7 and 11, are respectively 4, 2, 3 and 1. Integer 584 is the smallest solution.

$\overline{K}$. Thus here is an example of a precisely stated problem that can never be solved by mechanical means.

This discussion has of course been informal and nonrigorous. It is possible, however, to reformulate all the ideas and arguments with precise mathematical definitions and proofs. In fact, they have been formulated in a branch of mathematical logic called recursive function theory, established in the 1930's by Gödel, Church, Post, Kleene and Turing.

Now, what has all this to do with Diophantine equations? Simply this. Matyasevich has proved that every listable set has a corresponding Diophantine equation. More precisely, if S is a listable set, then there is a corresponding polynomial P, with integer coefficients and variables $x, y_1, y_2, \ldots, y_n$, which is denoted by $P_S(x, y_1, y_2, \ldots, y_n)$. Any integer, such as 17, belongs to set S if and only if the Diophantine equation $P(17, y_1, y_2, \ldots, y_n) = 0$ has a solution.

It might be thought that for some sets we would have to resort to inconceivably complicated polynomials, but this is not the case. The degree of P need not exceed the fourth power; the number of variables $y_1, y_2, \ldots, y_n$ need not exceed 14. (No one knows yet if both of these bounds can be achieved simultaneously.)

This result of Matyasevich' quickly leads to the conclusion that no Hilbert machine can exist. Recall the listable set K constructed a few paragraphs above. According to Matyasevich, there is a Diophantine equation, $P_K(x, y_1, y_2, \ldots, y_n) = 0$, associated with this set. If it were possible to build a Hilbert machine, that is, a green-light-red-light machine for testing Diophantine equations to see if they have solutions, then for any integer x we could determine whether or not there existed integers $y_1, y_2, \ldots, y_n$ such that the equation has a solution. In so determining, however, we would also be determining whether or not x belongs to K. In other words, a Hilbert machine applied to the Diophantine equation that describes K could be used as a green-light-red-light machine for K. We have proved, however, that K is not computable, so that no green-light-red-light machine can exist for K. The only way out of this dilemma is to conclude that there is no Hilbert machine. In other words, Hilbert's 10th problem is unsolvable!

The fact that a Diophantine equation is associated with every listable set is a positive result that is of great interest in itself, quite aside from its application to Hilbert's 10th problem. A particularly important and interesting set of integers is the set of prime numbers. A prime number is one that is factorable (divisible) only by 1 and by itself. Some examples are 2, 3, 5, 7, 11, 13 and 17. That they are listable is rather obvious. An algorithm for listing them has come down from the Greeks with the name of "the sieve of Eratosthenes." Combining Matyasevich' result with a device developed by Putnam, we obtain a Diophantine equation $Q(y_1,y_2,\ldots,y_n) = z$ such that a positive number z is a prime if and only if this equation has a positive integer solution $y_1,y_2,\ldots,y_n$. (The exact form of the polynomial Q is a bit too complicated to fully write out here.)

Another remarkable result can be proved by combining Matyasevich' theorem with Gödel's work on undecidability. If there is any system of axioms whatsoever from which information can be deduced about Diophantine equations, one can always obtain a particular Diophantine equation that has the following properties: (1) the equation has no positive integer solutions and (2) the fact that it has no positive integer solutions cannot be logically deduced from the given set of axioms. Of course, once the Diophantine equation is obtained we can make up a new set of axioms from which one can prove that the Diophantine equation has no solution. But then this new set of axioms will give rise to another Diophantine equation for which the same can be asserted.

What went into the proof of Matyasevich' theorem? In addition to the results from classical and even ancient number theory that we have already mentioned, there is a key result known as the Chinese remainder theorem. It will be helpful to illustrate the Chinese remainder theorem by a numerical example.

Suppose one wishes to find a number whose remainders, when divided by the numbers 10, 3, 7 and 11, are respectively 4, 2, 3 and 1 [*see bottom illustration on opposite page*]. The Chinese remainder theorem assures us that there must be such a number. (In fact, in this case 584 is such a number.) All that is required for the Chinese remainder theorem to work is that no pair of the divisors used have any common factor (except, of course, 1). There can be any number of divisors, and the desired remainders can be any positive integers whatsoever.

In 1931 Gödel showed how to use the Chinese remainder theorem as a coding trick, in which an arbitrary finite sequence of numbers can be encoded as a single number. From the code number one recovers the sequence in the same way that 4, 2, 3 and 1 are obtained from 584 in the example—as remainders in successive divisions. The divisors can be chosen to be in arithmetic progression.

The first attempt to prove that a Hilbert machine cannot exist was made by one of us (Davis) in his doctoral dissertation in 1950. Gödel's technique of using the Chinese remainder theorem as a coding device was applied to associate a Diophantine equation, $P_S(k,x,z,y_1,y_2,\ldots,y_n) = 0$, with every listable set S. Unfortunately the relation between the set and the equation turned out to be more complicated than what was needed for Hilbert's 10th problem. Specifically, the relation was: A positive integer x belongs to the set S if and only if for some positive integer value of z it is possible to find a solution for every one of the Diophantine equations obtained by substituting $k = 1$, then $k = 2$ and so on up to z into the equation $P_S(k,x,z,y_1,y_2,\ldots,y_n) = 0$. Although the result seemed tantalizingly close to what was needed, it was only a beginning.

At about the same time Robinson began her own investigations of sets that can be defined by Diophantine equations. She developed various ingenious techniques for dealing with equations whose solutions behaved like exponentials (grew like a power). In 1960 she, Davis and Putnam collaborated in proving another result. They made use of both her work and Davis' result to show that to any listable set there corresponded a Diophantine equation of an "extended" kind, extended in the sense that variables in the equation were allowed to occur as exponents. An example of such an equation is $2^t + x^2 = z^3$. Davis, Robinson and Putnam combined their work with some of Robinson's earlier results and discovered the following: If even one Diophantine equation could be found whose solutions behaved exponentially in an appropriate sense, then it would be possible to describe every listable set by a Diophantine equation. This would in turn show that Hilbert's 10th problem is unsolvable.

It took a decade to find a Diophantine equation whose solutions grow exponentially in the appropriate sense. In 1970 Matyasevich found such an equation by using what are known as the Fibonacci numbers. These celebrated numbers were discovered in A.D. 1202 by Leonardo of Pisa, who was also known as Fibonacci. He found them by computing the total number of pairs of descendants of one pair of rabbits if the original pair and each offspring pair reproduced itself once a month. The Fibonacci series is obtained by starting with 1 and 1 and successively adding the preceding two numbers to get the next: the first Fibonacci number is 1, the second is 1, the third is $1 + 1 = 2$, the fourth is $1 + 2 = 3$, the fifth is $2 + 3 = 5$ and so on. The property that is important for Hilbert's 10th problem is that the Fibonacci numbers grow exponentially. That is, the nth Fibonacci number is approximately proportional to the nth power of a certain fixed real number.

If one could find a Diophantine equation whose solutions relate n to the nth Fibonacci number, it would be the desired example of a Diophantine equation whose solutions behave exponentially. The solution of Hilbert's 10th problem would follow from this example. What Matyasevich did was to construct such a Diophantine equation [*see illustration above*]. Once he had shown that the set of Fibonacci numbers is associated in this way with a Diophantine equation, it followed immediately from the theorem of Davis, Robinson and Putnam that for every listable set there is an associated Diophantine equation, including in particular the set K, which is not computable. And so ends the story of Hilbert's 10th problem.

I. $u + w - v - 2 = 0$

II. $l - 2v - 2a - 1 = 0$

III. $l^2 - lz - z^2 - 1 = 0$

IV. $g - bl^2 = 0$

V. $g^2 - gh - h^2 - 1 = 0$

VI. $m - c(2h + g) - 3 = 0$

VII. $m - fl - 2 = 0$

VIII. $x^2 - mxy + y^2 - 1 = 0$

IX. $(d - 1)l + u - x - 1 = 0$

X. $x - v - (2h + g)(e - 1) = 0$

MATYASEVICH' SOLUTION to Hilbert's 10th problem involves a Diophantine equation that is obtained by squaring each of these 10 equations and then adding them together and setting the resulting complicated polynomial equal to zero. In these equations the values *u* and *v* in the solutions are related in such a way that *v* is the 2*u*th Fibonacci number. From the solution it followed that for every listable set there is an associated Diophantine equation. Since there exist listable sets whose complements are not listable, then not every listable set can have a green-light-red-light machine. Since having a green-light-red-light machine for a set is equivalent to having a Hilbert machine for Diophantine equations, Matyasevich' result means that no Hilbert machine can be built to test Diophantine equations.

THE FLYING LEAP OF THE FLEA

This flightless insect can jump 100 times its own length. It does so by the sudden release of energy stored in a rubberlike protein located at the site of the wing-hinge ligament in flying insects

by Miriam Rothschild, Y. Schlein, K. Parker, C. Neville and S. Sternberg

The saltatorial powers of the flea puzzled philosophers and poets even before the naturalist Pliny noted them "skipping so merrily in victualing houses." Not only is the jump a large one for the animal's size, but also the takeoff is bewilderingly sudden and too rapid for the human eye to follow. With the aid of high-speed motion-picture photography, combined with a careful study of the anatomy of the flea, we have been able to describe its movements while jumping and to put forward a hypothesis suggesting how the movements are executed.

The oriental rat flea *Xenopsylla cheopis* (Rothschild), which is a major vector of bubonic plague, measures one or two millimeters in length, weighs unfed some 210 micrograms and can jump to a height of 90 millimeters (3.6 inches). Good average performers consistently leap from 60 to 70 millimeters at a temperature of about 28 degrees Celsius (about 82 degrees Fahrenheit). Since the angle of takeoff is often nearly vertical, the horizontal distance covered by the flea is not so spectacular: about 180 millimeters for *X. cheopis* in our laboratory. Even so, the flea is jumping more than 100 times its own body length. Cat fleas and human fleas are more impressive performers, being capable of executing a standing leap 33 centimeters (a little more than a foot) high.

Unfed fleas can continue jumping about for long periods. The oriental rat flea, if it is stimulated by the presence of other fleas, will jump 600 times per hour for as long as 72 hours without a pause. On one occasion 10 fleas executed 10,000 jumps per hour, or one jump each every three and a half seconds. (The rate is measured by connecting a microphone and an amplifier to the jar in which the fleas are confined and recording the sound of each jump.)

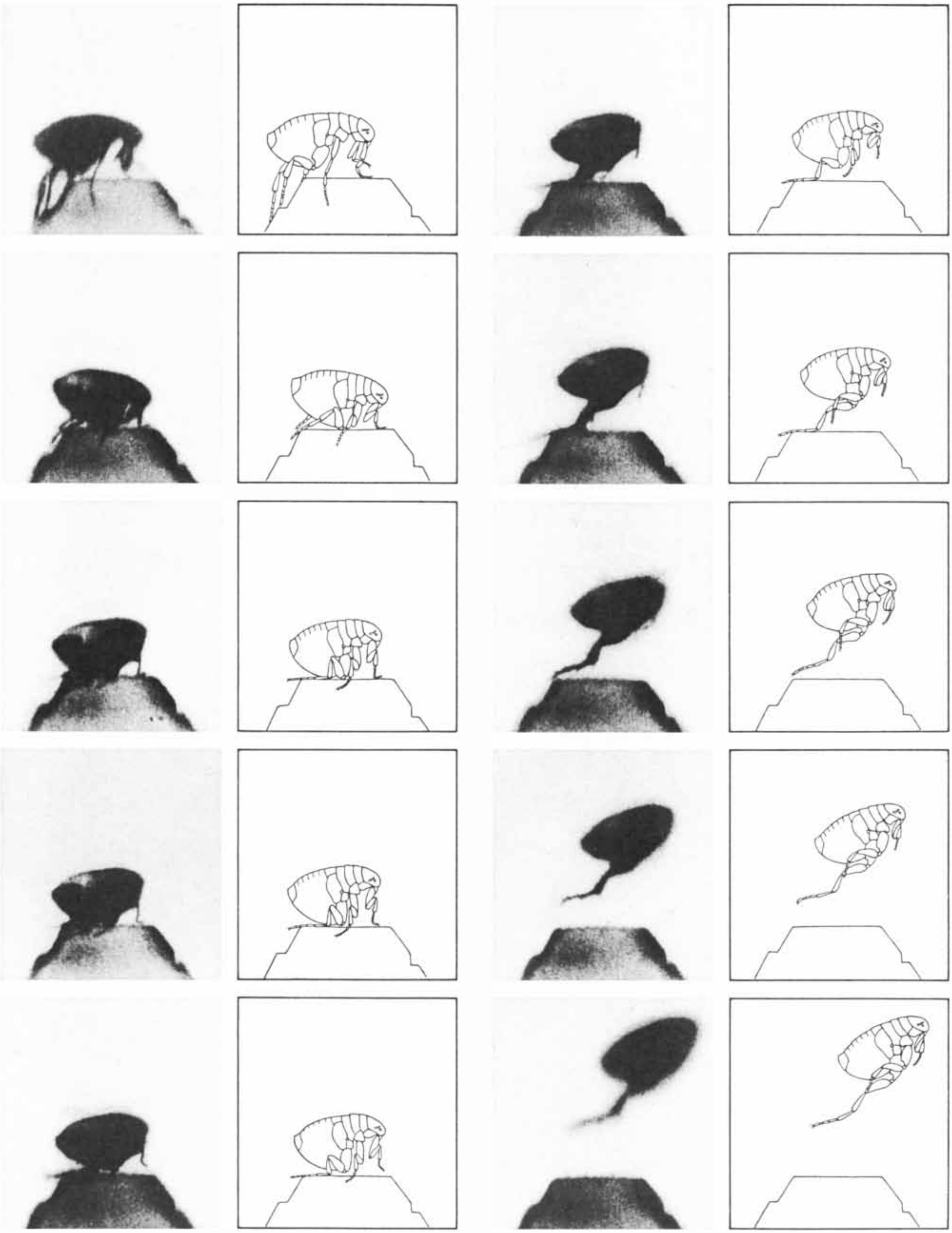
The early stages of a flea's life cycle are passed in the nest of its host or in debris on the ground. The crux of the flea's existence is making contact with the host after hatching from the pupa. Undoubtedly this hazardous operation, that is, the flying leap from the substratum to the passing animal, has set a premium on the evolution of the insect's highly specialized and efficient jumping mechanism. One may also note that the leap is an extremely effective means of escape.

Almost all groups of insects that have adopted the ectoparasitic way of life lack wings. It is evident that these structures present such a handicap to an animal living in fur or feathers that they are lost in the course of evolution. Although natural selection has disposed of the flea's wings, it has retained several important features of the flight mechanism found in many flying insects today, notably the wing-hinge ligament, the starter muscle (tergo-trochanteral depressor) of the leg and certain of the direct flight muscles (the subalars and basalars), all of which have been modified to produce the lightning takeoff.

A grasshopper, once it is in the air, has no control over its trajectory unless it can use its wings. It often turns somersaults in midair and even lands on its back. Photographs show that during the flea's descent its legs are spread out laterally, which no doubt assists in controlling its descent. It is remarkable that rat fleas land on their feet in 78 percent of their leaps if they come down on a reasonably flat surface. They often spin and somersault in midair, which is revealed not only in our films but also by the fact that 20 percent of the fleas jumping away from a strong source of light land exactly facing their point of departure and an additional 60 percent land at a right angle to it. Such fleas reorient themselves by shuffling around, and after a pause they jump again toward the shade. Females are always somewhat less successful than males in their directional landing performance, perhaps because their abdomen is larger and heavier.

Some years ago we photographed a jumping flea by means of a special camera that was triggered by the flea itself when it interrupted a narrow beam of light shined through a glass cell [see "Fleas," by Miriam Rothschild; SCIENTIFIC AMERICAN, December, 1965]. Although these photographs revealed certain interesting features of the jump in midair, we decided that only high-speed motion-picture photographs could show details of the takeoff. The British Broadcasting Corporation first undertook to produce such a film on our behalf, but although the speed of 1,000 frames per second at which the pictures were made (by E. C. A. Lucey) was fast enough to reveal the movements of the flea's individual limbs, the photographer was unable to overcome the difficulty of focusing accurately at short range on such a small and unpredictable performer. As a result few details could be seen in those early pictures, although calculations made from the film ruled out a single muscle contraction as the direct source of power for the jump.

The fastest single muscle twitches (in the locust, for example) take 15 milliseconds to reach peak force, and in developing an acceleration of 140 gravities a flea would require something over 10 times this speed. Therefore it was postulated by H. C. Bennet-Clark of the University of Edinburgh, and has been confirmed by us, that the flea's leap is powered not by muscle alone but is assisted by the elastic protein resilin. This protein, which is present in the wing-hinge



PLAGUE FLEA (*Xenopsylla cheopis*) climbs to the level top of an experimental jumping platform (*top left*) and begins to "collect" its legs, contract its body and crouch in preparation for a jump (*second from top*). Six successive exposures show jump, beginning (*third from top*) as a surge of energy drives the trochanters of the rear legs against the platform surface. After the recoil lifts the trochanters from the surface, rotation of the femurs keeps the tibiae thrusting at the surface (*fourth from top, top right and second from top*). The flea is airborne, its tibiae and tarsi clear of the surface, one millisecond after the jump starts (*third from top*) and follows a parabolic path (*fourth from top*) that will probably carry it some 90 millimeters into the air and 180 millimeters from the platform.

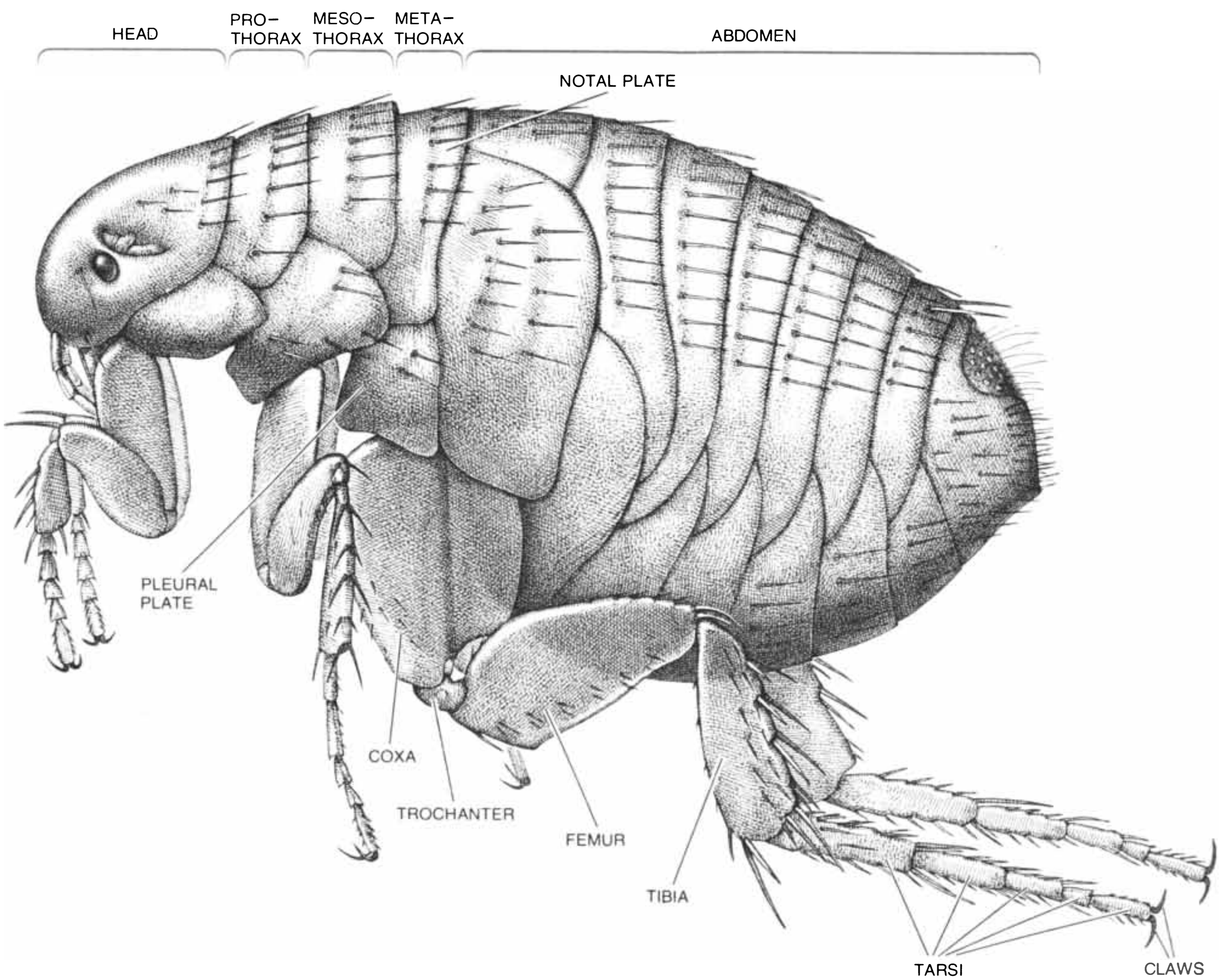
ligament of dragonflies and locusts, can store and release energy more efficiently than any known rubber and can deliver power faster than most actively contracting muscles.

Eventually the difficulty of obtaining high-magnification motion pictures was overcome by taking advantage of the flea's natural inclination to climb up small objects in its vicinity. A pyramid-shaped "hill" with a flat top was made of plastic and placed in a glass cell together with about 50 fleas. The camera was focused on the apex of the hill. The fleas obligingly climbed up the sloping sides and jumped from the top as they were jostled by the next in line. We obtained pictures with a relatively sharp image at a speed of 3,500 frames per second. A series of pictures made during the jump of a single flea shows the movements involved [*see illustration on preceding page*].

It was realized many years ago that the main thrust for the flea's leap comes from its powerful back legs, and it was also pointed out that the pleural arch—a small, transparent area located between the pleural and notal ridges of the thorax—was absent or greatly reduced in species of flea that are poor jumpers. Therefore the place to look for the energy store was the pleural arch. We found that this structure possesses various characteristics of dragonfly and locust resilin. First, it occupies the same position (at the top of the pleural ridge), suggesting that the structure is homologous with the wing-hinge ligament of scorpion flies, dragonflies and locusts. Second, it responds to various specific staining procedures in the same manner as resilin. Third, in the pleural arch of fleas it occurs as a region of pure resilin grading into an area where microfibrils of chitin are embedded in a resilin matrix. Finally, this protein occurs in the flea as a discrete ligament sandwiched between regions of solid cuticle and separated from them by characteristically sharp junctions. Hence even though the flea has no wings, it is possible to show that resilin is present in its pleural arch. Clearly the apparatus is derived from the flight system of some winged ancestor.

Resilin was first discovered in insect cuticle by Torkel Weis-Fogh of the University of Cambridge. It is a nearly perfect rubber, yielding 97 percent of stored energy when it is released after being stretched. There is thus a loss of only 3 percent of the applied kinetic energy, which is dissipated as heat. In most commercial rubbers the loss is 15 percent.

Resilin therefore appears to function as a true elastomer in the thermodynamic sense, with very little of the applied energy being wasted in straining covalent bonds. Instead the energy is



JUMP IS POWERED mainly by forces acting through the flea's large hind legs. The leg components are the coxa, the trochanter, the femur, the tibia, five tarsi and finally the claws. Only the hind legs are linked to a zone where energy for jumping is stored.

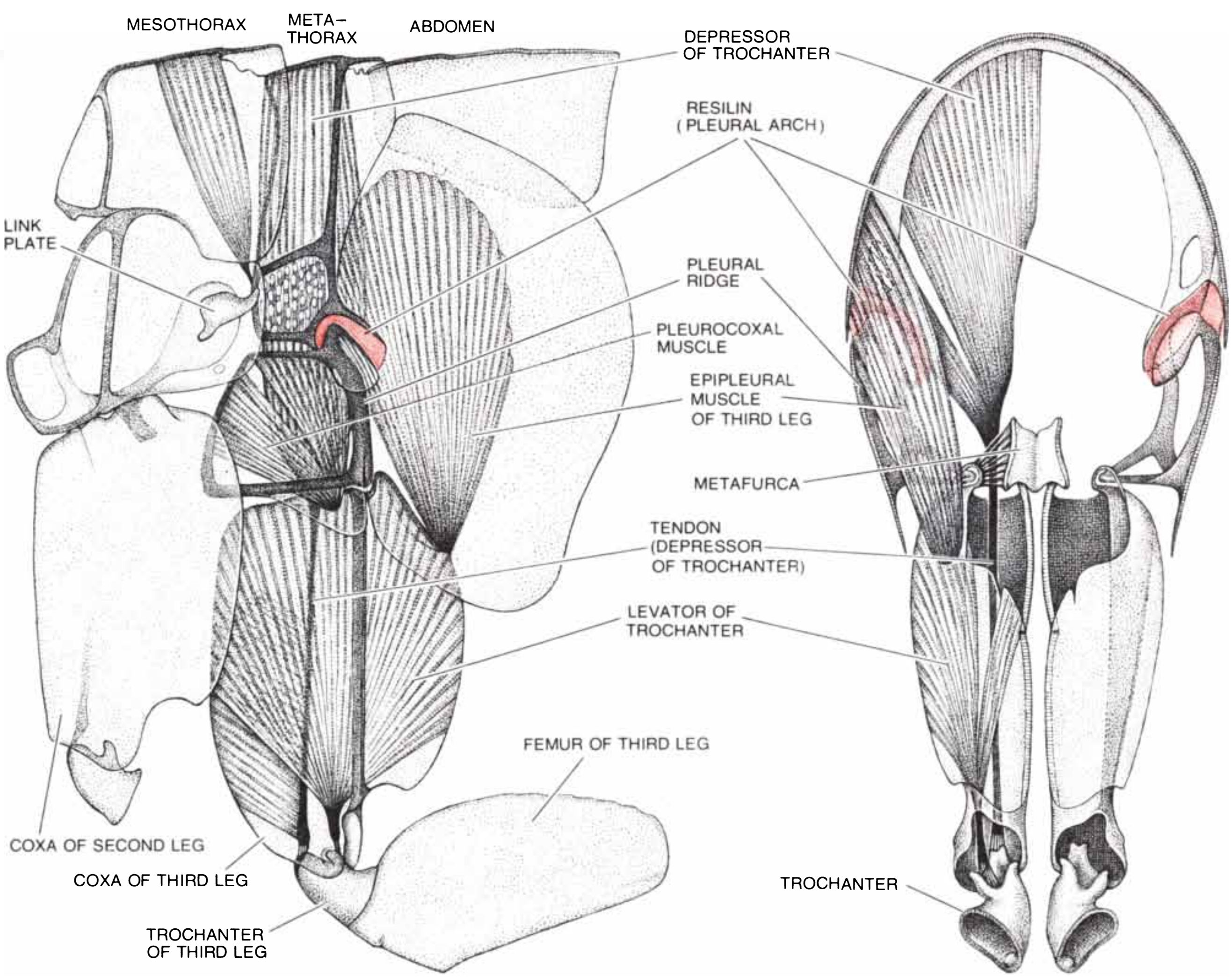
absorbed mainly in reducing the entropy of the system (that is, in bringing more order into it) by orienting randomly kinked polypeptide chains into a less probable conformation. When the chains are released, they oscillate rapidly back to a more random state. Resilin can be stretched to three times its equilibrium length for several months, and on release it recovers its former length within seconds, indicating a complete lack of flow in the molecular network.

By exploiting resilin fleas have overcome two of the major limitations of muscle: its relatively low rate of relaxation and contraction and its poor performance at low temperatures. (Warm-blooded animals deal with the second problem by constantly transforming energy to generate heat.) The fastest insect flight muscles are mechanogenic, with the muscle following the resonant mechanical frequency of its suspension system. Although some mechanogenic muscles attain contraction and relaxation times of less than half a millisecond, they cannot reach maximum amplitude in the first contraction, so that notwithstanding their high-speed rhythmical capability they are unsuitable for a single event as rapid as the flea's jump.

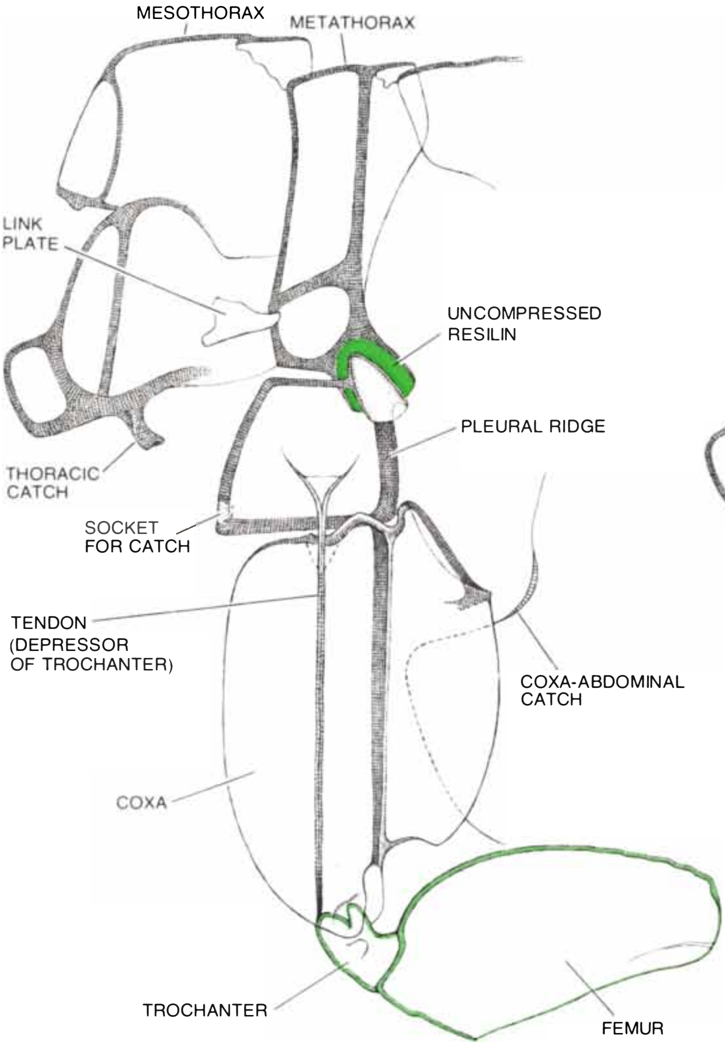
When energy is stored in elastic structures such as cuticle and the resilin of the pleural arch of fleas, however, even release by muscle action is fast enough—provided that a bistable "click" mechanism is present. (Such a mechanism is found in flying insects; the wing clicks suddenly from an upward to a downward position and vice versa when the applied force reaches a critical value.) It is important to realize that in addition resilin can store energy for prolonged periods with little loss of energy in the muscles compressing it. This economy is possible because the muscles involved are not then moving and because resilin shows no lasting deformation under prolonged strain. The flea can therefore remain "cocked" for some time before its sudden takeoff, and it can also sustain prolonged periods of jumping.

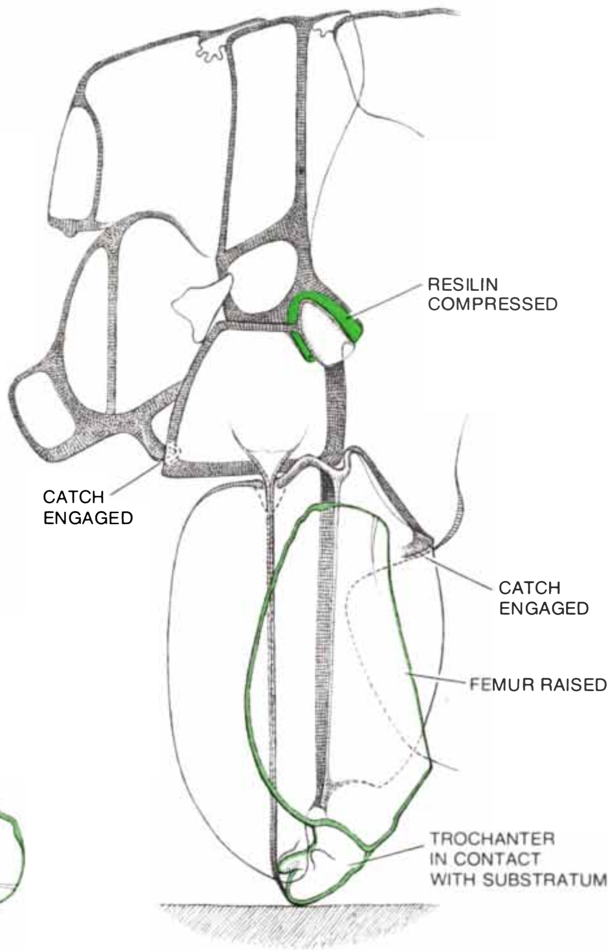
A reasonable hypothesis, if nothing more, can be put forward to explain the flea's jumping mechanism based on a careful study of the insect's exoskeleton (hard external covering) and musculature. One can obtain a fair idea of the remarkable arrangement revealed by this study if one imagines a man with a perfectly elastic rubber ball incorporated in his hip joint bouncing himself upward from a kneeling position.

The pleural arch of the flea is part of the cuticular exoskeleton, which forms a system of hardened struts and ridges strengthening the outer walls of the insect's thorax and the legs [*see illustration below*]. The epipleural muscles (direct flight muscles, each consisting of some 60 to 80 fibers) help to compress



PRINCIPAL MUSCLES of a flea's hind leg are seen in sagittal section (*left*) and transverse section (*right*), extending from the thorax to the trochanter. An area in the thorax, the pleural arch, contains the resilin (*color*), the rubberlike protein that powers the jump.





"UNCOCKED" LEG is seen in sagittal section with all its muscles removed except the tendon of the depressor of the trochanter. The resilin of the pleural arch is uncompressed and the flea's weight is being supported by the tarsal segments of the legs (*not shown*). Both the thoracic catch, which will link mesothorax and metathorax (*left*), and the coxa-abdominal catch, which is shown only schematically (*right*), are unengaged and the depressor tendon is relaxed. Dark areas are stiffened cuticle, not compartment walls.

"COCKED" LEG has its femur raised nearly upright so that the leg's contact with the ground is through the trochanter and tibia. Both the thoracic and the coxa-abdominal catches are engaged and the resilin in the pleural arch is compressed. The flea is ready to jump. Release of the catches lets the expanding resilin drive the trochanter forcibly against the ground and gives the flea its initial acceleration. The descending femur adds to the acceleration by thrusting the tibia against the ground for nearly a millisecond.

the resilin; some of them are attached to the thoracic ridges and pull them downward, others are attached to the arch itself and still others are inserted along the pleural-ridge fold, pulling the pleural ridge down into the rim of the coxa. The trochanteral depressor muscle (consisting of some 70 to 80 fibers in the thorax) originates on the notum, and its massive tendon is inserted in the heel of the trochanter. This tendon attains its maximum diameter in good jumpers such as the cat flea. It lies nearer the center of the flea's body than the epipleural muscles and is the principal muscle that depresses the tergum, arches the thoracic and coxal walls and also helps to compress the resilin.

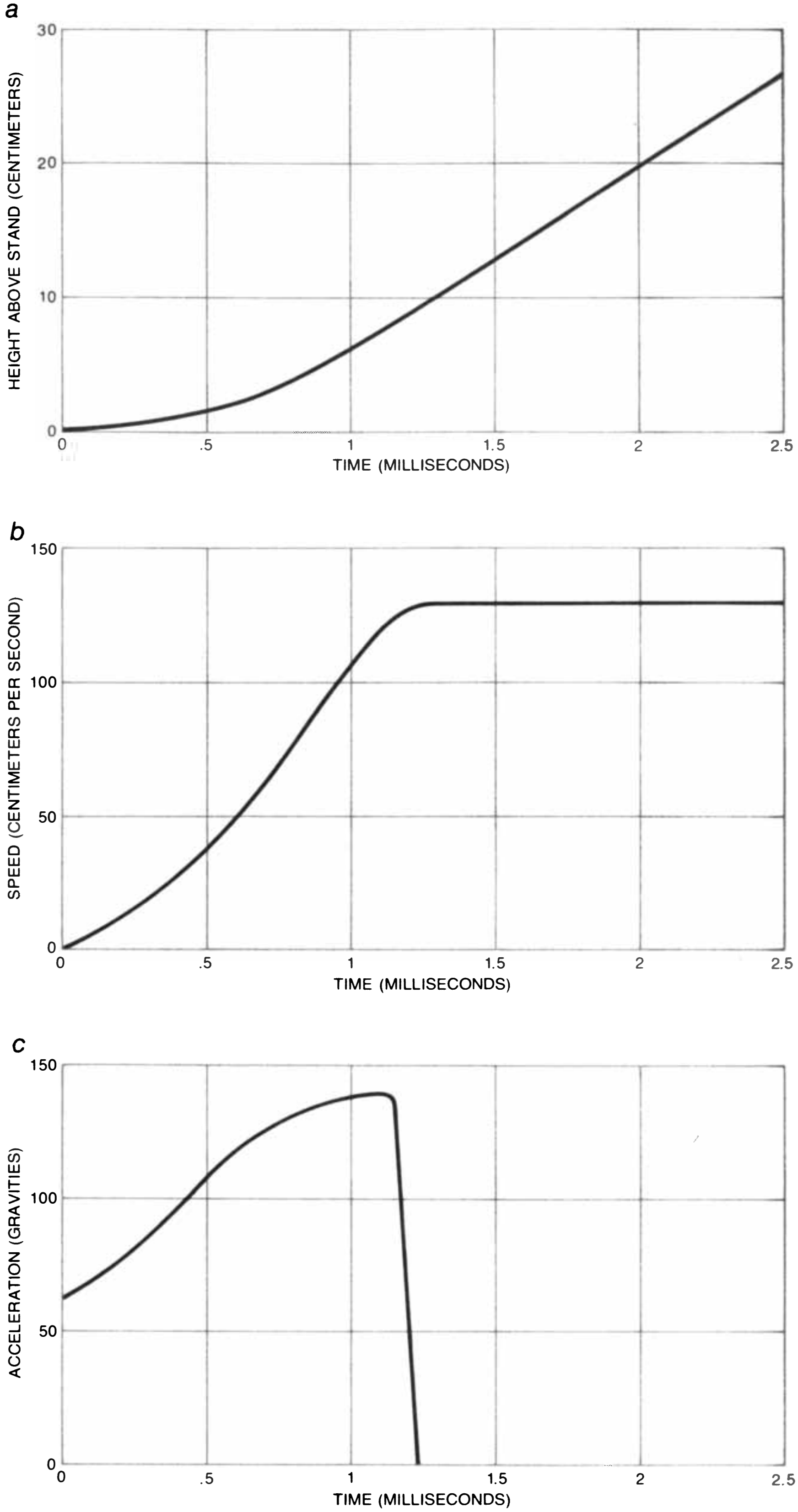
It is important that the tendon of the trochanteral depressor remain rigid and not be pulled out of position when the femur of the flea is raised and the tendon is rotated into the "heel" of the trochanter for takeoff. In all fleas there are devices that prevent any such movement. Where the tendon leaves the thorax and enters the coxa it runs through a funnel-shaped cuticular slot; horizontal muscles originating near the far end of the tendon become taut like tent ropes when the femur is raised and hold it firmly toward the front of the leg, while a knob of untanned cuticle (part of the trochanter) with a central depression clamps like a limpet against the back of the tendon, where it prevents overcentering and also acts as a shock absorber. Strong membranes are attached to the end of the pleural ridge near its articulation with the coxa. These ligaments probably provide a mechanism to help absorb the recoil forces that would result from the release of elastic energy. The muscles of the locust, for example, are so strong (they have about half the tensile strength of steel) that at maximum tension they can snap the tendons. Thus recoil mechanisms in powerful jumpers can be expected.

When the flea is about to jump, it crouches down, lowering its head and contracting its body. Some 30 years ago Helmut Jacobson of the University of Erlangen noted that fleas "collect" their legs before takeoff; this procedure, he remarked, took considerably longer than the jump itself. During the preparatory stage the femurs are raised by the contraction of the levator muscles, which also rotate the trochanteral hooks into their sockets in the distal end of the coxa, thus bringing the trochanters and the tibiae into contact with the substratum [*see illustration at right on opposite page*]. The force is then loaded into the pleural arch by contraction of the epipleural and trochanteral depressor muscles, which squeeze the resilin sandwiched between the pleural ridge and the sclerotized ridges of the thorax. The gap between the pleural ridge and the dorsal rim of the coxa is also closed by this action, thus making the thick ridges of the cuticle into a continuous line of force. The shortening of the body results from the contraction of the oblique dorsal and ventral longitudinal muscles, which in turn are responsible for the upward rotation of the link plates. In this way the three thoracic segments are clamped together and two of the principal cuticular catches concerned in the jump are engaged.

The most important catch is a single median hook securing the sternum of the mesothorax to the metathorax; a less powerful pair of catches, laterally situated, clamp the coxae to the abdomen. Two ventral transverse muscles in the third segment brace the tendon of the depressor against the metafurca. The epipleural muscles (and apparently the trochanteral depressor muscle as well) can now relax, but the resilin is still held compressed by the cuticular catches, which thus maintain the line of force. The flea in the crouched position is "cocked" for takeoff.

The jump is initiated by relaxation of the levator muscle, which allows the femur to descend. Simultaneously the energy stored in the resilin and the arched pleural and coxal walls is freed when the two cuticular catches are released by relaxation of the ventral longitudinal muscles. As the thoracic walls straighten, the trochanter is pressed downward against the substratum, with the force generated by the expansion of the resilin traveling along the lines of force and providing the initial acceleration. Additional acceleration is obtained from the force exerted through the tibia by the descending femur.

When the trochanter leaves the ground, its far end continues to move downward with respect to the notum, while the proximal end is restrained by the tendon of the depressor muscle. Since the femur forms a mechanical (but not rigid) unit with the trochanter, this movement is responsible for the descent



FIRST TWO MILLISECONDS of a flea's jump were analyzed with respect to position (*a*), velocity (*b*) and acceleration (*c*) by studying motion pictures made at a rate of 3,500 frames per second. The sudden expansion of the resilin compressed in the pleural arch gives the flea an instant acceleration rate of more than 50 *g* (50 times the acceleration of a body falling under the influence of the earth's gravity). Thereafter added acceleration comes from the thrust of the femur. The peak, 140 *g*, is reached in little more than a millisecond. By this time the flea has risen about .07 centimeter above the jumping stand and has reached a peak velocity of about three miles per hour. In the next millisecond it triples its altitude.

of the femur. There are various possible ways in which the trochanteral depressor (muscle and tendon) could supply this counterpoint for the trochanter-femur lever. The bracing of the tendon, within its socket, against the metafurca could be sufficient, or the viscous property of the muscle may passively resist the sudden shock.

As in insect flight systems, elasticity in both resilin and ordinary cuticle ensures very rapid movement, and it also simplifies coordination. Synchrony of fast-moving parts is achieved with greater effect by the release of mechanical click mechanisms, particularly the centrally located single control catch, than by coordination through nerves. Since it is impossible to prove how fleas jump, because insects of such tiny proportions do not lend themselves to experimental neurophysiology, we have to content ourselves with a hypothesis based on a morphological study.

The middle leg of the flea lacks both resilin and line of force, but the trochanteral depressor is present, and apart from acting as a support this leg often plays some role in the jump. Results of experiments involving amputation must be accepted with reserve, yet the removal of both hind legs of the rat flea proved beyond doubt that the flea was still capable of feeble jumping, using its middle and front legs only. It is thus jumping without the help of resilin, and in such circumstances it may well take off from its feet in a more conventional manner.

As one might expect, removal of the tarsi and claws from the back leg only had little effect on jumping performance. Removal of the tarsi of the middle leg seemed to reduce the number of jumps per half-hour slightly more than removal of the hind tarsi. These experiments also proved that the middle leg plays some part in the jump, probably a more or less important part according to circumstances and the species of flea.

Fleas are adapted to varied environments and can survive in deserts or within the Antarctic Circle, where, for as long as nine months of the year, they remain frozen under ice and snow. The oriental rat flea is adapted to hot climates. Temperature has a marked effect on the frequency of its jumping. The total number of jumps by 10 fleas confined in our testing apparatus rose from 3,976 per hour to 8,848 when the temperature was increased from 20 to 30 degrees C. (68 to 86 degrees F.).

The European rabbit flea becomes lethargic at high temperatures. On the other hand, this species can be frozen at about zero degrees C. for periods of up to nine months in a domestic refrigerator, and on being returned to room temperature the flea begins to jump around in a matter of minutes. Bird fleas have been observed leaping about on the snow in the Alps, and temperatures as low as six degrees C. seem to have little effect on their jumping ability. This phenomenon is characteristic of the release of energy stored in elastic structures—an entirely physical process that, unlike the chemically controlled release of energy in muscle contraction, is relatively independent of temperature.

The importance of elasticity in arthropod locomotion was first emphasized by Weis-Fogh following his discovery of resilin and the demonstration of its contribution to the energy balance in insect flight. This work greatly stimulated research on proteins that can act as rubbers in the thermodynamic sense. Two proteins besides resilin are known that behave as a rubber when they are swollen with water. They are the vertebrate counterpart, elastin, which provides the elasticity of arterial walls and the vocal cords, and the molluscan equivalent, abductin, which acts as an antagonist to the muscle closing the shell valves. There is also some evidence for an elastic energy-storing mechanism in the spring action of the foot in the contemporary horse.

Elasticity simplifies nervous coordination. In flying insects automation is found in a number of systems once flight begins. Examples include increased ventilation by elastic movements of the thorax, the fuel circulation in the blood bathing the flight muscles, elastic energy storage and release, active heating of flight muscles and the establishment of a resonant frequency in mechanogenic flight systems. A slow muscle in the dragonfly acts through an elastic tendon of resilin to provide wing-twisting when the insect flies backward. Because of the elastic tendon the wing twists automatically at just the right moment of each wing stroke. Similarly, in the jump of the flea a triggered click mechanism and stored elastic energy eliminate the neuromuscular complications that are normally associated with "startle reactions" in other animals.

The morphology of both adult and larval fleas suggests that their winged ancestors resembled scorpion flies or moss flies. Because of their ectoparasitic life-style, fleas have lost their wings and have returned to a more primitive form of locomotion. Nevertheless, they have retained some of the more spectac-

ular flight adaptions and now utilize them for perfecting their leap.

In order to slide easily through the host's hair they have also become flattened from side to side; thus the wing-hinge ligament, originally situated on the dorsum, has become displaced laterally, making its subsequent incorporation into the jumping mechanism possible. The shortening of the pleural ridge associated with the lateral position of the pleural arch is probably the most important modification involved in making the change feasible. The leg and thoracic muscles, which were once molded to produce wing power, are now further adapted to load energy into the resilin cap. No force is dissipated along the wings, since they have been suppressed, and the elastic energy is turned back, by way of the lines of force and the starter muscles, into the jump. The fact that resilin functions independently of temperature is an added advantage for the flea, which has to hop from the cold external environment to the warm-blooded host and then leave it again for mating and egg-laying in the nest.

Certain species of fleas have secondarily lost the resilin in the pleural arch and hence their flying leap, but some of them are still capable of feeble jumping; others are reduced to walking and crawling. One group of mole fleas, which live in underground nests, have lost the pleural ridge as well as the arch. Various bat fleas, which hatch in guano on the floors of caves and crawl up the walls to reach their roosting hosts, often have reduced arches or lack them altogether. Fleas parasitizing arboreal rodents or birds such as swallows, which build high aerial nests, tend to lose the resilin in their pleural arch. It will be readily appreciated that speculative jumpers in those situations would be rapidly eliminated from the breeding population. On the other hand, fleas that are associated with large hosts, particularly those without well-circumscribed nests or lairs, such as deer or sheep, the ubiquitous cat or man himself, have the largest pleural arch. Broadly speaking, the larger the host, the higher the jump and the larger the cap of resilin—a relation that serves to emphasize the importance of the flea's initial jump onto a passing animal.

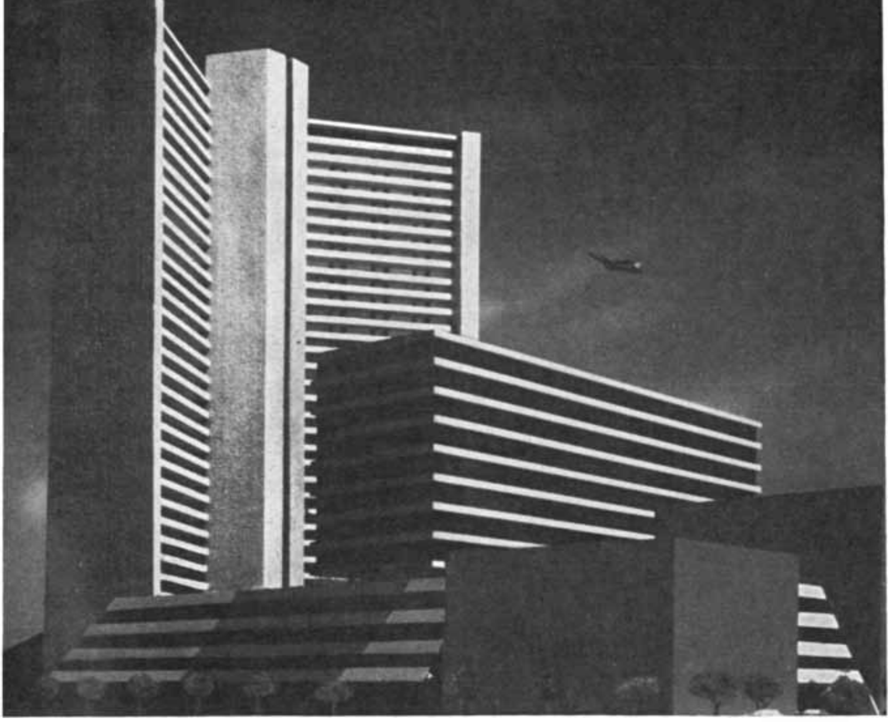
The flea appears to be the only insect so far described that "flies" with its legs or bounces itself upward off its trochanters. We found that the moss fly *Boreus* has resilin at the base of its degenerating wing stumps, and we suspect that the resilin has some function connected with the fly's habit of jumping about in the snow. *Boreus* may be following an evolutionary course vaguely reminiscent of the flea's.

Another insect, the click beetle (*Melanotus communis*), has exploited the same combination as the flea, namely resilin as an energy store combined with a click mechanism. When the click beetle jumps, it lies on its back, arches its body and suddenly flattens out, thus jackknifing itself into the air. The jump is nearly vertical and reaches a height of 300 millimeters—the beetle somersaulting several times after takeoff if the surface is at all uneven. Moreover, the insect develops an acceleration of 380 gravities.

It is not clear what produces the characteristic sound from which the click beetle gets its name, but it is thought to be caused when the peg on the prosternum snaps into the cavity or socket on the next segment, the mesosternum, and not, as is popularly supposed, when the beetle bangs its head on the ground.

The jump of the flea and the jump of the click beetle therefore have several features in common. Both insects have a resilin energy store; both can hold the "cocked" position for an appreciable time, and in both insects muscle relaxation releases cuticular catches, thereby releasing in turn the energy stored in the resilin and the bent cuticular walls of the thorax. The legs of the click beetle, however, play no part in the takeoff and apparently little, if any, in the landing. Whereas the flea usually lands on its feet, it appears to be a mere matter of chance whether *Melanotus* pitches on its feet or its back. The beetle always comes down more or less at its point of departure, and the flea, owing to the straightening of its femur and tibia after the takeoff, lands some distance away from its starting point. The beetle's jump is apparently an escape mechanism pure and simple; the flea's flying leap is locomotory, and although it may be highly speculative, it is often directed. Thus the flea may jump toward a shadow or away from the light.

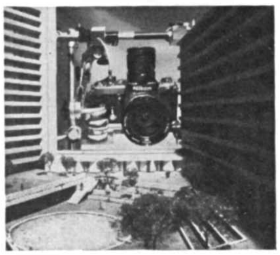
The acceleration developed by the beetle seems to be considerably greater than that measured for the rat flea, although possibly that of the cat flea and the deer flea will eventually prove to be more impressive. The uniqueness of the flea's jump lies not in its speed, acceleration, trajectory or sustained repetition, however, but rather in the subtle artistry by which natural selection has parsimoniously turned back the flight mechanism and incorporated it into the new jumping mechanism.



chitects: Langdon & Wilson. A I A.

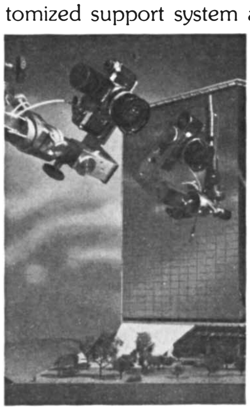
This building is 2 feet high

It's a model, of course. But, a model so well detailed and so superbly photographed that you can almost hear the hum of the elevators. If you were a prospective tenant, you could readily visualize its effect on the prestige of your business.



In achieving this exceptional realism, architectural photographer Jack Laxer has demonstrated for us not only his expertise but also his innovative and highly imaginative approach to this demanding application. When he recently acquired his new Nikon F2 cameras, he went to extraordinary lengths in exploiting their capabilities.

He designed ingenious support systems that allow totally flexible camera movement in all three directions and provides precise adjustments for about 800 different camera positions. It was a sizable investment in time, effort and money—but the unique combination of cameras, customized support system and talent has already paid off. The architects and builders, who are Laxer's clients, now can count on greater realism in photography for their promotional brochures. In fact, his Nikon photos are so accurate and so detailed that they enable clients to make design changes before the actual construction begins.



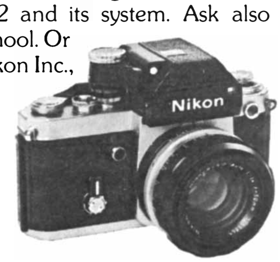
Apart from its legendary precision and reliability, the Nikon F2 offers special advantages for this type of work. Some are built in, such as the viewing system, the most accurate among 35mm cameras, and the multi-exposure provision that lets you add a background or put clouds in the sky.

Others are available in the Nikon System: The perspective-correcting PC-Nikkor lens, actually designed for architectural work. The 6X Magnifying Finder providing, as Laxer puts it, "an image virtually equal to 6x9 inches." The unique Nikon Bellow Attachment IV whose front standard movements permit corrective closeup photography. And, two of the nineteen interchangeable finder screens have grid patterns that simplify alignment of vertical and horizontal shapes in critical compositions.

This kind of precision photography has always been considered the exclusive domain of large, heavy studio view cameras. Yet, here it is done professionally with a 35mm camera that is best known for its speed and agility in handling split-second news, sports and similar action.

But, that is what Nikon versatility is all about. To help you make superior photographs of whatever interests you, whether a model just inches from the camera or a solar eclipse across millions of miles of space. To inspire you to look for new visions, new ways to picture your world and to test the capabilities of a camera whose limits have yet to be reached. And, to help you put it all together, there's the Nikon School of Photography, traveling to all parts of the U.S.A.

Your Nikon dealer will be glad to demonstrate the Nikon F2 and its system. Ask also about the Nikon School. Or write for folio 19, Nikon Inc., Garden City, New York 11530. Subsidiary/Ehrenreich Photo-Optical Industries, Inc. EPOI (In Canada: Anglo-Photo, Ltd., P.Q.)



The Evolution of the Pacific

Deep-sea drilling shows that the bottom of the western Pacific basin is different from the bottom of the eastern basin. The slow movement of the crust underlying the basin seems to account for the difference

by Bruce C. Heezen and Ian D. MacGregor

Advances in human knowledge often result from the coupling of two separate areas of human activity. An outstanding recent example of this phenomenon is the union of deep-water drilling technology and the scientific theories of continental drift, sea-floor spreading, plate tectonics and the youthfulness of the oceans. In 1968 the National Science Foundation inaugurated the Deep Sea Drilling Project, in which the ship *Glomar Challenger* began to systematically drill holes in the sediments that record the oceans' dynamic history. Among the oceans whose history has begun to emerge from this sampling program, coupled with the theoretical background, is the largest of all: the Pacific.

The *Glomar Challenger* is an unlikely hybrid: a ship with a hole in the middle and a 180-foot drilling rig amidships. With the ship's hull for a platform the men who operate her utilize the techniques of oil-well drilling to bring up the long cylindrical cores of deep-sea sediment. Holding a fixed position over a point on the sea bottom, compensating for the ship's vertical movements and guiding a fresh drill string into a hole only a foot in diameter several thousand feet below the ship's keel are only a few of the more formidable hurdles that the drillers have surmounted. The cores of sediment gathered by the *Glomar Challenger* from around the world constitute one volume after another, so to speak, in the growing library of oceanic history.

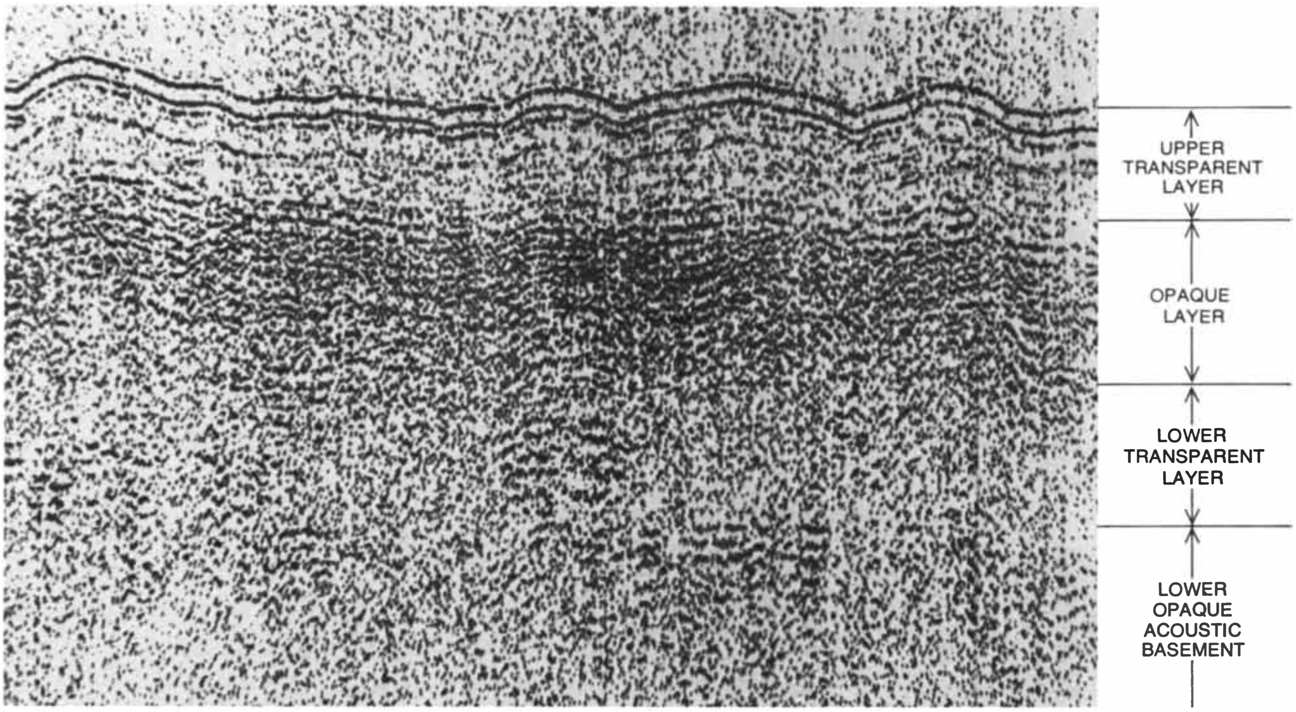
The deposition of sediments on the ocean bottom is a variable process. The continual rain of detritus, both organic and inorganic, that slowly sinks to the sea floor is not of the same composition everywhere, nor is it evenly distributed. Consider first the organic component. It is made up of fine fragments of the shells and skeletons and other parts of many different marine animal and plant species. The vast majority of these organisms are small or microscopic: planktonic protozoans and algae such as radiolarians, diatoms, foraminifera and coccoliths that inhabit the upper 400 meters of the ocean. The organisms are cosmopolitan in distribution, but they are found in the greatest numbers only in certain waters. For example, they are abundant in the nutrient-rich zones of deep oceanic upwelling along the western shores of continents. A similar but less well-known zone of abundance exists along the Equator, where the currents of the Northern and Southern hemispheres interact to produce another upwelling of nutrients. One planktonic group, the diatoms, is particularly numerous in two other zones of upwelling: the cold waters of the Arctic and the Antarctic.

Whether or not the calcareous and siliceous remains of these organisms ever actually reach the ocean floor depends on their solubility in seawater. Normally seawater is far from being saturated with calcium carbonate or silica. Because solubility increases with depth, almost all calcareous remains are completely dissolved before they can sink much deeper than 3,700 meters (a little more than two miles). This point is known as the carbonate compensation depth. The point where silica is completely dissolved is somewhat deeper. Compensation depths are, however, kinetic boundaries, and their location depends on both the rate of supply of organic detritus and its rate of dissolution. For example, in areas of high planktonic productivity such as the equatorial zone the carbonate compensation depth may be as much as 5,000 meters below the ocean surface (a little less than three miles). The accumulation of organic detritus on the ocean floor therefore varies from region to region, depending on the ocean depth and the planktonic productivity in the surface waters. As a result in a static model of the Pacific basin the accumulation of organic sediment on the ocean floor would be predictably greatest where the planktonic productivity was highest and the floor itself was shallower than the carbonate compensation level.

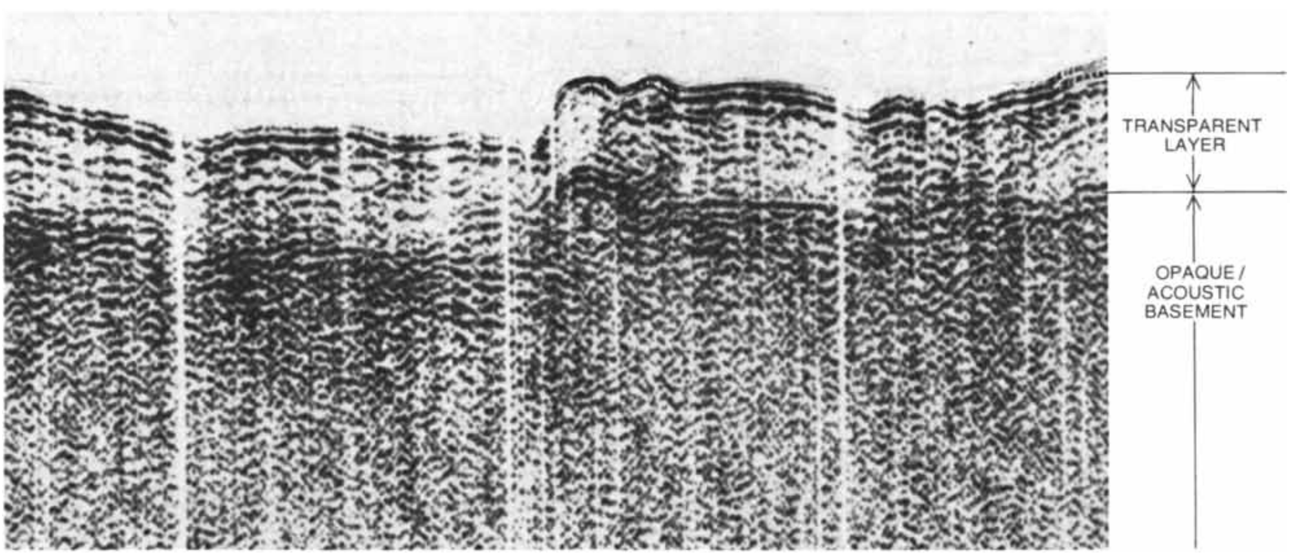
What about the sea-floor sediments formed from inorganic detritus? These substances move seaward along a variety of pathways before reaching the bottom. Essentially all inorganic detritus is derived either from continental processes of erosion or from volcanic activity along the margins of the continents. The sediments carried in suspension by streams and rivers make up one class of inorganic detritus. Some of the very smallest particles remain in suspension for a long time and become quite evenly distributed throughout an ocean basin before sinking to the bottom. The same is true of the fine dust that is carried far out to sea by the winds. These two deposition processes are responsible for a thin veneer of red, gray and green abyssal clays that covers much of the floor of all ocean basins. The clays are recognized as distinct sedimentary deposits, however, only where they have not been mingled with other sediments. For example, in the equatorial regions of the Pacific the detritus that elsewhere makes up the abyssal clays is only a trace dilutant, inseparable from the organic oozes that accumulate there in large quantities.

Coarser inorganic sediments reach the ocean in river runoff or as products of shore erosion. They do not usually travel far out to sea before sinking, and so they form a thick wedge of sand, silt and clay adjacent to their continental source. Some may be moved farther offshore by the action of turbidity currents (swift, self-propelled downslope flows of sediment-laden water) but their connection with the mainland remains obvious. Still another characteristic inorganic sediment of the Pacific basin is windblown ash from the active volcanic zones along the margins of the basin. The ash collects in the lee of the volcanoes, forming wedge-shaped deposits that usually thin out to the point of being no longer identifiable 1,000 kilometers or so from their source.

Each sediment, organic or inorganic, accumulates at a different rate. In regions of average planktonic activity the passing of a million years produces a layer of organic detritus 10 meters thick. In zones of greater activity, such as the equatorial region, the rate rises to 15 meters or more per million years. The abyssal clays are laid down at a much lower rate: about one meter per million years. For a few hundred kilometers downwind from an active volcano the average accumulation rate for volcanic ash is about 10 meters per million years. Sediments resulting from river runoff and shore erosion build up the most rapidly: from 50 to 500 meters per million years. (Even this process is not particularly fast; accumulation at the rate of 500 meters per million years means that



ACOUSTICAL PROFILE of the bottom of the Pacific in an area east of the Japan Trench shows four sedimentary strata, the number characteristic of the western Pacific. A "transparent" layer (*top*) overlies a darker "opaque" layer of sediments. Below the opaque layer lies a less well-defined transparent layer and below that is another opaque layer resting on the basalt oceanic crust.



DIFFERENT PROFILE characterizes the eastern Pacific basin. Here only two layers of sediment cover the oceanic crust: a transparent top layer and an underlying opaque one. Deep-sea drill cores confirm the acoustical data. The two profiles are not reproduced at the same scale. In the top one the top three layers are 2,000 feet deep; in the bottom one the top layer is 200 feet deep.

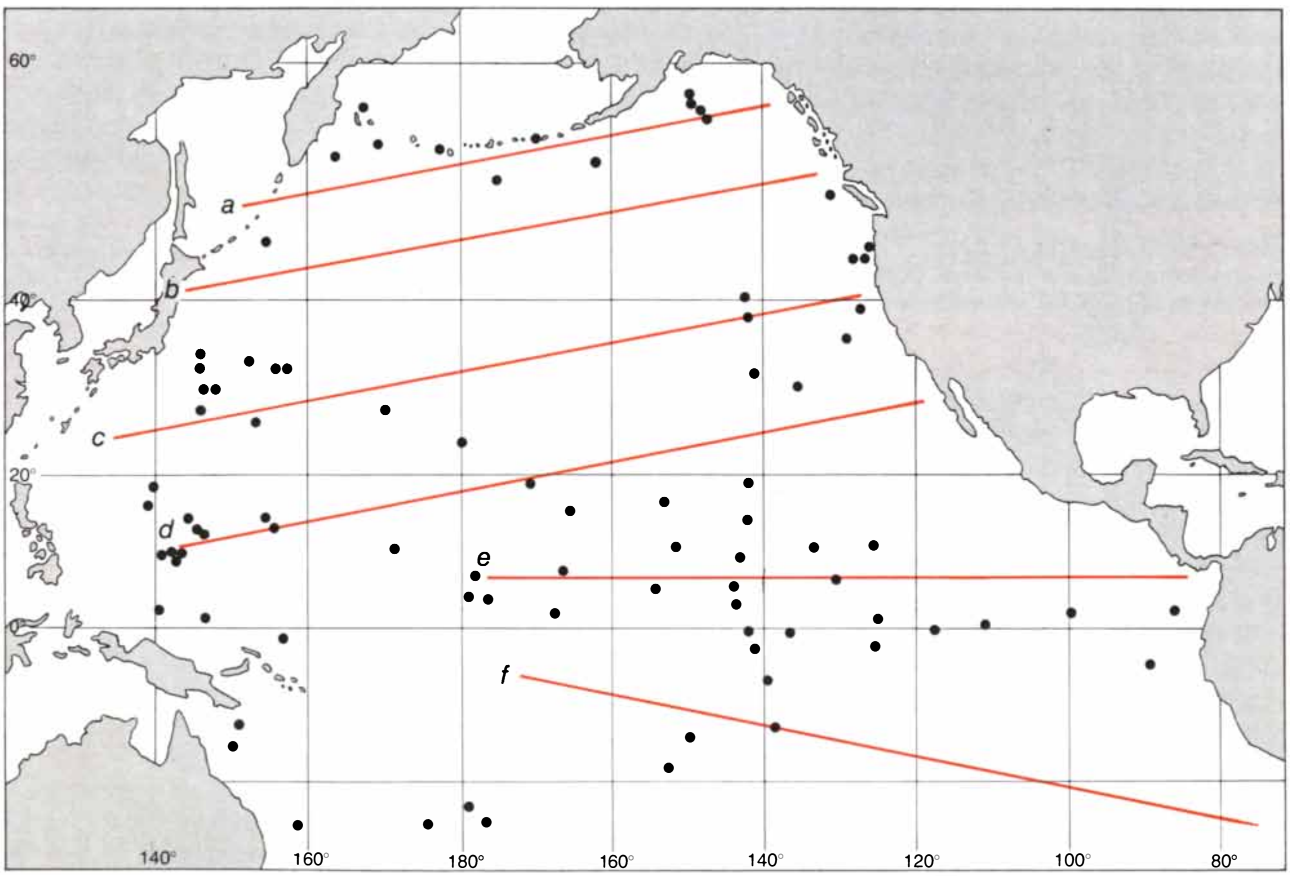
only a twentieth of a centimeter is deposited each year.)

Sediments of these kinds are accumulating today on the floor of the Pacific basin in a clearly patterned fashion [*see illustrations on these two pages*]. In the northern waters and all along the margins of the basin volcanic ash and sediments from runoff and shore erosion build up, mixed in the subpolar region with the tiny skeletons of the diatoms that thrive there. In a broad zone along the Equator the steady rain of planktonic detritus gives rise to a blanket of organic oozes on the sea floor. Between the largely inorganic northern zone and the organic equatorial zone, below waters that are too deep to allow the buildup of organic debris, lies a vast expanse of abyssal clays.

If the present pattern of sedimentation has been constant over the past 100 million years or so, a static model of the Pacific basin would predict that any core of sea-floor sediment taken from one or another of these major zones should consist uniformly of, say, abyssal clays or organic oozes, or sands, silts and ash. In actuality no drill core is that uniform. We are left to conclude either that the pattern of sedimentation has not been constant or that a static ocean-basin model is wrong. Suppose we assume that any changes in the sedimentation pattern over the past 100 million years have been trivial. Can we, by changing the static model into a dynamic one, explain the observed nonuniform nature of the sedimentary deposits in the Pacific basin? We believe the answer is yes.

Current plate-tectonics theory states that the basement rock of the ocean floor is being formed continuously. In the Pacific basin the area of formation is the "rift valley" at the summit of the East Pacific Ridge. The new crust migrates essentially westward, sinking deeper as it creeps along. The generation of new crust at the ridge is presumably balanced by a comparable consumption of old crust as it plunges under the island arcs of the western Pacific. The horizontal and vertical motions of the crust are coupled; the rate of sinking decreases logarithmically with age.

Let us visualize a dynamic model of part of the basin, centered on the East Pacific Ridge. The movement of the newly formed crust will be westward and downward [*see top illustration on page 106*]. From the time that the crust is formed at the top of the ridge, and for a considerable period thereafter, it will be above the carbonate compensation depth. As a result it will acquire a blanket of organic oozes. The sediments will increase in thickness until the outward and downward motion of the crust carries it below the compensation depth. Thereafter, as the crust continues to creep westward into deeper water, a slowly accumulating blanket of inorganic abyssal clays will begin to cover the initial organic deposit. Because the carbonate compensation depth lies 1,000 meters below the ridge crest and because the subsidence rate of the new crust is about 50 meters per million years, organic detritus will accumulate on the new crust over a period of 20 million years. Thus the basal layer of organic sediments, accumulating at a rate of 10 meters per million years, should under



PACIFIC SEDIMENTS WERE SAMPLED by deep-sea cores taken at the sites marked with the black dots on this map. The colored lines labeled *a*, *b*, *c*, *d*, *e* and *f* relate the cores to the cross sections of the sediments shown on the opposite page. The cores were obtained by drilling from the *Glomar Challenger* in the Deep Sea Drilling Project sponsored by the National Science Foundation.

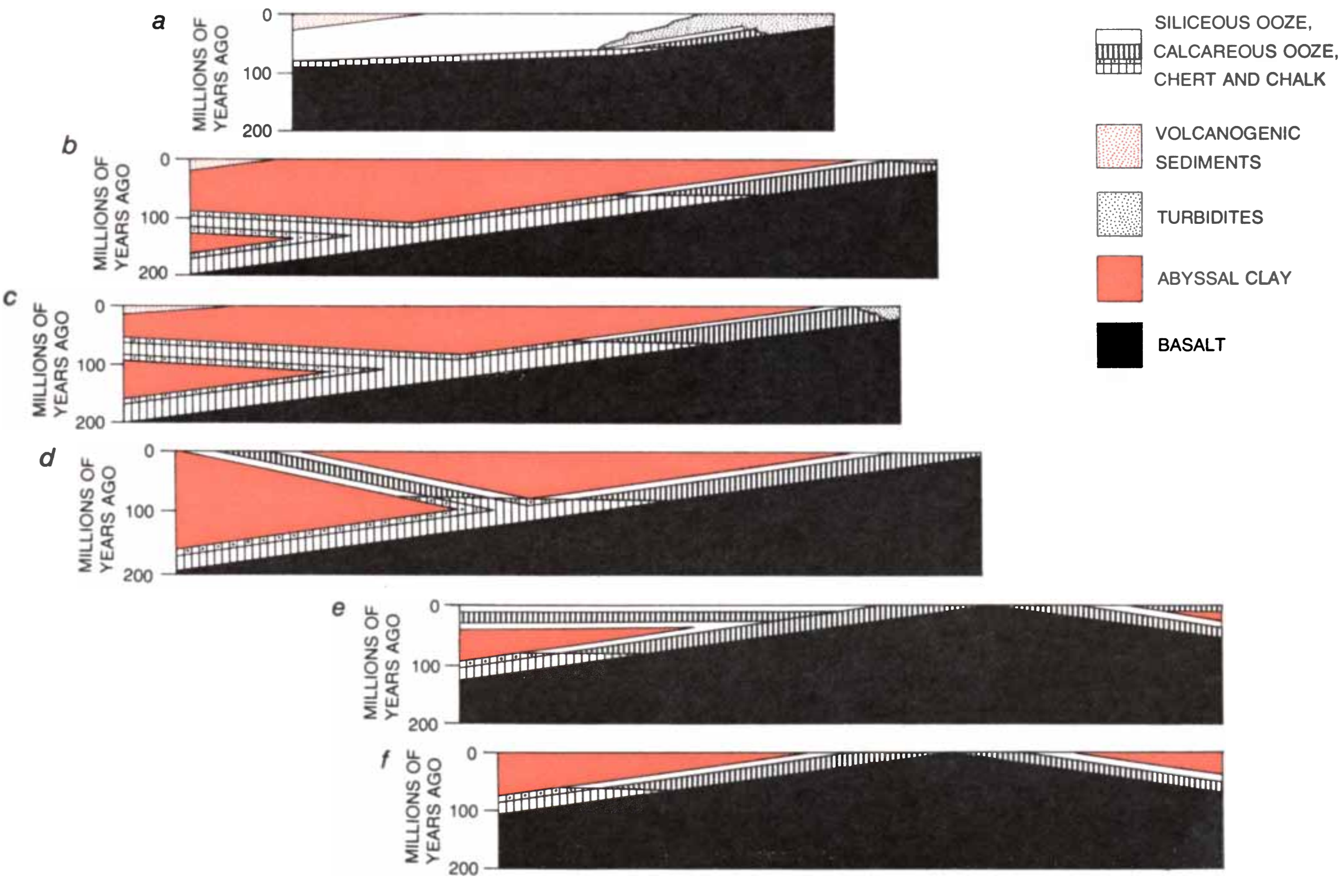
normal circumstances be no more than 200 meters thick.

Now, the sea floor in the western Pacific, the deepest and oldest part of the basin, is at least as old as early Cretaceous times, some 120 million years ago. If during the first 20 million years of its existence this part of the basin accumulated organic sediments at the rate proposed in our model, then 100 million more years were available for the deposition of abyssal clays on top of the organic oozes. At the rate of one meter of clay per million years, therefore, the western Pacific sea floor should have accumulated 100 meters of abyssal strata on top of the 200-meter organic layer. How do the sediments in the western Pacific compare with the predictions of the dynamic model? Before answering the question we must introduce another kind of information.

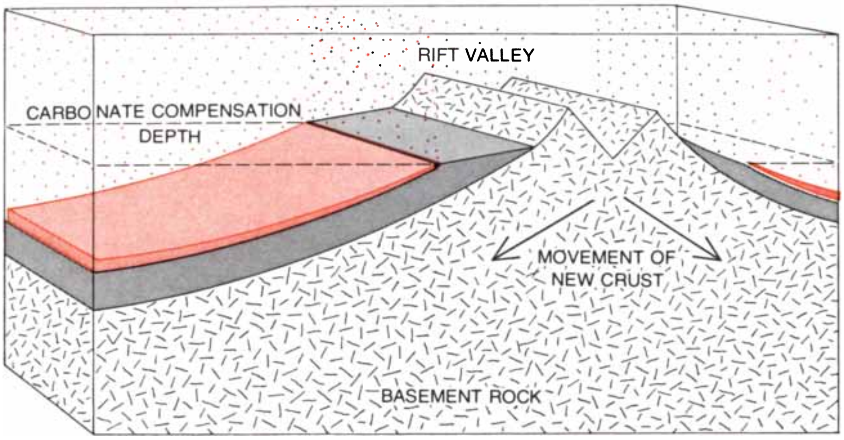
Drilling is the only way to collect samples of the layers of sediment below the ocean floor but it is by no means the only way to gather information about these strata. One method that has been used increasingly since World War II is acoustical profiling. In this method a strong acoustic signal is returned from the interfaces of layers of contrasty sea-floor sediment and is recorded in a continuous profile. Profiling thus makes it possible to trace the lateral distribution of different sedimentary layers for hundreds and even thousands of miles. In the Pacific profiling indicates that a layer of unconsolidated and relatively "transparent" sediments, less than 100 meters deep, covers the central part of the North Pacific floor [*see lower illustration on page 103*]. Such a region of shallow sediments would correspond to the area where our dynamic model predicts the buildup principally of abyssal clays. Along the northern and northwestern boundaries of the basin the profiles show a much thicker transparent layer; the greater thickness apparently corresponds to the rapid accumulation in these regions of continental debris and volcanic ash.

The acoustical profile along the Equator shows that the sediments there form a massive deposit, triangular in cross section and in places as much as 1,000 meters thick. Such a large accumulation corresponds to the high rate of sedimentation in a zone where upwelling nutrients produce an abundance of plankton. The acoustical profiles in the eastern Pacific are also what would be expected; the transparent layer of surface sediments, presumably abyssal clays, is present and is not unusually thick. In the western Pacific, however, not one or two but four separate sedimentary layers appear in the profiles. Just below the transparent layer on the sea-floor surface is an "opaque" layer. A second transparent layer underlies the opaque layer and a second opaque layer underlies the second transparent layer. This fourth layer rests directly on the oceanic crust [*see upper illustration on page 103*].

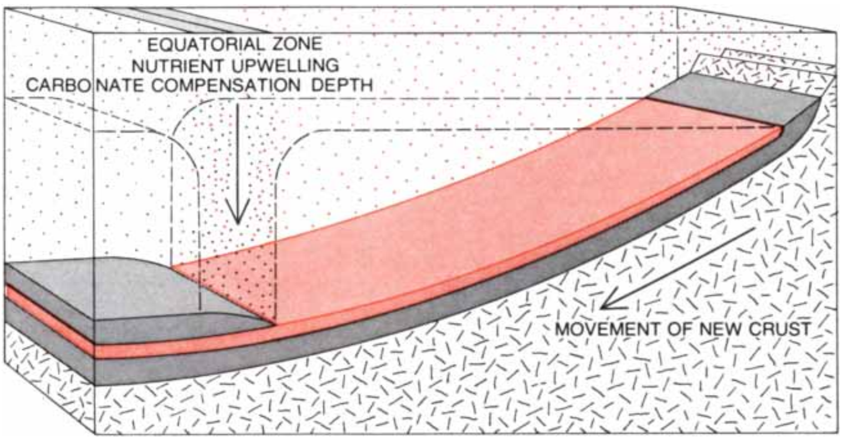
The acoustical data are in good agreement with the specific information about stratigraphic sequences provided by deep-sea drilling. The cores show that the bottom layer of sediments, resting directly on the oceanic crust both in the older, deeper western Pacific and in the younger, shallower eastern Pacific, is or-



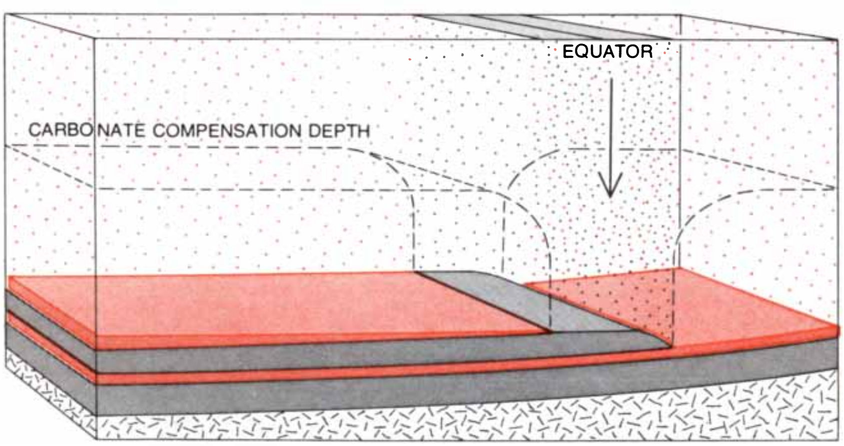
CROSS SECTIONS OF SEDIMENTS along the colored lines in the illustration on the opposite page are shown schematically. The wedge-shaped form of the cross sections results from the fact that, as the sediments were being deposited, the basalt crust below them was moving. The time at which the sediments were being deposited is given by the scale at the left side of each of the cross sections.



DYNAMIC MODEL of the Pacific basin in the vicinity of the East Pacific Ridge shows how different kinds of bottom sediment accumulate. Oceanic crust, newly generated in the rift valley at the crest of the ridge, slowly moves outward and downward. While the crust remains above the "carbonate compensation depth" most of the detritus that settles on it is organic (*black dots*) and a thick stratum of ooze is formed. When the crust enters deeper water, only inorganic detritus (*colored dots*) reaches it and forms a thin stratum of clay.



REVISED MODEL makes the oceanic crust travel through the equatorial zone, where abundant planktonic growth brings about an abrupt lowering of the carbonate compensation depth. The heavy "rain" of organic detritus (*black dots*) in this zone deposits a second layer of ooze on the abyssal clay. As a result three separate sedimentary strata overlie crust.



FOURTH LAYER OF SEDIMENTS accumulates above the three earlier layers after the oceanic crust passes beyond the equatorial zone. The crust is once again in water far below the carbonate compensation depth and only inorganic detritus (*colored dots*) reaches the floor of the ocean. As a result a thick second stratum of abyssal clays builds up, covering second stratum of organic ooze. This is the kind of layering characteristic of western Pacific.

ganic in origin. The only difference between west and east is that the older deposits have turned into chalks and cherts whereas the younger ones are still unlithified oozes. In both west and east, the cores show, the basal organic layer is overlain by a layer of abyssal clays.

The clays in the western Pacific are older than the clays in the eastern part of the basin. In addition the western clays are overlain by a second layer of chalks and cherts, and that organic layer is covered by a second layer of inorganic clays. Thus the cores show precisely the same four-layer sequence that is revealed by acoustical profiling, and it is obvious that serious discrepancies exist between the four-layer stratigraphy in the western Pacific and the two-layer prediction of our dynamic model. The model sequence—oceanic crust overlain first by organic oozes and then by inorganic clays—is correct as far as it goes, but it fails to describe the real sequence. There is another difficulty too. We find two depositions of organic sediments in the western Pacific rather than one, and the buildup of this double ration of sediment would have needed more time. The process may have taken 50 million years, or more than twice the 20 million years allotted in the model's timetable.

Can our defective dynamic model be repaired? The answer is yes, if we alter its direction of movement. Let us add a new component—northward drift—to the slow westward creep of the freshly formed oceanic crust. If we do that, the new crust formed north of the Equator will, as before, accumulate only two layers of sediment: first organic oozes and then abyssal clays. The same two layers will also build up on crust newly formed south of the Equator. When the southern crust slowly moves northwestward across the Equator, however, it will receive a further heavy rain of detritus from that nutrient-rich zone, thereby acquiring a second layer of organic oozes [*see middle illustration at left*]. When the crust once again enters waters that contain nothing but inorganic matter, it will start to acquire a second layer of abyssal clays on top of its second layer of organic oozes.

The predictions of our revised dynamic model compare favorably with the results of deep-sea drilling in the Pacific basin, although we should like to see cores taken at many more locations. In the cores from the western Pacific that are available now the oldest chalks and cherts are of lower Cretaceous age (and some are probably even older). This layer evidently represents the organic

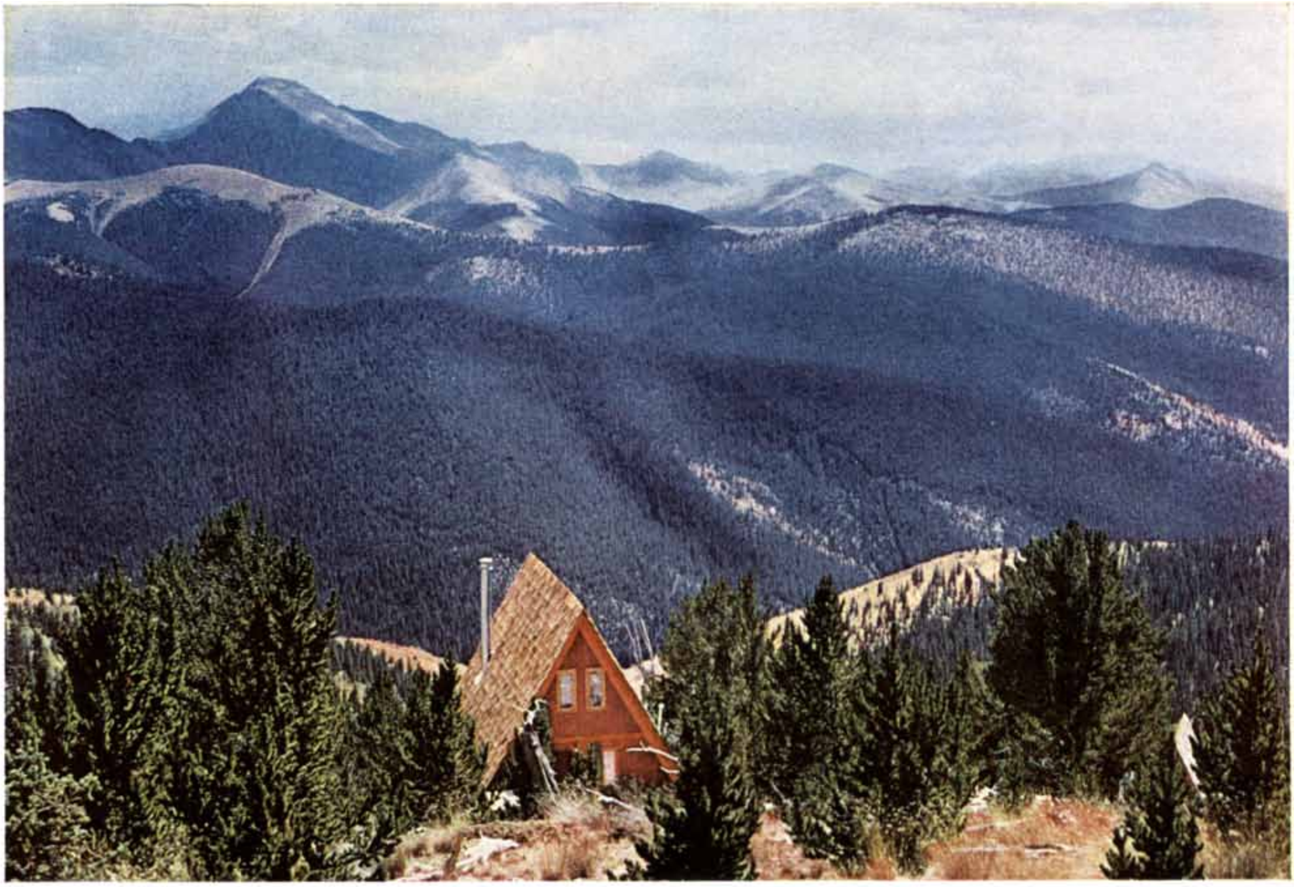


Photo of actual ranchland taken at Sangre de Cristo Ranches

Where the deer and the antelope play

Sangre de Cristo Ranches. Midway between Denver and Albuquerque. Where the Kit Carson Trail meets the Navajo Trail in beautiful southern Colorado.

Big country. Wide open spaces. Majestic mountains surrounding green hillsides of the sleepy San Luis Valley. Rugged, yet serene. Crisp, clean air, scented with sagebrush and pine.

You can own land here. Leg-stretching, five-acre minimums. Land for ranches. For vacations. Retirement. For settling down and getting away from it all. For long term investment. To leave to your kids and for them to leave to theirs.

This is good land. Land you can see for yourself. Land you can stand on and dig into. It's a subdivision of the gigantic, 260 square mile Forbes Trinchera Ranch. Forbes Inc., publishers of Forbes Magazine and owners of Sangre de Cristo Ranches Inc. acquired the Trinchera Ranch some years ago. One of the oldest of the remaining big ranches in America, it ranks among the best known preserves for deer, elk, game birds and other wildlife.

Located near the intersection of U.S. Route 160 and Colorado Highway 159 just outside Fort Garland, Colorado, Sangre de Cristo Ranches are spread over thousands of rolling acres which rise to the towering mountain peaks of the Trinchera and Mount Blanca.

Since Forbes Inc. began offering this section of its big ranch for homesites, thousands of people have come from all over the nation to see this magnificent land, to purchase their own Sangre de Cristo Ranch.

Important guarantees backed by Forbes Magazine's distinguished reputation have contributed much to the Sangre de Cristo success story. Every investor is double protected: he has twelve months after his first payment during which he can—and is urged—to visit his ranch. And if it doesn't fully meet expectations, he can have all payments promptly refunded. Or if he simply changes his mind within sixty days of making his first payment, his money will be returned promptly.

You can own part of this breathtaking land for just $5,000 total cash price. Easy credit terms are available. Also, other Sangre de Cristo ranch sites are available at prices ranging from $5,000 to $10,000. This offer presents a wonderful opportunity for long term investment in the great southwest.

For complete details on these ranches, without obligation, simply fill in and mail the coupon below to Sangre de Cristo Ranches, Box 2000, New York, N.Y. 10011.

SANGRE DE CRISTO RANCHES INC., is a duly registered subdivision developer in and under the laws of the State of Colorado. This registration is not to be construed to imply state endorsement of the subdivision.

Sangre de Cristo Ranches Inc.

DIVISION OF FORBES MAGAZINE
Box 2000, Old Chelsea Station, New York, N.Y. 10011

Without obligation, please send me FREE, more information on the Sangre de Cristo Ranches.

Name ______

Address ______

City ______ State ______ Zip ______

Telephone ______

Price Range: ☐ $5,000 ☐ $7,500 ☐ $10,000

A statement and offering statement has been filed with the Department of State of the State of New York. The filing does not constitute approval of the sale or lease or offer for sale or lease by the Department of State or any officer thereof or that the Department of State has in any way passed upon the merits of such offering. A copy of the offering statement is available, upon request, from the subdivider. NYA 73-975 Not available to residents of states where prohibited by law.

A635 AD 6609 (b)

detritus accumulated on the new crust during its initial subsidence. The abyssal clays that overlie the basal organic layer similarly represent the deposition of inorganic detritus after the new crust sank below the normal carbonate compensation depth. The chalks and cherts in the layer above the lower abyssal clays are middle Cretaceous in age; they evidently record the passage of this part of the Pacific crustal plate northwestward across the Equator. Finally, the abyssal clays above the second organic layer are late Cretaceous or Cenozoic in age; they must represent the further accumulation of inorganic sediments after the equatorial zone was left behind.

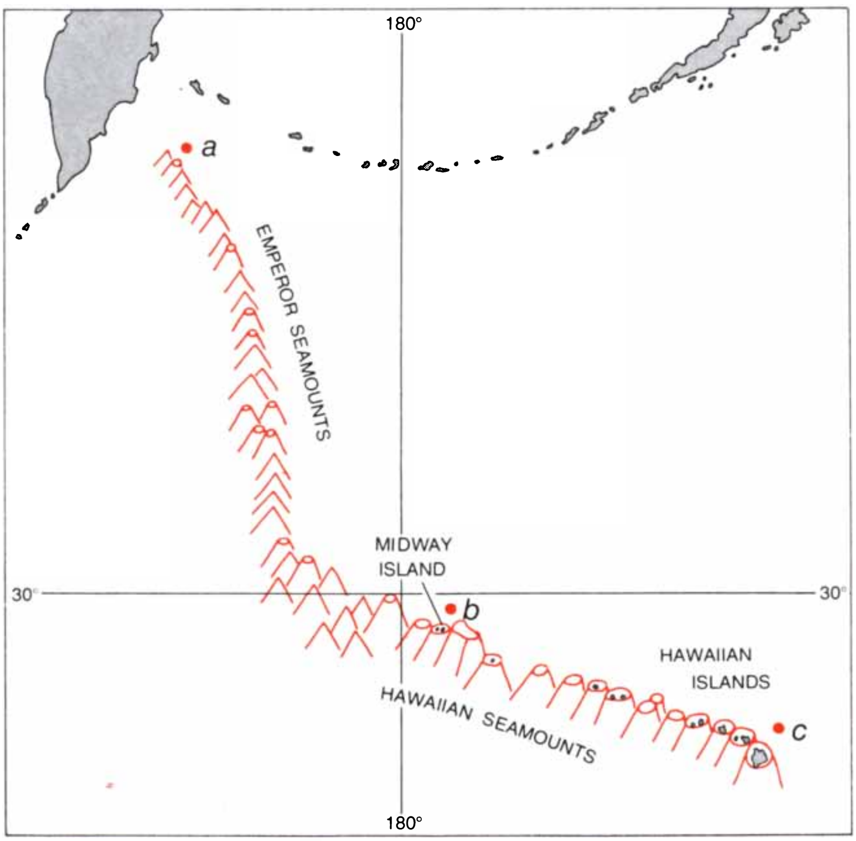
Other kinds of evidence support the concept illustrated by our revised dynamic model. Provocative evidence on the probable directions of motion of the Pacific plate is provided by two long lines of submerged Pacific seamounts: the Hawaiian and the Emperor chains [*see illustration on this page*]. It has been proposed that the seamounts are the truncated remains of former volcanoes that were successively brought into existence by the action of a "hot spot" rooted to the deep mantle below the Pacific crustal plate. In principle the two long seamount chains combined represent a kind of fossil record of the slow movement of the Pacific plate across the hot spot over millions of years.

As a result of drilling and dredging the ages of such features can be estimated. The oldest seamount, at the northern end of the Emperor chain, was evidently submerged some 70 million years ago. Midway Island, near the western end of the Hawaiian chain, is only some 20 million years old. The volcano Kilauea, on the island of Hawaii at the eastern end of the chain, is still erupting. The ages of seamounts located at the "bend" where the two chains meet have not yet been determined, but on the basis of extrapolation they are thought to be from about 30 to 40 million years. Since the ages increase to the north and west it would appear that the Pacific plate has moved west-northwest in a direction parallel to the Hawaiian chain for the past 30 to 40 million years, and that for 30 to 40 million years before that time it had moved in the more northerly direction parallel to the Emperor chain.

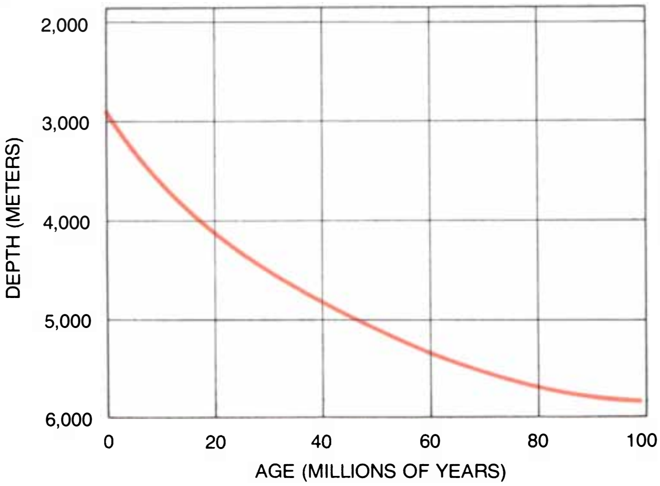
It is possible to estimate on the basis of the deep-sea drilling data the average northward component of the plate's motion. The point in any core where the bottom layer of the youngest abyssal clays is in contact with the top layer of the underlying chalks and cherts can be interpreted as a temporal boundary; it indicates approximately when that part of the plate completed its transit of the equatorial zone. In cores taken at a variety of latitudes in the Pacific basin the age of this temporal boundary has been determined from the fossils preserved in the uppermost chalks and cherts. The results suggest that, for a period of some 80 million years beginning in the late Mesozoic and continuing until middle Cenozoic times, the northward movement of the Pacific plate proceeded at a rate of some 4.4 centimeters per year. Since then, for the past 30 million years or so, the northward component of motion has been smaller: some two centimeters per year [*see top illustration on next page*].

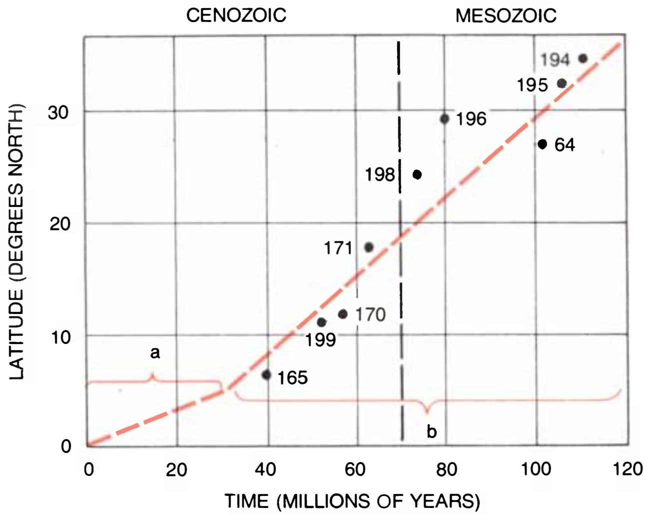
Still further evidence with respect to the motion of the Pacific plate comes from the observation that in oceanic crust generated along an east-west axis the bottom is "striped" with zones of alternating magnetic polarity, the result of past reversals of the earth's magnetic field. The magnetic striping is either nonexistent or poorly developed, however, in crust that has been generated in the equatorial zone; here the crust is magnetically quiet. Now surveys show that a huge area of the North Pacific is also magnetically quiet. The region starts near the present intersection of the Equator and the East Pacific Ridge, in the vicinity of 100 degrees west longitude, and continues generally north by west in a wide swath that terminates off the Kamchatka Peninsula in the northwest Pacific [*see bottom illustration on next page*].

Up to now the existence of this large quiet area has been explained as evidence that during a considerable part of the Cretaceous there were no reversals of the earth's magnetic field. Although there may have been such a period, our model predicts that a magnetically quiet zone should be present. A better explanation, in our opinion, is that the quiet region represents crust that was formed within the equatorial zone and that has

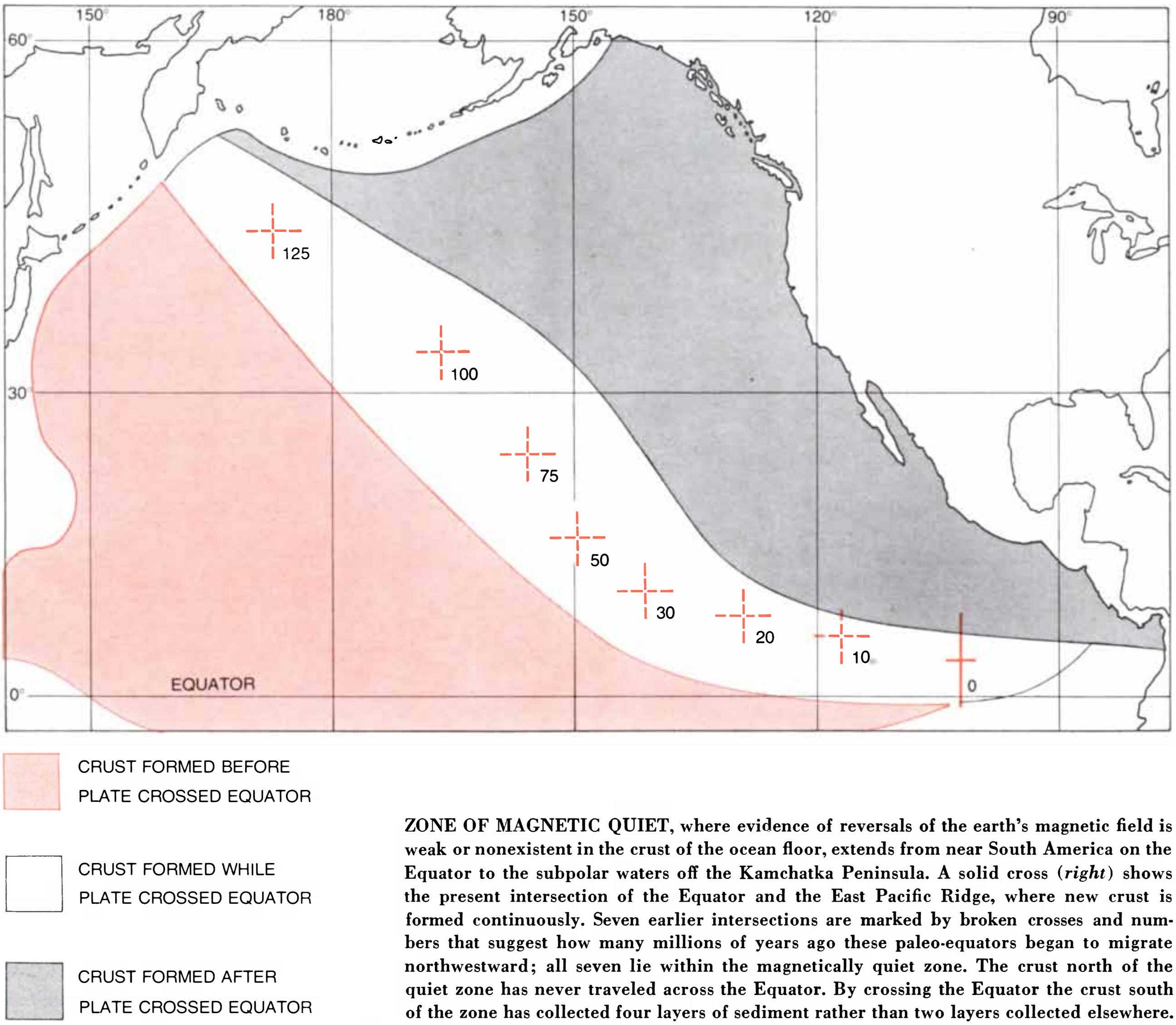


LINE OF SEAMOUNTS, the Emperor chain (*left*) and the Hawaiian chain (*right*), traces the past movement of the Pacific plate over a "hot spot" rooted in the mantle near Hawaii. As the plate moved northwestward over the hot spot, volcanoes appeared; the seamounts are their submerged remnants. A seamount (*a*) at the northern end of the Emperor chain is the oldest known: 70 million years. This contrasts with an age of 20 million years for Midway Island (*b*) and one of less than a million years for the volcano Kilauea on Hawaii (*c*).





AGE OF FOSSILS in the uppermost organic strata of sea-floor sediments in the Pacific increases with depth (*left*) and distance north of the Equator (*right*). Two rates of plate movement become apparent (*broken line*) when the trend is averaged. Between 120 and 30 million years ago (*b*) the rate was some 4.4 centimeters per year. Since then (*a*) it has diminished to some two centimeters.



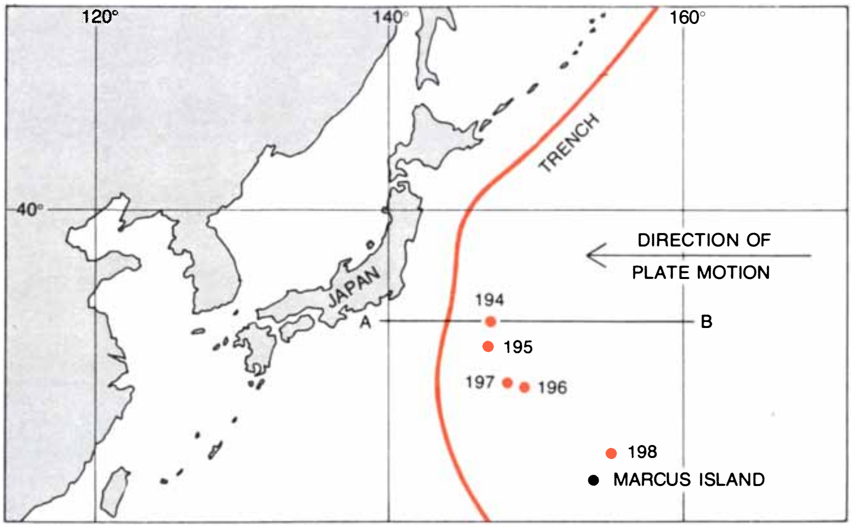
ZONE OF MAGNETIC QUIET, where evidence of reversals of the earth's magnetic field is weak or nonexistent in the crust of the ocean floor, extends from near South America on the Equator to the subpolar waters off the Kamchatka Peninsula. A solid cross (*right*) shows the present intersection of the Equator and the East Pacific Ridge, where new crust is formed continuously. Seven earlier intersections are marked by broken crosses and numbers that suggest how many millions of years ago these paleo-equators began to migrate northwestward; all seven lie within the magnetically quiet zone. The crust north of the quiet zone has never traveled across the Equator. By crossing the Equator the crust south of the zone has collected four layers of sediment rather than two layers collected elsewhere.

reached its present position as a result of the slow northwestward movement of the Pacific plate. An eastward extrapolation of paleo-equator positions determined from deep-sea drilling, together with a westward extrapolation of crustal age based on dated magnetic stripes, enables us to estimate the location and age of a series of points where the East Pacific Ridge and earlier "paleo-equators" once intersected.

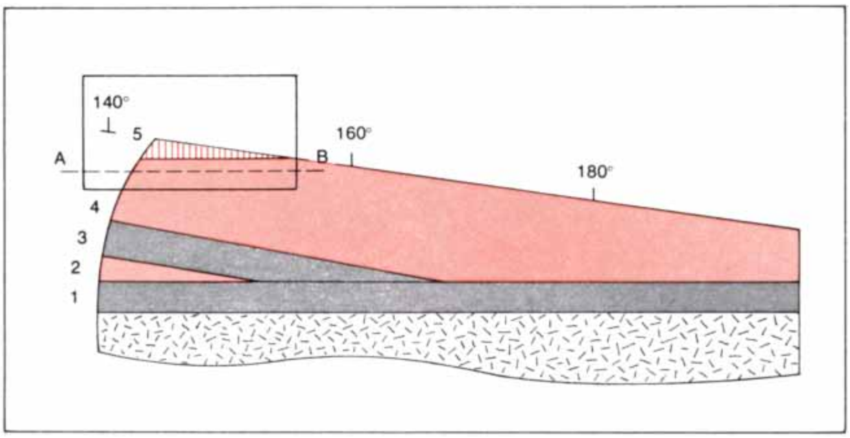
In our interpretation the magnetically quiet region represents a boundary between the two principal components of the Pacific crust. North of the quiet region's northern boundary is crust that has never made an equatorial crossing. South of the southern boundary is crust that was formed before a crossing, so that during its subsequent equatorial transit it received a second ration of organic sediments. The quiet belt between these regions would of course be crust formed during the equatorial transit.

Our refined model applies to an oceanographic problem a particular kind of geological reasoning: the facies concept. The concept arises from the observation that sedimentary deposits reflect the circumstances and environment of their origin. In brief, it states that just because similar sets of strata are found in two separate sedimentary deposits this does not necessarily mean that the strata accumulated at the same time. It is just as likely that their similarity is due to the existence of similar environmental conditions at quite different times. For example, if the margin of a continent is subsiding, a shoreline environment will "migrate" inland; the sedimentary deposits of each successive shoreline will become a rock formation essentially identical with the one preceding it, but what may appear to be a uniform lithological unit will not have been deposited at the same time. In the same way our model of the Pacific basin assumes that the ocean's various sedimentary environments have remained constant while the plate comprising the floor of the basin has migrated. The strata that are similar are not contemporaneous; instead they represent the plate's passage through a series of constant environments.

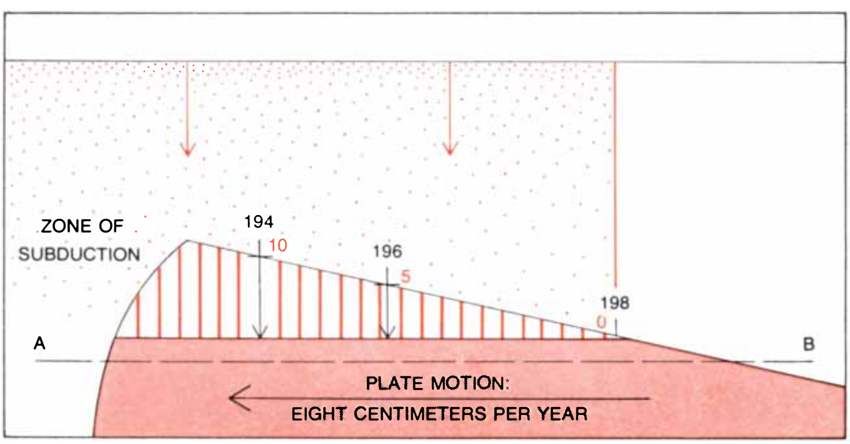
It should be emphasized that the simplicity of our model necessarily gives it a number of shortcomings. For example, it is a depositional model that ignores episodes of erosion, changes in sea level and possible shifts in compensation depths. It scarcely considers such variations in plate motion as rotation or changes in direction and the rate of



DEEP-SEA CORES from an area in the western Pacific between the Japan Trench and Marcus Island reveal the presence of a wedge of volcanic ash that overlies a part of the four-layer sequence of sediments in this area. In the first illustration below is a cross section.



LAYER OF ASH is a fifth sedimentary deposit on the sea-floor crust. The lowest layer is the first deposit of organic detritus, accumulated after the crust was formed. The small wedge above it is the first deposit of inorganic abyssal oozes, which here pinches out. Above that are a second organic layer, which pinches out farther to the east, and a second, very thick inorganic layer that comprises the ocean bottom. A closer view of section is shown below.



THICKNESS OF LAYER increases from east to west. Assuming a sedimentation rate of 15 meters per million years, the lowest ashes in drill hole No. 196 would be some five million years old and those in hole No. 194 would be some 10 million years old. Ash much older than that has been lost in the zone of subduction at western edge of moving Pacific plate.

drift. Neither does it particularly concern itself with the great mass of sediments that comes from continental runoff, shore erosion and volcanic activity.

Concerning these continental sediments, however, one point should be made. The northward and westward motions of the Pacific plate and its subduction under the island arcs of the northern and western Pacific do affect the size and shape of the great sedimentary wedges that extend out from the continents. At the boundary of the subduction zone the dimensions of the wedge will be determined by the combined rates of sedimentation and subduction. A rapid rate of sedimentation will tend to advance the wedge seaward, whereas subduction will counteract this tendency by continuously moving the sediments back toward and under the continent.

Drill cores taken in the vicinity of the Japan Trench, northwest of Marcus Island, provide good examples of the layers of volcanic ash that began to build up on the surface of the abyssal mud in this area late in Miocene times, some 12 million years ago. The ash here has not yet reached the subduction zone, and the volcanic accumulation is some 150 meters thick [*see middle illustration at left*]. Some 800 kilometers farther to the southeast the wedge of ash thins to the vanishing point, leaving the uppermost layer of abyssal clay uncovered. Drillings in between show that the wedge is intermediate in thickness. If we assume that the sedimentation rate for volcanic ash has averaged about 15 meters per million years, then it appears that the Pacific plate is moving into the subduction zone at a rate of eight centimeters per year and that the volcanic ash deposited on the Pacific sea floor in middle Miocene and earlier times has already vanished under the continental plate.

In summary, the fact that our simple model gives a good first-order fit with the data so far available suggests that the evolution of the Pacific basin has been governed by a few equally simple and systematic factors. Among these are the carbonate and silica compensation depths, the gradual sinking of the oceanic crust as it creeps westward from the East Pacific Ridge and the simultaneous northward motion of the crust. Data already exist, of course, that deviate from the model, and future observations will surely produce more such data. Such deviations are welcome. Not only should they reveal local variations in the sedimentary history of the Pacific basin but also they may demonstrate irregularities in the motion of the Pacific plate itself.

EVERYWHERE

you find people working with numbers, words, data and information, you'll find Olivetti. In offices, factories, banks, schools. In airports and hotels. In government offices. In every country of the world!

IN AMERICA

the Olivetti sales and service organization covers the country - with 80 district offices, a network of agents and dealers, and service centers coast to coast staffed with Olivetti-trained personnel. We also have teams of specialists whose sole job is to help you find solutions to specific problems and show you how to get the most out of our machines and systems.

FOR IN-HOUSE PROCESSING OF DATA

Olivetti leads the way. We created a world market for microcomputers by bringing programmed electronic calculation facilities right to the desk. Our famous Programma 101 was the world's first desk-top computer to use programs stored on magnetic cards. Since then we have exploited each new technological advance, developing new machines and new systems with ever increasing power and flexibility. And now we're proud to announce our latest achievement:

P652 SYSTEM FOR ADVANCED TECHNICAL AND SCIENTIFIC APPLICATIONS

allows you to process data for scientific research and technical projects on the spot, quickly and economically. The P652 has an exceptional storage and calculating capacity. Its keyboard is simple, functional; doesn't require any special operating skill. And it's an open-end system. Which means it can be extended at any time by adding peripherals using a variety of input/output media - from paper tape to magnetic cassettes.

olivetti

president Express Diners Club card.

DINERS CLUB
10 Columbus Circle, New York, N.Y. 10019
[PLEASE PRINT ANSWERS TO ALL QUESTIONS]

20-101-991

FOR OFFICE USE ONLY

1. ☐ Personal Account—mail bill to residence ☐ Personal Account—mail bill to office ☐ Company Account—mail bill to office

2. ☐ Mr. ☐ Miss ☐ Mrs. ☐ Ms. | First Name | Middle | Last | Date of Birth: Month Day Year | Spouse (first name)

Home Address: Street | City | State | ZIP Code

Years at Present Address | Own Home ☐ Rent ☐ | Telephone (include area code) | Number of Dependent Children | Social Security Number

Previous Home Address: Street | City | State | ZIP Code | Years there

3. Firm Name or Employer | Nature of Business | Position

Address: Street | City | State | ZIP Code

Telephone (include area code) | Years with firm | Annual Earnings $ | NOTE: If annual earnings are less than $10,000 indicate source of other income below

Amount and Source of other income

Previous Employer (if employed by above less than 3 years) or College/University if Recent Graduate | Yrs. with firm or yr. graduated

Spouse's Employer | Address | Position | Annual Earnings $

Name and Address of Personal Reference (not living with you)

4. Name of Bank | City & State | Acct. Number
Type of Account: ☐ Checking ☐ Savings ☐ Loan

Name of Bank | City & State | Acct. Number
Type of Account: ☐ Checking ☐ Savings ☐ Loan

Department Store Accounts
1. ____
2. ____

Oil Company Accounts
1. ____
2. ____

Other Credit Accts:
☐ American Express
☐ Carte Blanche
☐ Bank Americard
☐ Master Charge
Other: ____

Former Diners Club member? ☐ Yes ☐ No

5. $17 FEE: covers 12 months' membership from date card is issued at $15 plus 1 year's subscription to Signature, the Diners Club Magazine, at $2. Subscription optional. Indicate choice below. Do **NOT** enclose check—we will bill you later.
☐ $17 Fee (Includes Signature Magazine) ☐ $15 Fee (Membersnip Only)

Send me an additional Diners Club Card at $7.50 for a member of my immediate family.
First Name | Middle | Last | Signature of Add-on applicant | Relationship
Please send me____forms for additional cards for members of my firm/family.

Applicant authorizes exchange of credit information, the issuance of credit cards as indicated and renewal and replacement thereof. Applicant agrees that each cardholder assumes joint and several responsibility for all charges with company applicant, if Company Account, or with primary applicant if Personal Account.

Signature of Company Officer for Authorization of Company Account | Title

Signature of Individual Applicant | Date

[HAVE YOU SIGNED THIS APPLICATION?]

MATHEMATICAL GAMES

Fantastic patterns traced by programmed "worms"

by Martin Gardner

Fashionable methods of teaching mathematics to children come and go. (At the moment the modern-math fad is mostly going.) One of these days mathematics teachers will discover what John Dewey tried to tell them 75 years ago: Children learn best by doing something they enjoy. With this in mind, Seymour A. Papert, a former assistant to Jean Piaget who is now working in the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology, has designed a variety of animal robots that can be controlled by a desk computer. One of them is a "turtle" with a pen on its underside. Suitably programmed, the turtle draws geometric figures by crawling across large sheets of paper on the floor.

Defining geometric figures as paths generated by a moving point is an ancient idea. Take a square. Instead of calling it a four-sided polygon with equal sides and angles, call it the path traced by a worm crawling over a plane according to the following rule: Go straight for distance k, turn 90 degrees left and repeat until the path returns to its origin.

The idealized worm (moving point) can obviously be programmed to generate any pattern of lines. A challenging recreational task now presents itself. What kinds of program, with extremely simple rules, give interesting or beautiful patterns? A good way to simplify rules is to restrict the worm to paths along a regular lattice. This enables one to experiment with rules by drawing paths on square or isometric graph paper. Better still, if one has access to a computer with a display screen, one can write simple programs, then enjoy the spectacle of watching a path of light grow on the screen.

Frank C. Odds, a British biochemist, recently proposed a class of rules for generating patterns that he named spirolaterals. The worm crawls a distance of one unit, turns, crawls two units, turns, crawls three units and so on, traversing distances in counting order until the length of a path segment reaches a specified integer, n, when the procedure starts over. The turning angle is always the same, but the direction of turn may be right or left according to the worm's program. The number n, which is the number of segments in the counting series and also the number of turns before the series repeats, is called the order of the spirolateral.

Two examples will make this clear. If the order is 1, the angle is 90 degrees and all turns are in the same direction, the spirolateral is a square. If the order is 7, the angle is 90 and all turns are the same, the spirolateral has a pleasing closed pattern. The top illustration on page 118 shows all the right-angle spirolaterals of orders 1 through 8, when all turns are the same. It is easy to see why Odds chose the name spirolateral: "lateral" for flat surface and "spiro" for the square spirals that generate the figures.

Note that spirolaterals of orders 4 and 8 do not close (return to the origin). As Odds puts it, they meander jerkily to infinity. Indeed, it is not hard to prove that there is no closure for orders in the series 4, 8, 12, 16..., that two repetitions of the square spiral will close the figure (producing twofold symmetry) if the order is in the series 2, 6, 10, 14..., and that all other orders close after four repetitions and have fourfold symmetry.

There is no reason why all turns must be the same way, but when they vary, the situation becomes complicated enough to call for a compact notation. Odds suggests writing the angle as a subscript of the order number and using a superscript on the left or the right to show which turns are left and which right. For example, ${}^{1,7}9_{90}$ defines the spirolateral generated by nine turns of 90 degrees, with left turns preceding segments 1 and 7 and (by implication) all the other segments turning right. The same spirolateral could be defined by putting a superscript on the right to indicate right turns, $9_{90}^{2,3,4,5,6,8,9}$, but clearly the first notation is preferable.

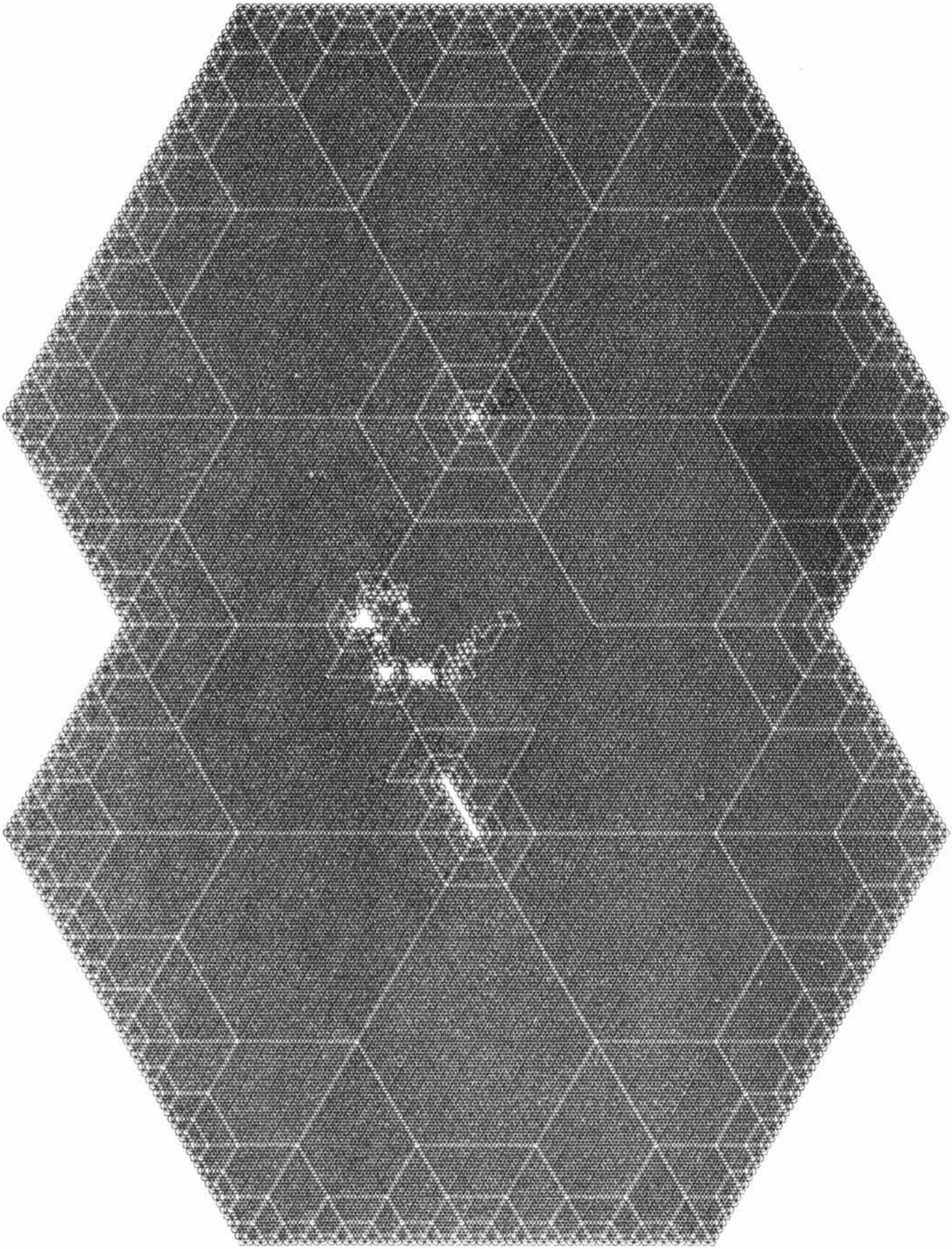
The middle illustration on page 118 shows two right-angle spirolaterals with mixed left and right turns. The bottom illustration shows spirolaterals with angles of 36, 45 and 60 degrees. Spirolaterals with 60-degree turns can be easily drawn on isometric paper. Those with 45-degree angles are not hard to construct with a ruler and a draftsman's triangle, but those with other angles take more doing. Any angle that is an exact divisor of 180 degrees will generate a spirolateral.

Not much is known about spirolaterals. Sometimes two different rules produce spirolaterals that are mirror images (for example ${}^{1,2,3}7_{90}$ and ${}^{5,6,7}7_{90}$), but no one knows how to predict this without drawing the figures. Nor is it known how to tell from looking at a spirolateral's formula, except in special cases, whether or not the figure will close and, if it does, how many repetitions are needed to close it.

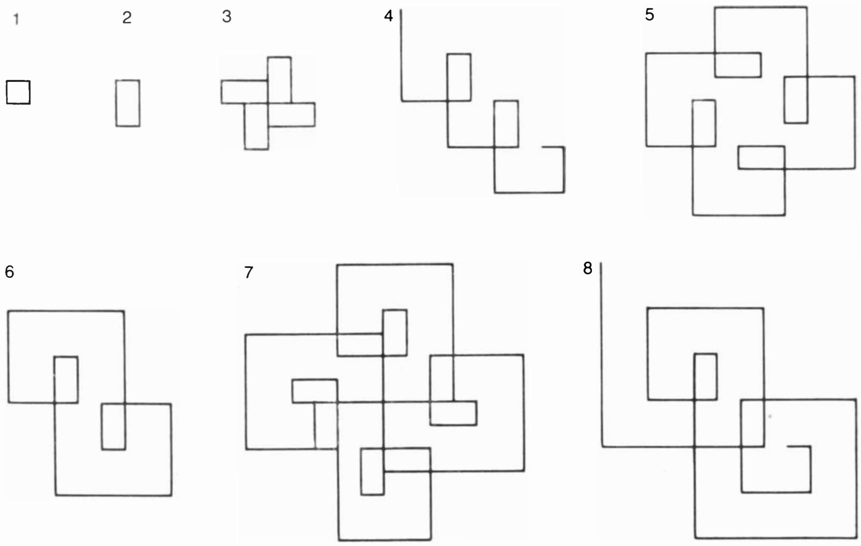
A few years ago John Horton Conway of the University of Cambridge suggested a new approach to worm paths. Instead of viewing the worm as an explorer, view it as an eater. Food is confined to the lines of an arbitrarily large grid. The worm hatches from an egg at one node and then starts to crawl along grid lines, eating as it goes. A fixed set of rules determines the worm's decision at each node. It is assumed that the worm never traverses a segment already eaten. No segment lengths need to be specified, only the direction of turn, the direction depending solely on the state (eaten or uneaten) of all segments meeting at that node.

"Paterson's Worm," by Michael Beeler, a memorandum issued in June of this year by M.I.T.'s Artificial Intelligence Laboratory, deals entirely with such worm tracks. What follows, either directly quoted or paraphrased, is taken from the memorandum with the permission of Beeler and Marvin L. Minsky, who heads the laboratory. (Readers can obtain the document by sending 80 cents to Room 912, 545 Main Street, Cambridge, Mass. 02139.)

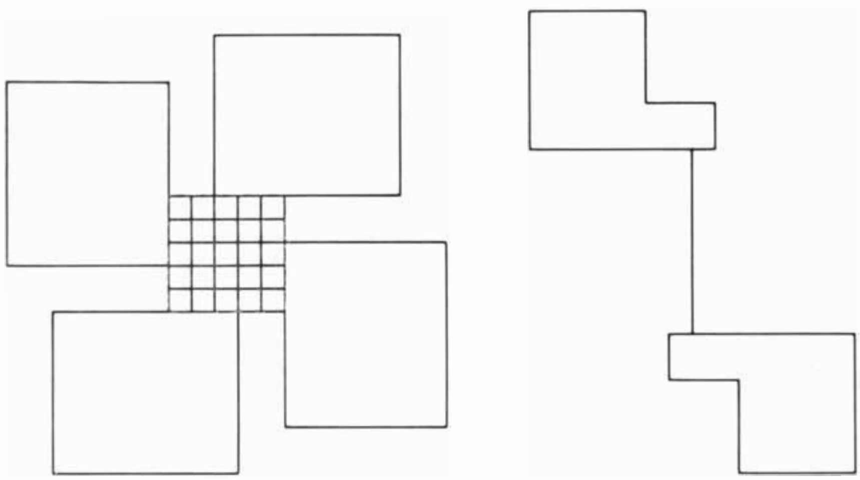
"Certain prehistoric worms fed on sediment in the mud at the bottom of ponds," Beeler's memorandum begins. "For efficiency, they would not retrace paths which had already been traveled, since little food was left there. Yet food probably occurred in patches, so it was desirable to stay near previous trails. Worms had innate 'rules' regarding how close to 'eaten paths' to stay, how far to



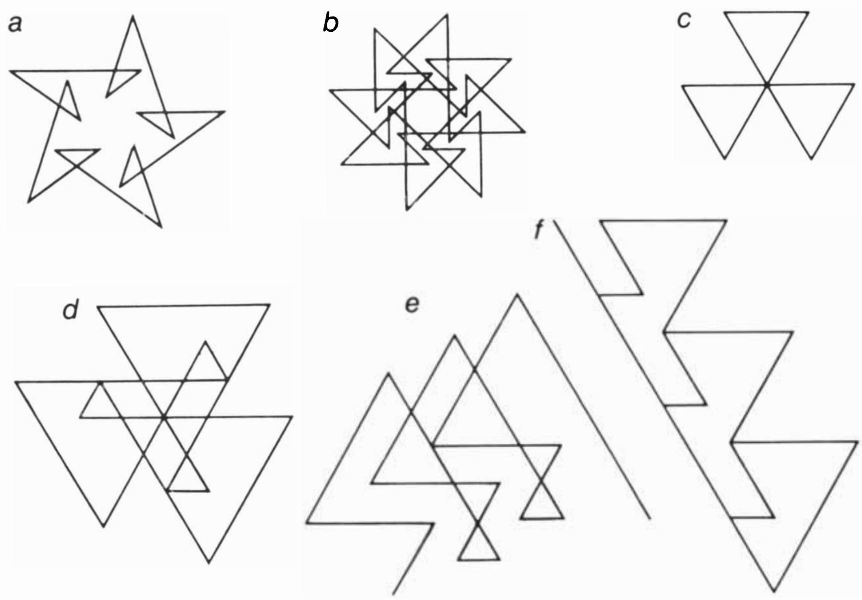
Longest known finite path generated by a simple isometric worm



Spirolaterals of orders 1 through 8 with all turns in same direction



Ninety-degree spirolaterals with mixed turns: 9_{90}^{6} (left) *and* $8_{90}^{1,4,8}$ (right)



Spirolaterals (a) 3_{36}, (b) ${}^{1}3_{45}$, (c) 2_{60}, (d) ${}^{1,2}5_{60}$, (e) ${}^{3}5_{60}$ *and* (f) ${}^{2}5_{60}$

go before turning around, how sharp a turn to make, etc. These rules varied from species to species, and paleontologists can trace the development of species and determine the similarity of different species by comparing fossil records of worm tracks. (See *Science* magazine, 21 November 1969, for further details and a discussion of computer simulation of natural worm tracks.)

"Early in 1971, Michael Paterson [a computer scientist at the University of Warwick] mentioned to me a mathematical idealization of the prehistoric worm. He and John Conway had been interested in a worm constrained to eat food only along the grid lines of graph paper....

"If a worm, arriving at a node with no segments eaten (except of course the one it just ate), should find in its rules, 'For this distribution, go straight,' then the worm will go straight forever. Since this is neither interesting to us, nor very useful to a real worm, which would quickly reach the edge of its food patch, we discard it. We require that, upon discovering a virgin node, all sets of rules must say to turn. To avoid mirror-image duplication, we require that the turn be to the worm's right (clockwise as seen from above)."

Consider what Beeler calls a "simple quadrille worm": a worm crawling on a square grid and turning right at each node. What happens after it traces a square? It cannot turn right again because this would take it along an eaten segment. It has only two choices. If it is programmed to turn left if, and only if, it cannot turn right, it will trace two squares [*see "a" in top illustration on opposite page*]. No uneaten segments are available, and so the worm dies. If it is programmed to go straight when it cannot go right, and left when it cannot go straight or right, it will trace five squares before it expires at its origin ["*b*"]. These two fossil tracks exhaust the variety of species of simple quadrille worms.

To avoid a worm's early demise, Conway proposed that quadrille worms have the ability to "look ahead" and see the distribution of eaten and uneaten segments at each adjacent node. For example, a worm could be programmed to turn right if, and only if, it senses that this will take it to a node where four uneaten segments meet. Otherwise it goes straight. The result is a simple square spiral. If the rule is to turn left whenever a right turn leads to a node with an eaten segment, the path is a more interesting spiral [*see middle illustration* *at right*]. It is easy, of course, to produce more elaborate paths by complicating the rules.

The rules may allow anything. What happens when rules allow look-ahead worms to hop? What happens when barriers are suitably placed or when the grid is bounded on all sides? What happens when two or more worms of the same or different species interact? What happens when a newly hatched worm crawls a short distance along a defined path (such as a straight line of three units) before its repetitive behavior starts? How about two armies of worms crawling toward each other, each army obeying a different program? Are there possibilities here for competitive games? Are there interesting paths or patterns in three dimensions and higher dimensions?

Beeler avoids such difficult questions by confining his attention to what I shall call "simple isometric worms": worms that feed along an isometric grid of unit equilateral triangles. They are simple worms because they do not look ahead. With the isometric grid, however, six segments (not four) meet at each node. It seems to be a trivial difference, but as Beeler makes clear the possibilities for variant rules allow the definition of no fewer than 1,296 species.

All simple worms, quadrille or isometric, obey three general rules:

1. If no segments have been eaten at a node (except the segment just traveled), the worm turns right.

2. If all segments at the node have been eaten, the worm dies.

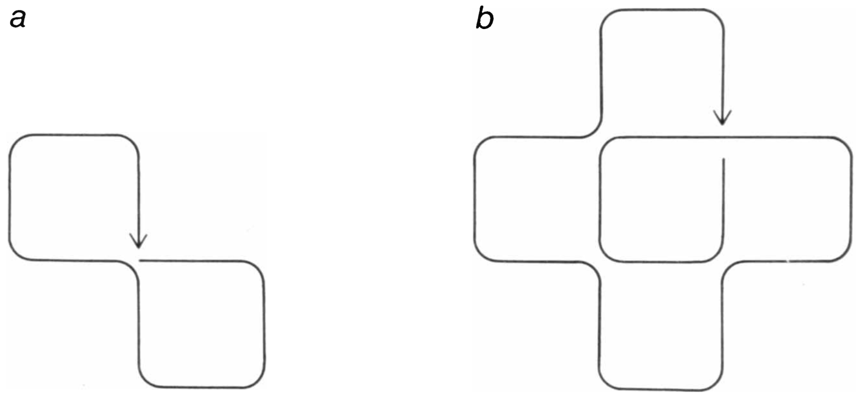
3. If only one segment at the node is uneaten, the worm takes it.

As we have seen, a simple quadrille worm, following the above rules, encounters only one "field" in which it must choose. Since it has only two choices, we can define only two species. The isometric grid, in contrast, offers a simple worm four major fields (one consisting of four subfields) in which decisions must be made. It is the behavior of these simple isometric worms that Paterson became interested in, and Beeler's memorandum also is primarily concerned with their behavior.

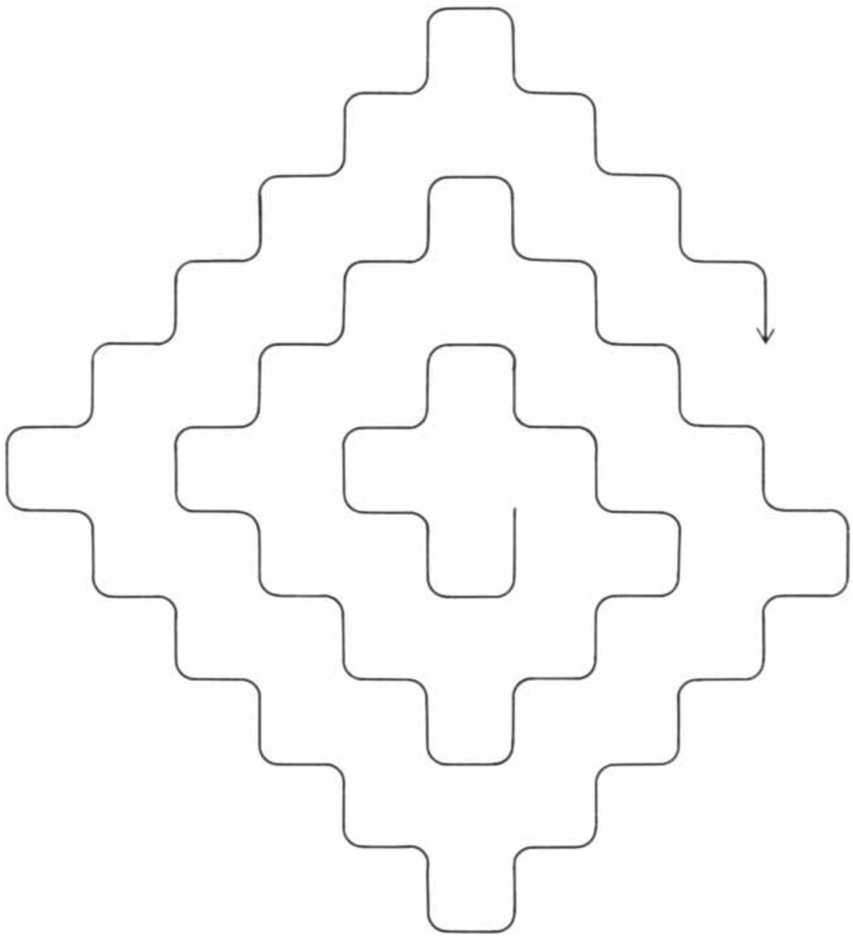
The four major fields, and all their choices, are shown in the illustration on the next page. The black lines indicate uneaten segments, the light colored lines are eaten segments of the worm's path and the colored arrows show how the worm approaches and leaves a node.

The four fields are:

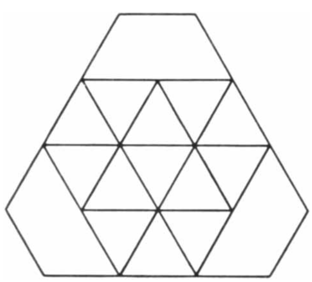
1. The worm approaches a node with no eaten segments other than the one it has just traversed. Its right turn can be



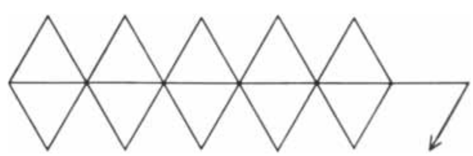
Fossil tracks of the only two species of simple quadrille worms



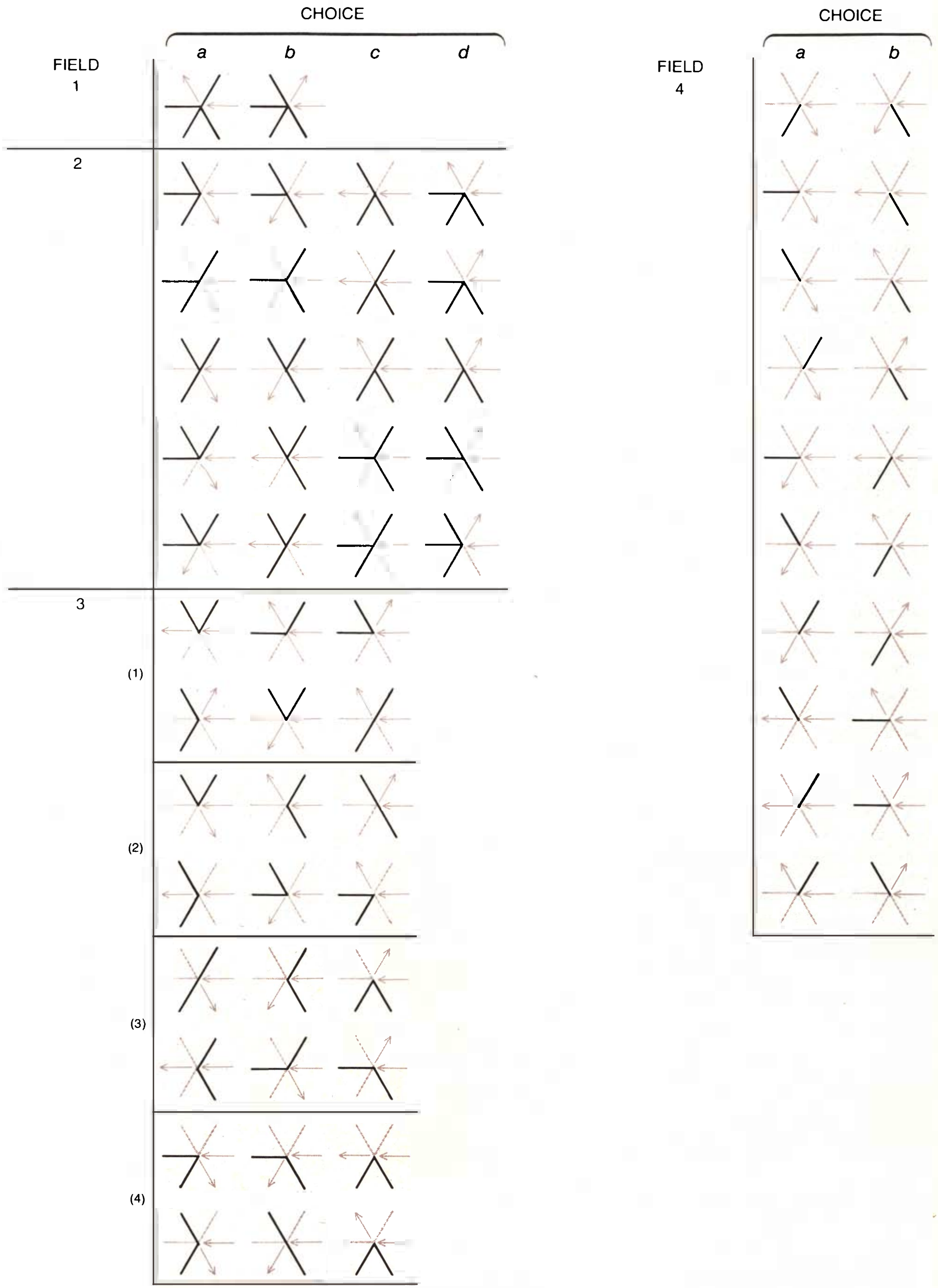
Infinite path of a "look ahead" quadrille worm that turns left when it cannot turn right



Path traced by 14 simple isometric worms



Zipper path of 54 sharp worms



Fields of choice for simple isometric worms either "gentle" (120 degrees) or "sharp" (60 degrees). Number of choices: two.

2. The worm meets one eaten segment as it returns for the first time to its origin. As the chart shows, it can approach this node along any of five different segments. For each approach there is a choice of leaving by one of four uneaten segments. Number of choices: four.

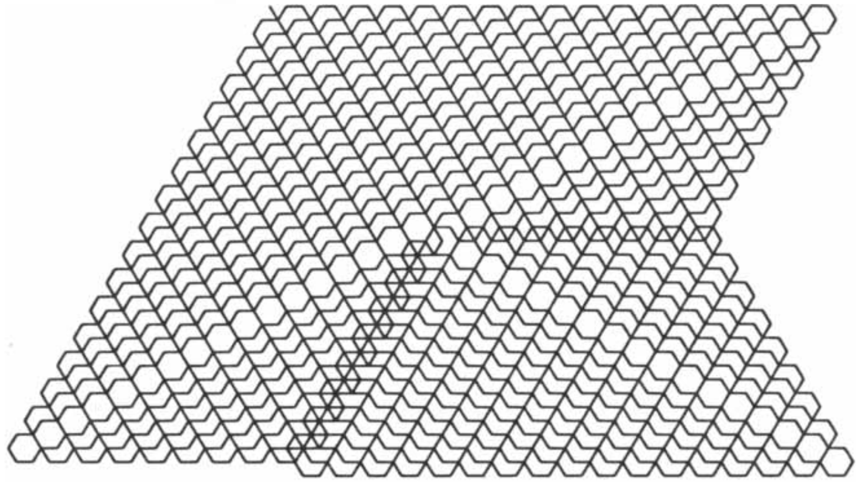
3. The worm meets two eaten segments as it returns to a point along its path. The node will be the vertex of either a sharp turn or a gentle one. In either case the approach can be made in four ways. For each way the worm can leave by one of three uneaten segments. Number of choices: $3 \times 3 \times 3 \times 3 = 81$.

4. The worm meets three eaten segments as it returns for the second time to its origin. This can happen in 10 ways, but each requires only a choice between two ways of leaving the node. Number of choices: two.

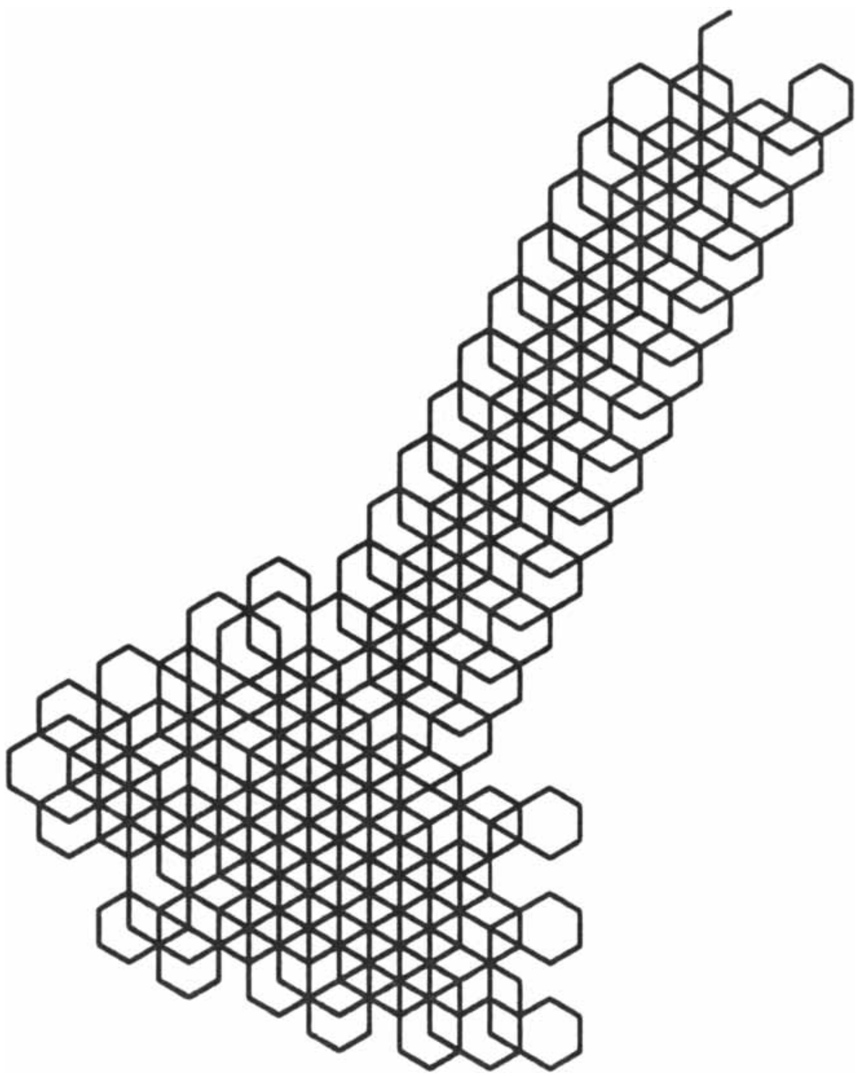
Meeting four or five eaten segments does not offer the worm a choice. In the first case it must take the only remaining uneaten segment. In the second case it has no segment to take, and so it dies. Thus there are $2 \times 4 \times 81 \times 2 = 1{,}296$ sets of rules, each defining a distinct species of simple isometric worm.

Beeler uses the pattern shown in the bottom illustration at the left on page 119 to explain how the rules work. The notation adopted here (Beeler uses a more compact notation based on octal and binary digits) identifies this as a fossil path generated by a $1_a2_b3_{acac}4_b$ worm. The formula tells us that when the worm faces a choice in Field 1, it follows Rule *a*. (It makes a gentle, not a sharp, turn.) For Field 2 it chooses *b*. The four subscripts after 3 refer to the four subfields of Field 3, where choices *a, c, a, c* are in the worm's feeding program. Finally, for Field 4 the worm selects *b*. As Beeler suggests, the reader might pause at this point to see if he can trace the path on a sheet of isometric paper, following the seven rules given by the formula. (It is necessary at each node to rotate the paper until the pattern at that node corresponds to the proper diagram on the chart.)

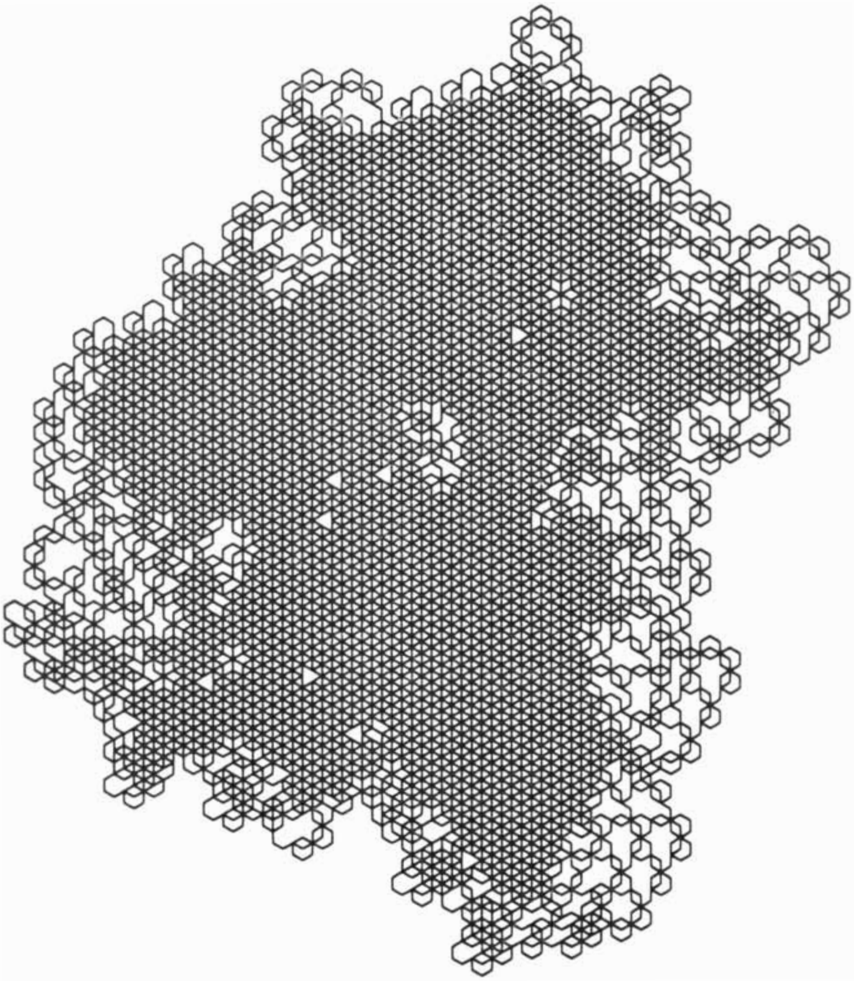
The path just described is generated by 14 worms. Some trace it exactly the same way. For example, the choice presented by the first subfield of Field 3 never arises, therefore it does not matter which of the three choices, *a, b* or *c,* is in the program. We indicate this in the formula by putting the three choices inside parentheses: $1_a2_b3_{(abc)cac}4_b$. The formula now describes three different



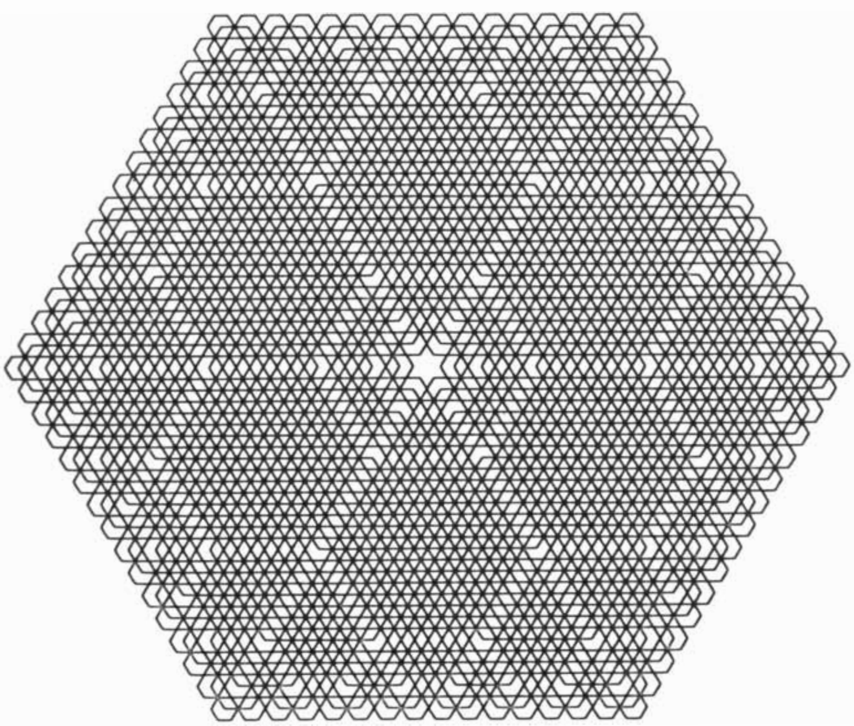
Three-pointed spiral by a $1_a2_c3_{acba}4_a$ *worm*



Shoot grower by a $1_a2_d3_{caaa}4_b$ *worm*



Cloud path generated by a gentle worm $1_a2_{abaa}3_c4$



The superdoily ($1_a2_d3_{cbaa}4_b$)

worms. Each draws the same path the same way. Other worms draw the same path different ways.

Beeler's computer program explored the behavior of all 1,296 species of simple isometric worm. Results show that 209 species generate paths that are unique, in the sense that no other species generates them. Forty-six paths are each characteristic of two species, and 44 paths belong to more than two. Thus there are 299 distinct paths all together.

The simplest of the 299 paths is the radioactivity symbol [*see "c" in bottom illustration on page 118*]. It has the smallest length (nine units) and also the fewest number of nodes (seven). It is the path traced by the largest number of worms, no less than 162 species. Using parentheses to show alternate choices, one formula defines all 162:

$$1_b2_{(ac)}3_{(abc)(abc)(abc)(abc)}4$$

The formula shows at once that it describes $2 \times 3 \times 3 \times 3 \times 3 = 162$ sets of rules. If the reader will test this formula on isometric paper, he will find that, regardless of the choice he makes inside each pair of parentheses, the radioactivity symbol will result.

Is there a longest path? No, because many paths never return to their origin a third time and are therefore infinite. A trivial infinite path of a type called the zipper is produced by 54 sharp worms with the formula $1_b2_d3_{(abc)(abc)b}4_{(ab)}$ [*see bottom illustration at right on page 119*].

Other worms spiral forever around their origin. The spiral shapes vary enormously: hexagons, diamonds, triangles, stars and asymmetric shapes such as the one shown in the top illustration on the preceding page. "Shoot growers," another class of infinite paths, start in a conventional way, and then the worm falls into a curious spiraling action that keeps repeating itself with regular displacements, creating a rod that shoots off to infinity [*see bottom illustration on preceding page*].

The longest of the known finite paths is shown in the illustration on page 117. The path is generated by a $1_a2_d3_{cbac}4_b$ worm and has 220,142 units. Note the crystalline regularity of the border and the lines that crisscross near the border, and how strikingly this symmetry contrasts with the dishevelment near the center. The point of origin, not identifiable, is at a spot just to the left of the center of the pattern.

One speaks of "known" finite paths because about a dozen worms have paths so long that no one yet knows whether they are finite or infinite. The path generated by $1_b2_a3_{bcaa}4_b$, for instance, was tracked by Beeler to a length of 10 million units without revealing whether the worm dies or goes on gorging itself forever.

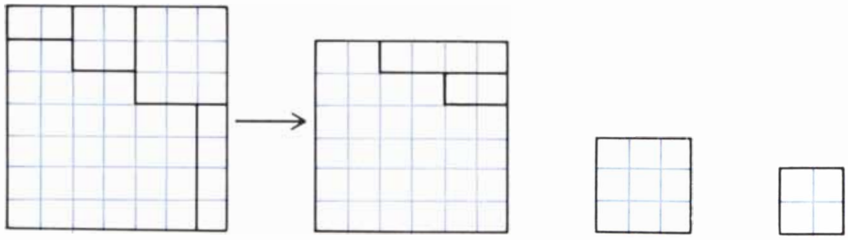
Some path patterns have outlines as unstructured as a cloud [*see top illustration on opposite page*]. Others, such as the "superdoily," display the rigid sixfold symmetry of a snow crystal [*see bottom illustration on opposite page*]. Note the six-pointed star in the center. Occasionally one such star or more appear as randomly situated white spots inside a gray plenum of closely packed unit triangles. Sometimes the stars are separated; sometimes they are partially merged to form binary or ternary systems. The situation is not unlike that of the physical world, in which low-order mathematical laws generate a rich assortment of both patterned and unpatterned objects, as well as objects in which order and disorder, symmetry and asymmetry are crazily mixed.

Beeler's work clearly has just scratched the surface of isometric-worm pathology. His program investigated only the simplest genus, with no attempt to complicate the genetic rules or to find out what happens when simple worms interact with barriers, boundaries or other worms of the same or different species.

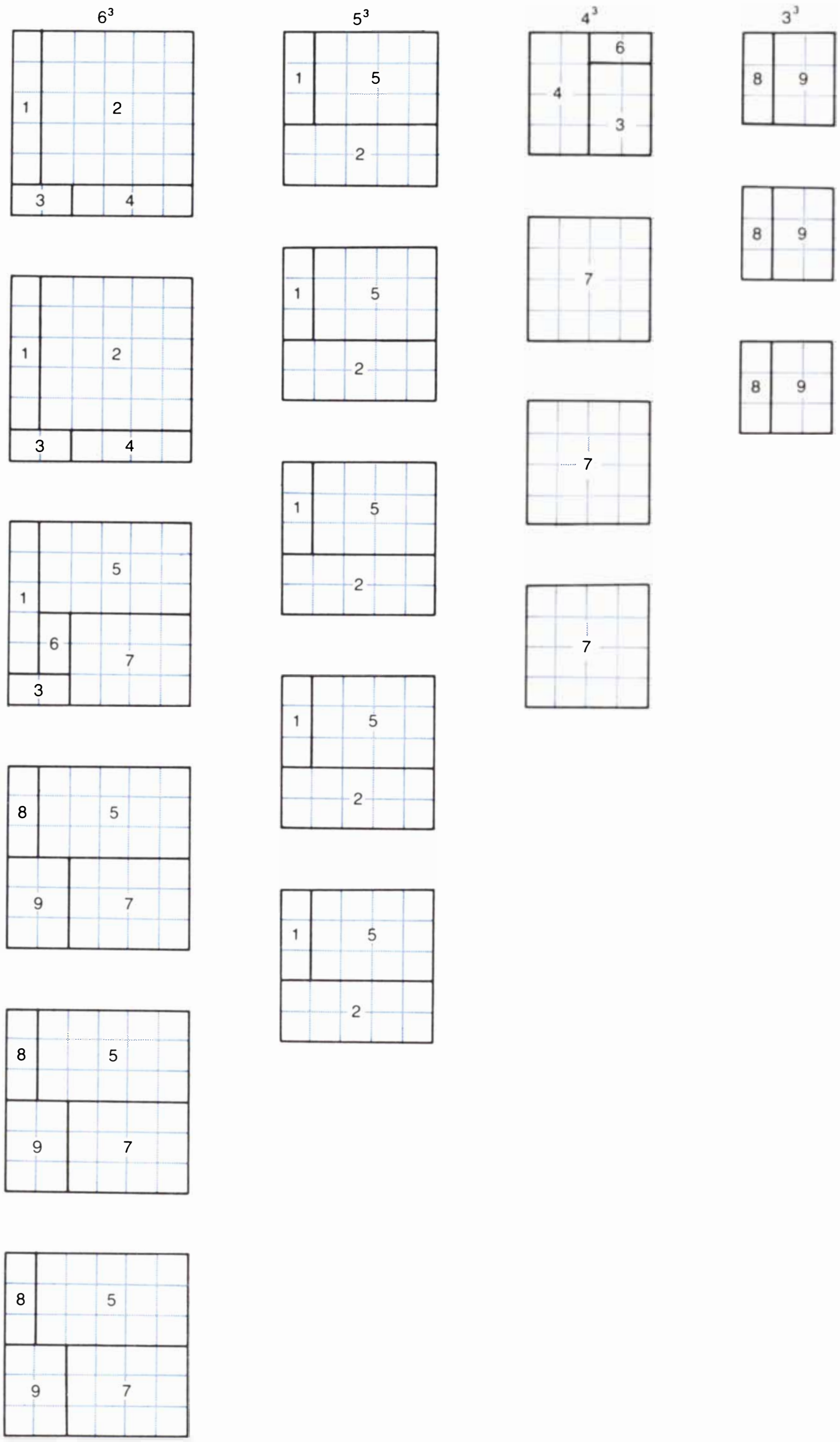
On August 13 *Newsweek* ran a story about Papert's turtle. A girl who had ranked near the bottom of her mathematics class was programming the turtle to draw a certain design. "That math must be fun," said a passing visitor. "There ain't *nothin'* fun in math," the girl replied. She had no idea she was doing math, reported *Newsweek*, and Papert saw no reason to tell her.

Last month's problem was to improve on a dissection of the 7-square into six pieces that form squares of sides 2, 3 and 6. A five-piece solution is shown in the top illustration at right. This dissection has the minimal cutting length of 16 units.

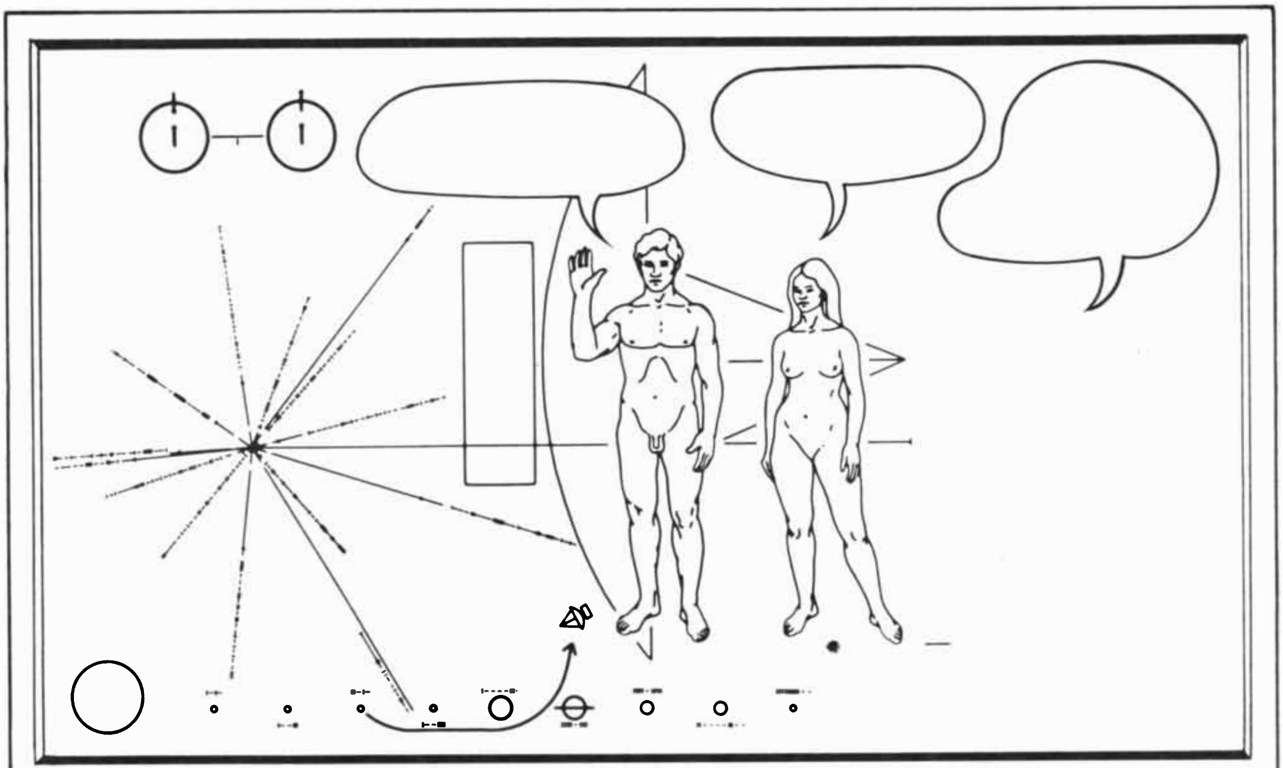
The bottom illustration at the right shows in six cross sections how Thomas H. O'Beirne sliced a 6-cube into nine rectangular blocks (the minimum) that can be reassembled to make separate cubes of sides 3, 4 and 5. No two "bricks" in this remarkable dissection are alike. O'Beirne has shown that a dissection in eight blocks is not possible, but it is not yet known whether or not the nine-piece dissection is unique.



Five-piece dissection for $7^2 = 6^2 + 3^2 + 2^2$



Nine-block dissection for $6^3 = 5^3 + 4^3 + 3^3$



TRW ENCOUNTER CONTEST

Are you witty, original, brilliant? Of course you are. Would a dummy be reading an ad like this? Here's a chance to demonstrate your creative prowess and perhaps even to win one of the many prizes we are about to bestow on bright people like you.

On December 3 of this year, the Pioneer 10 spacecraft (which TRW built for NASA-Ames Research Center) will encounter the giant planet Jupiter, half a billion miles from earth. Jupiter is so gigantic (it contains 70% of the planetary mass of the solar system) that it will fling Pioneer 10 up and out of the solar system.

Because Pioneer 10 will be the first man-made object to leave the solar system, it carries a plaque (see above) with a message to other forms of life which might exist in space. (One current scientific guess is that there are at least a billion such places in the universe.) The plaque tells where and when the spacecraft originated and shows the relative size and appearances of man and woman.

Whom might they meet? What might they say? In the balloons we have added to the plaque, ***write in your proposed conversation*** between man, woman, and the creature they might encounter. In the space provided ***draw the creature.*** (You may use call-outs to identify its parts.)

MAIL YOUR ENTRY TO US here at Space Park postmarked before midnight on NOVEMBER 15th. Entries will be judged on the basis of originality, imaginativeness, and humor. The decisions of the judges are final and all entries become the property of TRW Systems. No TRW employees or their immediate families are eligible. Contest void where prohibited by law. See the list of prizes (below) and mail in your entry **TODAY,** you clever person.

1ST PRIZE
$1,000.00 DOLLARS

2ND PRIZES
50 specially mounted replicas of the Pioneer 10 plaque

3RD PRIZES
100 Fisher Space pens, used by the Apollo astronauts to write under zero-g conditions.

TRW SYSTEMS GROUP

OFFICIAL TRW ENCOUNTER CONTEST ENTRY BLANK

Name (print carefully) ________________

Address ________________

City ________ State ______ Zip ______

One Space Park E2/9043 • Redondo Beach, California 90278

THE AMATEUR SCIENTIST

A sensitive mercury tiltmeter that serves as a seismometer

Conducted by C. L. Stong

Although no one has learned how to predict when an earthquake will occur, Japanese investigators have suggested that in some instances the surface of the earth may tilt slightly along a fault line somewhat in advance of a seismic event. Various seismologists are exploring the idea with an instrument known as a tiltmeter. In one of several possible forms the sensing elements of the device include two cylindrical cups of mercury spaced a few feet apart on a rigid, horizontal base. The cups are interconnected at the bottom by a length of metal pipe. The air spaces above the mercury are similarly interconnected by plastic tubing.

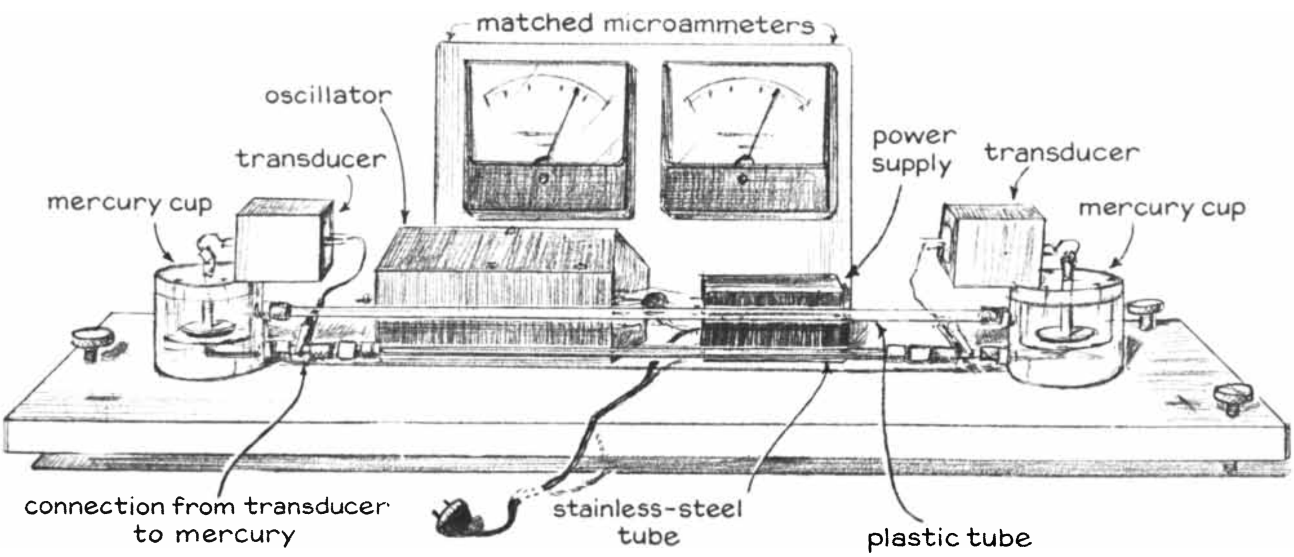
When the assembly tilts like a seesaw, some of the mercury drains from the upper cup into the lower cup. The relative change in the level of the mercury is measured continuously by an electrical circuit in which the surfaces of the metal in the cups function as the movable plate of a differential capacitor. One simple version of the instrument can detect a tilt of only .002 second of arc, which is about the angle that would be subtended by a dime at a distance of 500 miles.

A tiltmeter of this kind can also function as an inexpensive seismometer. The device is particularly sensitive to earthquake waves that vibrate in the direction of the tube that connects the cups of mercury. The acceleration of the tube and the cups causes metal to flow from one cup into the other. In effect the mercury reacts mechanically like the pendulum of a conventional seismometer, but it is far more stable and easier to adjust. For these reasons it is ideal for amateur use. The construction of such an instrument and some of the seismic effects it is designed to measure are discussed by William W. Gile, a research engineer on the staff of the Seismological Laboratory of the California Institute of Technology:

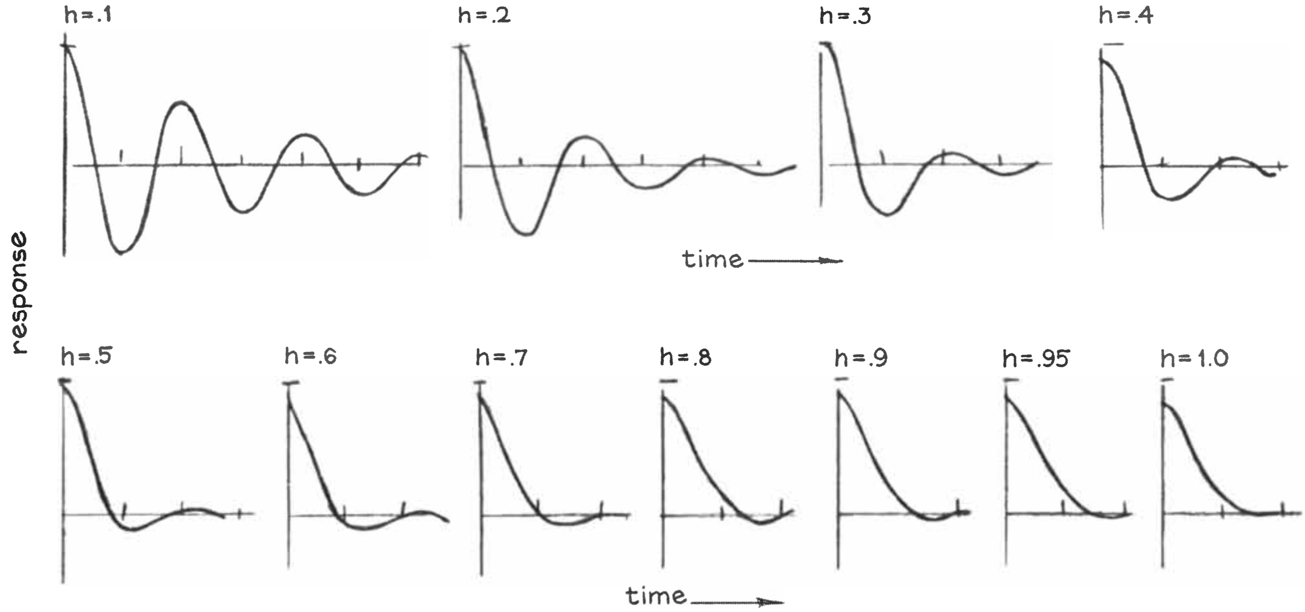
"In southern California the San Andreas fault is mentioned more frequently in daily conversation than the weather, although we can control the behavior of neither. On the seaward side of the San Andreas fault the crust of the earth appears to be creeping northward at a rate of about two inches per year. The underlying rock bends in response to the movement. Ultimately it snaps and the San Andreas fault lurches.

"The readjustment, in terms of feet and inches, can be impressive. For example, in 1857 a circular sheep corral about 20 feet in diameter straddled the fault near our part of the state. On January 8 of that year the accumulated stresses at the sides of the big crack reached the critical limit. Somewhere a rock fractured. When the dust settled, the sheep corral had the form of the letter S. During the subsequent 116 years stresses have again accumulated. If the relative motion along the crack in our vicinity is indeed two inches per year, the slabs of rock must now have a bend of about 19 feet.

"For such reasons earthquake prediction has become a focus of attention in many branches of geophysics and seismology. A number of prediction techniques have shown promise. It must be emphasized that although no technique exists for predicting earthquakes, the present findings do not exclude the possibility of developing one. It is this real



The tiltmeter-seismometer devised by a group at the California Institute of Technology



Progressive increases in h, *the damping factor*

but slender hope that continuously renews the enthusiasm of those of us who work in the field of seismic instrumentation.

"The effort has been rewarded in recent years by the development of some remarkably sensitive apparatus. For example, strain gages now routinely measure forces equivalent to those that would be developed if the width of the U.S. were increased by the thickness of a human hair, and temperature probes for use in deep holes can measure thermal variations across fault lines that amount to less than .001 degree Celsius. Finally, we have long-baseline tiltmeters that can detect angular changes as small as 10^{-12} radian, which is about the angle that the period at the end of this sentence would subtend if it were placed on the moon and viewed from the earth.

"Some of the devices are relatively inexpensive and easy to adjust and to maintain. The short-baseline tiltmeter [*see illustration on opposite page*] is an example. We have operated versions of the instrument on the island of Hawaii to measure changes in the tilt of the earth's crust caused by the eruption of Halemaumau volcano. The instrument also works well as a sensitive seismometer. When it is combined with an appropriately adjusted pen recorder, it can also be used for monitoring earth tides: the daily rise and fall of the crust in response to tidal forces exerted by the moon and sun. The low cost of the apparatus and its ease of construction, together with its high stability and sensitivity, make the instrument ideal for both amateurs and professionals.

"Conventional seismometers of the pendulum type depend on critical adjustments that, following seismic disturbance, determine the rate at which the pendulum swings, the number of oscillations it completes before coming to rest and the position at which it stops. Although the mercury of the tiltmeter is analogous to a conventional pendulum, the instrument does not require these tedious adjustments. The period of its oscillation and the damping (the rate at which it stops oscillating after a disturbance) are determined by the physical properties of the mercury, together with the size of the cups and the interconnecting tubing. These dimensions do not change significantly with time. Accordingly the instrument is stable.

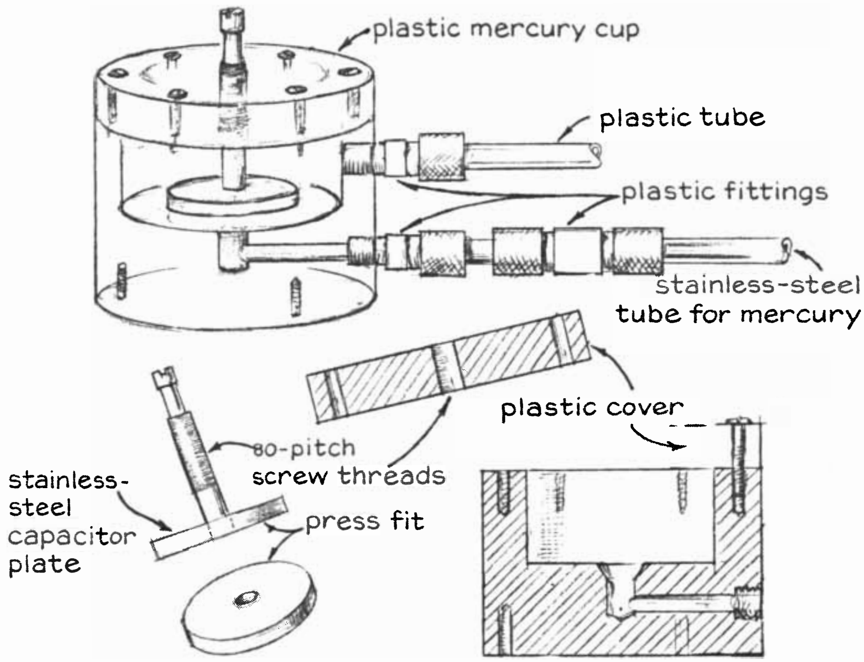
"A normal rule of thumb in seismic instrumentation is that the period of the pendulum—the interval required for the pendulum to make one full excursion in both directions—must exceed the period of the longest seismic wave that is to be observed. The necessity for the long period is explained by the role of the pendulum in measuring the length of seismic vibrations. In effect the bob of a pendulum will stand still, even though the framework from which it is suspended vibrates, if the period of the pendulum is long with respect to the period of the vibration. In other words, the stationary mass of the bob serves as a reference for measuring the relative motion of the earth's crust. The principle can be demonstrated by an interesting experiment.

"Make a pendulum by tying a foot or so of string to any small weight. Hold the string in one hand. Dangle the weight about 10 inches below the point where you grasp the string. Jiggle your hand back and forth an inch or so at the rate of about five vibrations per second. Observe that the weight stands fairly still. If your hand were the earth and the weight were the bob of a seismograph pendulum, the earth's relative motion could be observed just as readily.

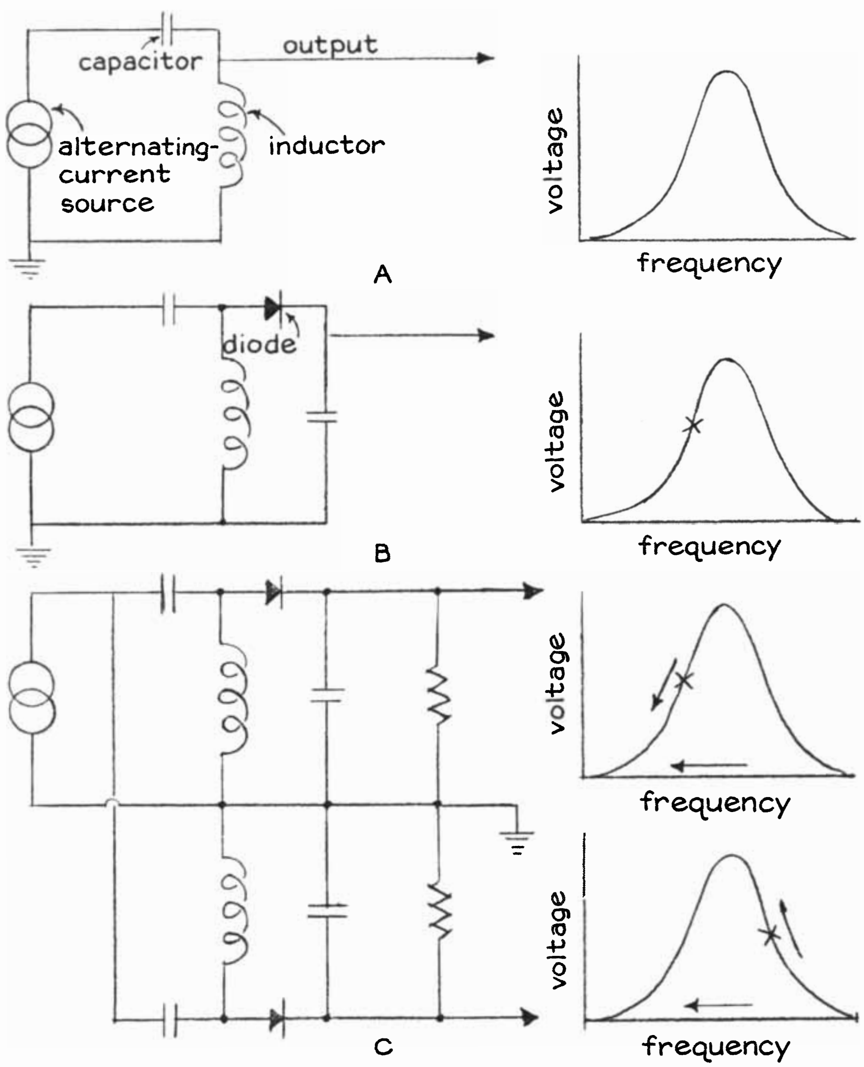
"Now decrease the rate at which you shake your hand to about one second per cycle. Within a very few vibrations of your hand the weight will be swinging back and forth violently. The reason is that one second is about the period at which a 10-inch pendulum vibrates naturally, that is, one second is the frequency of vibration to which the pendulum is resonant. If the pendulum of a seismograph were tuned to the period of the earth's vibration, it would yield a grossly exaggerated measurement of the seismic event.

"Stop moving your hand. Observe that the pendulum continues to swing for a time but with diminishing amplitude. It is said to be poorly damped. The damping effect results from the suppression of a pendulum's motion by the loss of energy through friction. A pendulum that comes to rest after about one vibration is said to be critically damped. Good seismographs of the type that employ mechanical pendulums must be fitted with adjustable gadgets for dissipating unwanted energy.

"Finally, move your hand back and forth slowly, at a rate of about 10 seconds per cycle. Observe that the motion of the weight follows the motion of your



Assembly of a mercury cup



Evolution of the basic discriminator circuit

hand. If the pendulum of a seismograph were similarly adjusted, it would be insensitive to seismic vibrations of long period. Indeed, in the case of a displacement seismograph, as is illustrated by this experiment, sensitivity declines at the rate of 12 decibels per octave for seismic vibrations of lower frequency than those of the freely vibrating pendulum. That is to say, each time the seismic period doubles, the pendulum moves only a fourth as much with respect to its motion at the free period.

"The earth has one free period of about 53 minutes per cycle. This is the frequency at which the earth 'rings like a bell' following a big earthquake. A conventional pendulum of this period would need to be more than 1,500 miles long. The period is equal to 6.28 × $(l/32)^{1/2}$, in which l is the length of the pendulum in feet and the period is in seconds.

"As I have mentioned, the mercury of the short-baseline seismometer is analogous to a pendulum. Its free period is determined by the inside diameter of the cups and of the interconnecting tubing and by the distance between the cups. To find the period in seconds, divide the inside diameter of the cups as expressed in feet by the inside diameter of the tubing and multiply the quotient by 6.28. Next divide the center-to-center distance between the cups by 64. Find the square root of the quotient. Multiply the square root by the product of the preceding calculation to find the period. The procedure is expressed by the formula $T = 2\pi D_1/D_2(L/2g)^{1/2}$, in which T is the period in seconds, $\pi = 3.1416$, D_1 is the inside diameter of the cups, D_2 is the inside diameter of the interconnecting tubing, L is the center-to-center distance between the cups and g, the acceleration of gravity, has the value 980 if lengths are expressed in centimeters and 32 if they are expressed in feet.

"An instrument of convenient size and excellent performance that is particularly appropriate for amateur construction would measure two feet from center to center of cups having an inside diameter of two inches. The cups would be interconnected by a stainless-steel tube with an inside diameter of 3/32 inch. The period of the recommended seismometer can be calculated by substituting these dimensions in the formula: $T = 2 \times 3.1416 \times 2/(3/32) \times (2/2 \times 32)^{1/2} = 6.28 \times (64/3) \times (2/64)^{1/2} = 6.28 \times 21.33 \times .177 = 23.7$ seconds.

"Why not make the cups wider with respect to the diameter of the interconnecting tube and thereby achieve a period of an hour or more? In theory the period could be thus increased without limit. As a practical matter, however, the mercury moves less in the end tanks as the ratio of diameters is increased. The sensitivity of the instrument to long seismic waves does not increase with the period. Instead its sensitivity decreases to vibrations of shorter period.

"It is useful to know the method of calculating the damping factor of the pendulum. As I have mentioned, the term damping suggests the tendency of a pendulum to vibrate at a diminishing amplitude after it has been set in motion. The amount by which a pendulum is damped is expressed by a factor h that approaches zero in the extreme case of an undamped pendulum and unity at the other extreme when the pendulum returns to its equilibrium or mid-position in exactly half the interval of its free period [*see illustration on page 125*].

"In the case of the mercury pendulum the damping factor h is equal to $2uT/d\pi(D_2/2)^2$, in which $u = .01554$ (the viscosity of mercury in poise units), T is the period of the pendulum in seconds, $d = 13.55$ (the density of mercury in grams per milliliter) and D_2 is the inside diameter of the interconnecting tube in centimeters. The damping factor of the recommended seismometer can be calculated by substituting appropriate quantities in the formula: $h = 2 \times .01554 \times 23.7/13.55 \times 3.1416 \times (.238/2)^2 = 1.22$. The pendulum is somewhat overdamped, which is to say that it stops oscillating in less time than is required for it to complete one full cycle of vibration.

"The proportions of the instrument can be varied to alter both the period and the damping factor. The damping factor of most professional instruments ranges between .7 and 1. Calculations indicate that a period of 25 seconds and a damping factor of .73 can be achieved by interconnecting the two-inch cups at a center-to-center distance of four feet with stainless-steel tubing of 1/8 inch inside diameter. Experimenters can manipulate the dimensions to achieve still other characteristics.

"The cylindrical cups can be made of clear plastic. If the experimenter does not have access to a lathe, the cups can be fabricated of flat stock and of tubing with a thick wall [*see top illustration on opposite page*]. The cups must be cylindrical in form, not rectangular, to prevent the development of nonlinear turbulence. Nonlinear turbulence is not canceled differentially. Therefore it adds to the 'noise' of the system.

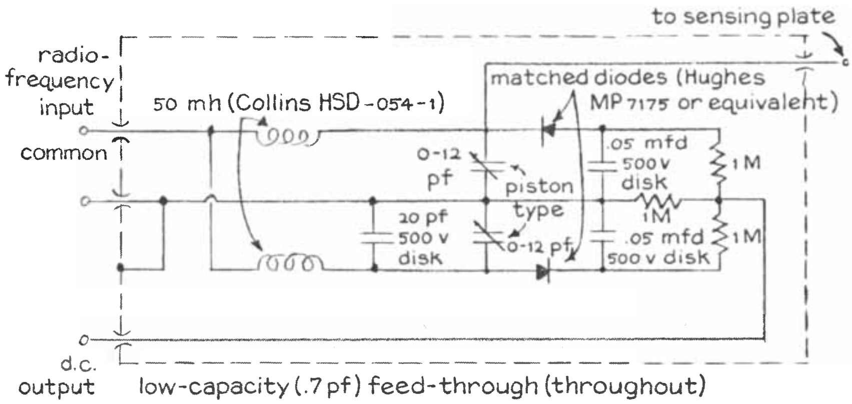
"Stainless-steel tubing interconnects the cups. This material was selected to provide an electrical contact with the mercury while avoiding an amalgamating reaction at the ohmic contact. The second tube, which is situated above the mercury level, equalizes gas pressure in the tanks. This tube can be of plastic.

"I flood the tanks of long-baseline instruments with nitrogen to prevent the mercury surface from oxidizing. Incidentally, the mercury cups of larger instruments have baselines of 30 meters (98.4 feet). They are interconnected by butyrate plastic. The coefficient of thermal expansion of this material approximates that of mercury. Electrical contact with the metal is provided by stainless-steel fittings that couple the plastic tubing to the cups.

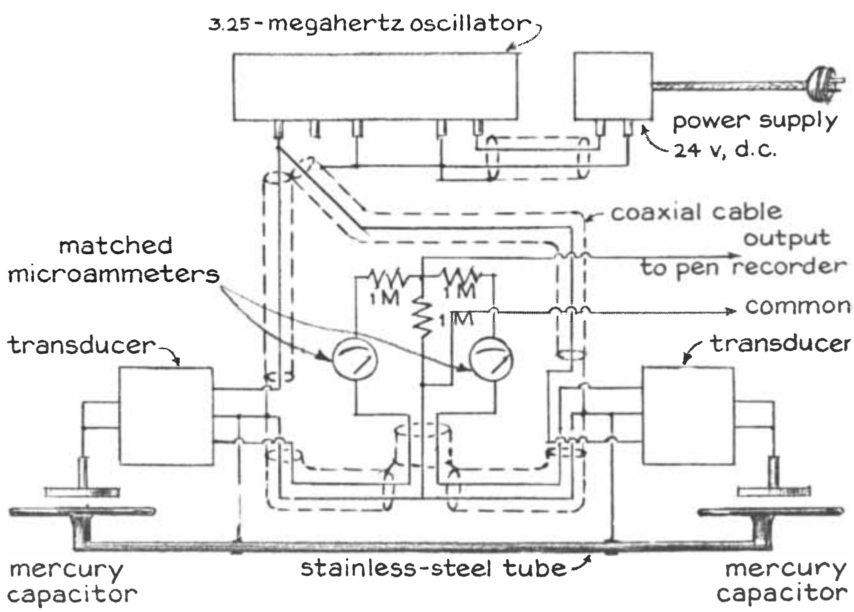
"A fixed electrode in the form of a polished disk of stainless steel is supported above the surface of the mercury in each cup by a threaded shaft that passes through the center of the lid. These electrodes constitute the fixed or sensing plates of the mercury-pendulum capacitor. The surface of the sensing electrode that faces the mercury must make a true right angle with respect to the axis of its supporting shaft.

"The shaft also functions as a screw for adjusting the gap between the surface of the electrode and the surface of the mercury. The threads must be as fine as the experimenter can cut. I use 80 threads per inch. If you do not have access to an engine lathe, have the rods threaded to this pitch by a local machine shop.

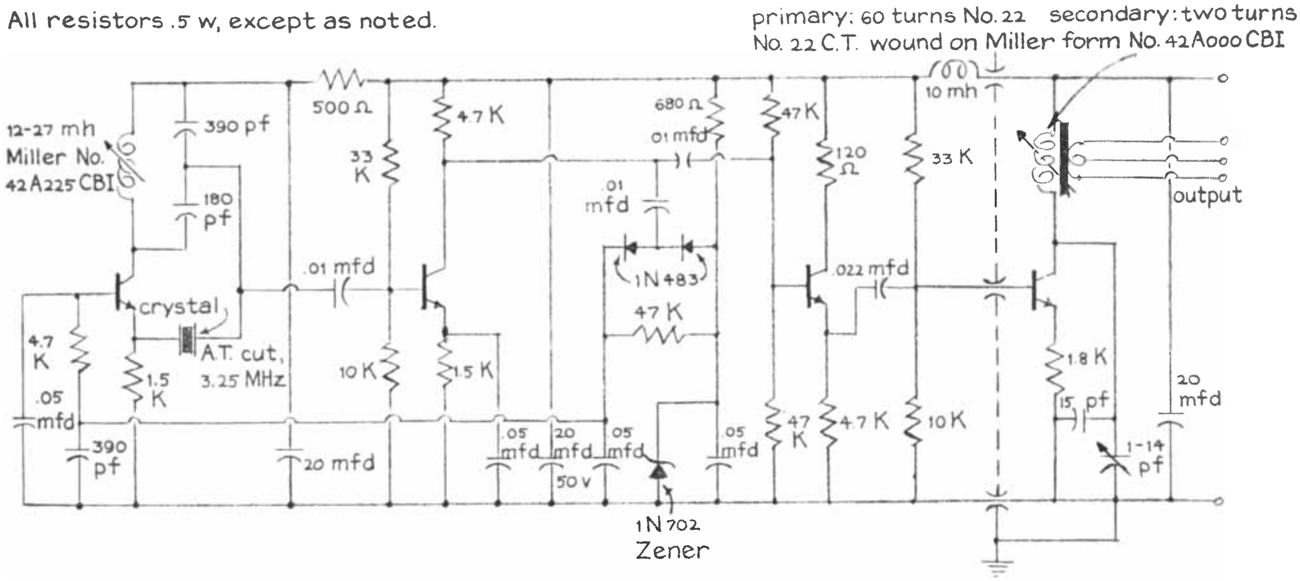
"If the cover of the cup is made of a slab of half-inch plastic and drilled about .001 inch narrower than the diameter of the thread, it will not be necessary to tap the hole. The rod will cut a shallow thread of adequate strength in the plastic. The base, provided with leveling screws for supporting the instrument,



Circuitry of the inverse discriminator



Wiring diagram of the tiltmeter-seismometer



Circuitry of 3.25-megahertz crystal oscillator

can be improvised from any rigid aluminum section, such as the rectangular aluminum tubing that is commonly used as door framing in office buildings. Adequate lengths of the material are usually available at low cost from dealers in aluminum scrap.

"An electronic transducer, which I refer to as an 'inverse' discriminator, converts the displacement of mercury into a correspondingly varying signal voltage. The voltage can be used to operate a pen recorder of the kind previously described in this space [see "The Amateur Scientist," SCIENTIFIC AMERICAN, March, 1972]. Essentially the transducer consists of a circuit similar to the one used in radio sets for detecting frequency-modulated signals. Such circuits are known as discriminators. The circuit of a conventional discriminator includes a fixed capacitor and a coil that help to detect signals carried by a variable frequency. My device works in reverse: an electric current of fixed frequency helps to detect 'signals' that are represented by the minute motion of a differential capacitor—the mercury pendulum.

"Discriminator circuits are essentially resonant systems. They are analogous to pendulums, the strings of musical instruments and similar mechanisms that vibrate at maximum amplitude when they are tuned to a preferred period. The basic principle is illustrated by an electrical circuit consisting of a capacitor and an inductor. A circuit of this kind is naturally resonant at some frequency. When a quantity of electrons is deposited on one plate of the capacitor, they immediately flow through the inductor, pile up on the other plate of the capacitor and return to the first plate for another round trip, just as a pendulum, when it is given a push, vibrates continuously at a characteristic rate and for exactly analogous reasons.

"The inductor of such circuits usually consists of a small coil of insulated wire. Assume that the capacitor and the inductor are connected in series to a source of current that alternates at a frequency lower than the frequency at which the circuit is resonant and that the frequency of the source is gradually increased until it is higher than the resonant frequency. Assume also that a method is provided for measuring voltage that appears across the coil. The experimenter will observe that voltage across the coil is minimum when the frequency of the imposed current is lower than the resonant frequency of the circuit, that the voltage becomes maximum at resonance and that it declines as the frequency of the source exceeds the resonant frequency of the circuit [*see bottom illustration on page 126*].

"The alternating potential across the coil can be converted to unidirectional voltage at the output terminal of the device by inserting a diode in the circuit, and it can be further converted into a reasonably constant potential by connecting in the circuit a second capacitor that acts as a filter. This simple device is known as a slope detector. It was used in inexpensive radio and television sets for detecting frequency-modulation signals. In that application the slope detector was tuned to a point such that the center frequency of the radio station was located about halfway up the resonant characteristic of the circuit, as indicated by the *X* on the graph of the detector. Excursions of the imposed frequency from the midpoint appeared as proportional excursions of voltage at the output of the slope detector. The voltage and the linearity available at the output of the device can be improved by connecting two of the circuits back to back.

"My inverse discriminator works in reverse. Alternating current of fixed frequency is imposed across a circuit of variable resonance. The signal consists of the movement of a differential variable capacitor in the form of the mercury pendulum. I tune the transducers to 50 percent of the peak resonant output. For an input potential of one volt peak to peak the peak output of each transducer is about 30 volts. The outputs of the transducers are detected, filtered and combined through three resistors to provide a single-end output. Note that in the practical circuit each transducer circuit is shunted by a trimming capacitor of the piston type. Minor variations in the characteristics of the coils and stray capacitances can be compensated for by adjusting the trimmers.

"The constant frequency is generated by a crystal-controlled oscillator. In this circuit one can use any silicon transistor that is rated for a collector-to-emitter potential of 45 volts and a dissipation of .3 watt. The collector should be rated at .05 ampere or more, and the small-signal, common-emitter current-transfer ratio (h_{fe}) should be about two at a frequency of 30 megahertz. Surplus catalogues list many transistors that meet or exceed these specifications. Experimenters who are not familiar with the techniques of constructing resonant circuits and transistor oscillators may want to enlist the cooperation of a radio ham.

"Install the completed instrument on a solid foundation in an environment of the most constant temperature available. Our seismographs and tiltmeters operate in tunnels several hundred feet underground and are provided with servomechanisms that automatically compensate for temperature changes. Nonetheless, an instrument of remarkable performance can be installed in a good basement, particularly if it is insulated in a styrofoam box with thick walls. Use clean mercury of National Formulary grade or better. About six ounces of mercury will be required for an instrument 24 inches long.

"Lift one end of the instrument about a foot to flush air bubbles from the mercury in the interconnecting tubing, and lower it quickly. Adjust the base to the horizontal position as accurately as possible with the aid of a bull's-eye level. Rotate the piston-type trimmer capacitors of the transducers to minimum capacitance. Install a jumper between the junction of the three one-megohm resistors and the common point.

"Switch on the power supply to the oscillator. By means of the adjustment screw decrease the air gap between the mercury level and the sensing plate of the capacitor until the corresponding meter indicates approximately half a microampere. Similarly adjust the other mercury cup.

"Adjust the piston-trimming capacitors to the point where each meter indicates 3.5 microamperes. Replace the jumper with a high-impedance voltmeter. The meter should indicate zero volts. If the meter indicates an offset voltage, rotate either one of the sensing-capacitor plates to the point of zero indication. At this adjustment the instrument is essentially balanced.

"Finally, rotate one of the sensing-capacitor plates a full turn. The voltmeter should now indicate a potential of 17.36 millivolts. If the output exceeds this potential, the sensitivity of the instrument is too high. Increase the air gap in both cups between the surface of the mercury and the sensing plates. Next increase the capacitance of the piston capacitors to the point where the meters again indicate 3.5 microamperes. If the output voltage is less than 17.36 millivolts, which indicates inadequate sensitivity, reverse the procedure. If the instrument is to be used exclusively as a seismometer and the highest possible sensitivity is wanted, adjust the piston capacitors to minimum capacity and the sensing capacitors to maximum."

GET THE BRAND NEW ONES–OR CATCH UP ON THE FEW GREAT ONES YOU'VE MISSED

by starting your trial membership now in

THE LIBRARY OF SCIENCE

It's the easy—and economical—way to keep up with the very best in scientific reading and reference. More, too. Books to think about. Books for sheer enjoyment. All here first. In the one scientific book club that covers *every* discipline in the expanding world of science.

(retail prices shown)

44190. ELEMENTARY PARTICLES AND THEIR CURRENTS. *Jeremy Bernstein.* Brilliant new and unified text on elementary particles, concentrating on currents and interactions. **$12.00**

62801. MODERN COSMOLOGY. *D. W. Sciama.* "This book is a masterpiece. From the standpoint of readability, orderliness, continuity, evenness of level, and its ability to hold one's attention, the book is three standard deviations above the mean."—*Physics Today* **$9.95**

56320. INTRODUCTION TO CALCULUS AND ANALYSIS. *Richard Courant and Fritz John.* A lucid and up-to-date introduction stressing the interaction between mathematical analysis and its applications, as well as the reciprocity between differential and integral calculus. **$14.95**

63131. MOLECULES AND LIFE. *Joseph S. Fruton.* The interplay of chemical and biological ideas, with full documentation from about 1800 to 1950, with reference to earlier and later developments. *Counts as 2 of your 3 choices.* **$19.95**

55561. THE INSECT SOCIETIES. *Edward O. Wilson.* "For anyone, layman or specialist, interested in a single, concise, lucid, and authoritative account of the most significant facts and theories about insect societies . . ."—*Natural History. Counts as 2 of your 3 choices.* **$20.00**

61360. MATHEMATICAL STATISTICS WITH APPLICATIONS. *William Mendenhall and Richard L. Scheaffer.* A uniquely practical work which stresses inference as the objective of statistics and emphasizes the role of probability in inference making. **$12.95**

50291. A GOD WITHIN. *René Dubos.* Moving testament to man's destiny. Particularly for ecologists and environmentalists, it creates a new, scientific theology which man must have to survive. **$8.95**

58830. THE LONG AFRICAN DAY. *Norman Myers.* This big, handsome book may well be the most powerful account ever published on the state of animal life in Africa. Captures the rich beauty of a majestic continent. *Counts as 2 of your 3 choices.* **$25.00**

84451. TOWARD A GENERAL SCIENCE OF VIABLE SYSTEMS. *Arthur S. Iberall.* Introduces a science that attempts to unify our understanding of all autonomous systems—natural, organic or man-made—proposing a continuum ranging from nuclear particles at one end of the scale to the cosmos at the other. **$14.50**

88020. X-RAYING THE PHARAOHS. *James E. Harris and Kent R. Weeks.* Not since Howard Carter opened the tomb of Tutankhamen, a half-century ago, has science made archaeological discoveries to equal those revealed for the first time in this handsome new volume. Illustrated in color and black and white. **$10.00**

45471. ENVIRONMENTAL GEOSCIENCE. *Arthur N. Strahler and Alan H. Strahler.* Explores the physical and biological processes involved in interactions between Man and the environment of energy and matter which surrounds him. **$12.95**

Take any 3 books

(values as high as $47.85)

all for only $3.95

if you will join now and agree to take only 3 more books—all at handsome discount—over the next 12 months.

The Library of Science, Riverside, New Jersey 08075 2-50Y

Please accept my application for membership and send me the three volumes indicated billing me only $3.95 for all three. I agree to purchase at least three additional Selections or Alternates during the first year I am a member, at special members' prices. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments.

Send no money. Members are billed when books arrive.

Indicate by number the 3 books you want.

[ ] [ ] [ ]

Some expensive books count as 2 selections.

Name____________________

Address____________________

City__________ State________ Zip________

Book selections purchased for professional purposes may be a tax-deductible expense.

48400. THE FASCINATION OF REPTILES. *Maurice Richardson.* A colorful and handsomely illustrated appreciation of these scientifically neglected and often maligned animals. Includes the care and feeding of reptiles. **$10.00**

53870. THE HUDSON RIVER: A NATURAL AND UNNATURAL HISTORY. *Robert H. Boyle.* Best-selling account of the destruction of one of America's mighty rivers. An eye-opening book. **$7.95**

68730. PHYSICS OF THE EARTH. *Frank D. Stacey.* Offers a wide range of new concepts for everyone who wishes to know more of the earth sciences. **$12.95**

36280. A BIOGRAPHICAL DICTIONARY OF SCIENTISTS. *Edited by Trevor I. Williams.* Over 1000 biographies of scientists in every field, from agriculture to zoology, their lives, their contributions and their relation to contemporaries, successors. **$9.95**

33430. THE ANALYSIS OF INFORMATION SYSTEMS. Revised Edition. *Charles T. Meadow.* Offers an all-inclusive treatment of the four aspects of information processing and retrieval. **$14.95**

38310. CLASSICAL ELECTRODYNAMICS. *John David Jackson.* Treats the subect of electromagnetism in a fairly conventional, but thorough, fashion, showing the unity of the subject in both its physical and mathematical aspects. Classic work adopted by 201 institutions of higher learning. **$17.95**

69930. POPULATIONS, SPECIES, AND EVOLUTION. *Ernst Mayr.* A brilliant abridgment and updating of the monumental 1963 classic, *Animal Species and Evolution.* **$10.00**

44350. ELEMENTS OF CARTOGRAPHY. Third Edition. *Arthur H. Robinson and Randall D. Sale.* Completely revised edition includes a wealth of new material, techniques of air-mapping, and the most recent experiments in design and reproduction. **$12.75**

HOW THE CLUB OPERATES. • The Book Club **News,** describing the coming Main Selection and Alternate Selections, will be sent to you 15 times a year at three to four week intervals. • If you wish to purchase the Main Selection, do nothing and it will be shipped to you automatically. • If you prefer one of the Alternates, or no book at all, simply indicate your decision on the reply form always enclosed with the **News** and mail it so we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News,** you should ever receive a Main Selection without having had the 10-day consideration period, that Selection may be returned at Club expense. • After completing your trial membership, you will be entitled to take advantage of our fabulous bonus plan which offers savings up to 70% and more. • (Offer good in Continental United States and Canada only. Prices are slightly higher in Canada.)

BOOKS

Are flies really guilty of spreading disease?

by Philip Morrison

FLIES AND DISEASE, VOLUME I: ECOLOGY, CLASSIFICATION AND BIOTIC ASSOCIATIONS, by Bernard Greenberg ($30); VOLUME II: BIOLOGY AND DISEASE TRANSMISSION, by Bernard Greenberg ($18). Princeton University Press. The order of the two-winged insects includes some 100,000 species of flies, mosquitoes and gnats. Fifteen full-page color paintings in the first volume almost persuade one that archangels too belong to the Diptera; these fly portraits, floating without scale or context on a pure white page and rendered bristle by bristle in loving and iconic detail, are as glowing, hieratic and majestic as Byzantine church art. They are the work of František Gregor of the Czechoslovak Academy of Science in Brno, who, with two Brno colleagues and an Englishman, an Australian and a Texan, has contributed to the first thick volume of this definitive work, assembled by the Chicago biologist who is the author-editor.

Tens of millenniums ago human beings created a specific animal community. The "domestic animals [are] its products" and flies, mice and rats are "its spontaneous members." The 346 species of flies named in the longest lists here are such members, more or less intimately living with us, held between toleration and war, and the 15 portrayed display the most common or interesting of these forms worldwide. The subject is an old one: cylinder seals of 3000 B.C. depict flies, perhaps as "symbols...to ward off evil"; a text of Hammurabi's time (based on Sumerian sources that were ancient even then) mentions about 10 fly forms, most of them quite identifiable today. The story is by no means ended: In 1952 for the first recorded time a European face fly was taken in North America, which it may first have viewed "through the porthole of an airplane"; by 1967 the face fly had spread from coast to coast in both Canada and the U.S. Its larvae breed in fresh cow pats and possibly in bison and swine feces as well; most other fecal droppings are too coarse or too dry. Expectant female face flies feed on the watery secretions of the eyes and nostrils of cattle and on the bloody wounds inflicted by other biting flies, and annoy us biped cattle rearers as well; the males appear to feed mainly on flower nectar, like many of their mosquito kindred.

The evolutionary story seems clear. The true housefly, the cosmopolitan *Musca domestica domestica,* was a highly plastic species, adapted to feeding on the feces of herds of large ungulates. From the wild horses and cattle *Homo* bred the tame ones, and the flies then moved from pasture to stable to house, finally adapting (with some 10 generations per year) to live on the food and wastes of the human household. Their ability to overwinter in all stages and to develop at room temperature is unique among flies; although housefly populations are decimated in a temperate winter, some individuals always survive. (This treatise explicitly elides detail in the case of the housefly, since the literature is vast. Luther West's *The Housefly,* a book of 1951, was a stimulus to Greenberg.)

Consider the lesser housefly, *Fannia canicularis.* Its larvae prefer damper sites, such as excreta wet with urine; the adults abound in our houses from Samarkand to Massachusetts (the curves are here), particularly in early and late summer. The males show a hovering, zigzagging "irregular dancing flight" around ceiling fixtures and wall projections; they alight mainly on ceilings and walls. The green bottle fly too has followed man, to become dominant in African and Australian cities and suburbs; it is most common over the temperate Northern Hemisphere. It frequents garbage of all kinds—and now dog feces in cities—and also feeds on foliage and flowers. It enters houses to alight mainly on food rather than on walls. The loud-buzzing, slow-flying bluebottle fly, abundant in the Middle West in spring and fall, is a carrion breeder and feeder. The stable fly, a worldwide biting pest of livestock (you can see its five-toothed proboscis here in a scanning electron micrograph), prefers cattle to man but will bite and suck blood from both. By night it is house- or stablebound; by day it basks in the sunshine on gate and fence. Its maggots prefer rotting plant material, such as windrows of weeds along a shore; the adults of both sexes are well known to bathers in late summer in our country. One conjectures that it developed first in the African Tropics, where the largest numbers of related species are found today. Once stables were contrived to shelter the big herbivores in colder climates, the "sharp-mouthed" *Stomoxys calcitrans* spread with them, and us, to the cool world.

Then there are the tiny eye flies of our South and Southwest, mainly *Hippelates pusio.* Their larvae feed on decaying organic matter, often in newly turned soil where there is intensive cultivation, and hence on lawns, ditches, river edges and the like. Weighing a fiftieth as much as a common housefly, they can fly strongly all day. They find man and beast, approach and land quietly, and then make their way by crawling or by brief flights to the eyes and other watery orifices or to cuts and sores, where they sip gently without biting.

So far we have described the natural history of some synanthropic flies as examples, familiar in everyday life, of the work of that subtle engineer we call natural selection. As the title of these volumes makes plain, however, the learned authors had a more urgent purpose: Do these flies—ubiquitous, mobile, restless and so patently identified with filth—cause disease in man or in his livestock?

One by one, over some 55 diseases of man and animals, the indictment of the varied fly species is drawn up. The little tsetse fly, which bars cattle from the Gambia and the Zambesi valleys and many an African clearing in between, is of course guilty: its bite spreads the trypanosomes that are deadly to cattle and often infect human beings. But it is a creature of the forest edge, not at all a cohabitant with mankind, and it is merely mentioned in this volume. For the rest most of the indictments stand unproved, although in many cases the trial is still in progress. Like all ecology, epidemiology is no simple subject. Sometimes flies abound and disease is rampant. Then both die away; the flies may return but not the disease, or the disease but not the flies. It is easy to demonstrate through direct feeding that flies pick up poliomyelitis virus and no harder to show a strong correlation between the number of infected flies and the presence of the virus in humans. "There is ample evidence that human populations readily infect flies with poliovirus. But we are woefully ignorant whether and to what extent flies return the favor."

For enteric diseases such as typhoid and dysentery the case is stronger. The sentence was passed before World War I: the housefly was dubbed "the typhoid fly." Flies seem to go with the disease worldwide, in hospital ward, regiment and city district. Yet the case is still not always airtight; there are many other potential vehicles. Fly populations are often out of phase with case numbers, even though both grow during the summer. Poor sanitation in general may breed both flies and disease, so that eliminating flies may (or may not) end the disease by chance. Flies are important vectors, but serious outbreaks of these diseases may occur entirely without them. Fair is fair: poverty and the careless handling of the food chain from farm to latrine are grave social faults that lie not in our flies but in ourselves. Actually it is a nematode disease of horses, habronemiasis, that offers the clearest example of the housefly proved a vector!

Even the eye flies cannot yet be convicted: "The critical experiments have not been done." It is not easy to find big doses of infectious microbes on flies; flight kills the external passenger germs, and the inner fly is often bactericidal. Housefly maggots cannot grow without the microbes that teem in their normal rearing medium, yet "at the moment an emerging fly wrestles free of the puparium it may actually be sterile."

These volumes are not meant as accounts for the general reader; about two-thirds of the total text is given over to taxonomic keys, lists of associations of flies with other organisms and a comprehensive bibliography. There is still a lot to read and ponder. The Russian proverb goes: "Fly in April, dead child in July." Certainly no one need advocate the risk, but it is not so simple. It is too bad that the Chinese "away with all pests" experience came too recently to receive treatment in this work. Overall a reader is again sobered by the complexity of the network that links living forms and by the subtlety of the relations among truth, belief and human action.

An Illustrated History of Brain Function, by Edwin Clarke and Kenneth Dewhurst. University of California Press ($14). The visage of a bald-headed man faces the reader directly. It is presented as the outline of a piece of sculpture, but over its surface a map of qualities has been drawn. The left side has some 160 numbered areas, and the right has an alphabetful besides. Just below the corner of the left eye there is a region, half the size of a postage stamp, associated with "wave motion"; at the lower left corner of the jaw there is an area marked "149. Republicanism." This is the preposterous work of a New York physician, J. W. Redfield, "the author of the most elaborate system of physiognomy ever conceived," which was first published at the end of the Civil War and was still current 30 years later. This volume is not a collection of such naïve and bizarre manifestations, although there are more than a few within it; it is a scholarly, visually remarkable pursuit of the growth of knowledge of cerebral localization and its graphic representation.

"Tell me where is fancy bred, / Or in the heart or in the head?" The introspective confusion lasted from pre-Galenic times through Descartes and beyond. Here is a careful text covering the entire period to 1972, illustrated throughout with a wide variety of drawings and photographs. There are Hellenistic manuscripts, medieval Persian texts, an early Dutch transmission of Chinese tradition, a photograph of the brain of Sir William Osler, Descartes's hydraulic reflex circuit via the pineal gland, a fine color reproduction of Rembrandt's *Anatomy Lesson of Dr. Deyman* (in 1656 the artist was more faithful to the brain than the anatomists were), Wilder Penfield's modern homunculus—in sum, nearly 160 figures in a large, thin, well-made volume.

To begin with, it was the ventricles, the cavities in the brain filled with fluid, that were considered the centers of function. Galen had not been wholly consistent but the Church Fathers tidied up his works and made it plain how these "cells" held reason, common sense and imagination. By the time of Leonardo and Vesalius the matter was much less clear, although the anatomy was surer. Vesalius had at first accepted even Galen's pig-and-dog results for the arterial network at the base of the brain, the *rete mirabile,* which is not found in man, but he recanted in the text of his great anatomy.

It was Thomas Willis and his contemporary Sylvius who first proposed (there had been a Greek foreshadowing) that portions of the brain substance itself had specific functions. Certain plates in Willis' 1664 anatomy of the brain are famous; "Dr. Wren, on account of his singular humanity, wherewith he abounds, was pleased to delineate with his own most skilful hands many figures of the Brain and Skull." The cranial nerves and the arterial circle of Willis (which replaces the *rete* in man) are clear. The Royal Society again proved its early vigor, here prenatally.

Still the convolutions of the outer layer of the brain, the cortex, were all but ignored. Descartes drew them rather well, even though he gave them no importance; Wren drew them rather vaguely at best. By 1820 the phrenologist Franz Joseph Gall, an excellent and sincere anatomist, had introduced his localized functions of the brain and the corresponding overlying bumps on the skull. His scheme was initially rather modest, but his followers made it laughable. Their intuitions and bombast were nonetheless the stimulus for real studies of the convolutions of the cortex. At first merely morphologically, then with the insight of Darwinian comparative anatomy, the anatomists mapped the folds and fissures with care and accuracy. Photography entered to add veracity. By 1870 the static map was about complete; by 1881 a diagram of a cortex no less obsessive than Redfield's of the head could be published by a German localizer who had studied brain lesions and the kind of defects the patients showed, but evidently not statistics. The physiologists and the neurosurgeons began to work on the living brain in animals and in man. After a century of investigation we have still only a first idea of the degree and meaning of human cerebral localization. Roger Sperry and Michael Gazzaniga have interrogated the two halves of the brain separately; Wilder Penfield has elicited vivid specific memories on the operating table, but in the last figures of the book the theorists still designate big areas for "writing fluency" and "melody." Such investigative tools as brain injury, epilepsy, "phantom limbs" and electrodes make for slow work. It clearly beats labeling an area for Republicanism, but it is a long way from localizing function.

The two British scholars who have put together this book have given both the trained and the general reader a valuable gift; one expects to see these pictures reproduced in a wide variety of contexts. The references, both historical and substantial, are rich.

ELECTROSTATICS AND ITS APPLICATIONS, edited by A. D. Moore. Foreword by Vincent J. Schaefer. John Wiley & Sons, Inc. ($24.95). In the era of Franklin of Philadelphia, to whom this fine book is dedicated, electrical science was electrostatics. Of course, charges must move a little if anything interesting is to happen on a human scale, but much charge and generally not much motion—a high impedance—roughly defines the domain of electrostatics. Volta invented current, Faraday electromagnetism, and by 1900 "electromagnetics had virtually eclipsed electrostatics" in applications.

The reason seems fundamental, although the enthusiastic editor of this collection of 20 chapters, contributed by 18 experts of the widest learning and experience, is not yet prepared to concede the point. All matter has charges that are mobile, given enough force. An epidemic breakdown occurs, so that charges move too freely, once a certain limiting electric field is present. That field determines the maximum net charge on any macroscopic surface area. The energy density in that maximum stable field is less, in free air, by four orders of magnitude than the energy density for a practical magnetic-field strength. Other materials, even the vacuum, help some, and cryogenic temperatures may one day become important, but so far no big electrostatic machines are around at high power. The moving-belt generators of Van de Graaff are commercial and scientific successes for high voltages (now up to 20 million volts and more) but the power is modest: a few tens of kilowatts. Noel Felici of Grenoble has designed electrostatic generators, now realized commercially with plastic cylinders spinning fast in high-pressure hydrogen, that also produce only tens of kilowatts in machines of modest size. Motors too exist, their rotors charged by a corona spray. There is a neat one pictured here, all Lucite and foil and only about five inches long, delivering 75 watts of power. Bigger sizes are foreseen, but 10 kilovolts at low current is not attractive.

On the other hand, the spinning of shafts is not the sole aim of engineering. Observe that the charge density on a surface is limited by the breakdown field. But then the electrostatic force on a charged grain is proportional to its area, whereas weight varies with the volume. Place any sufficiently small macroparticle in an electric field, then, and its motion is controlled not by gravity but by the charge you have placed on it. That charge depends on its history, its chemistry, its environment. Here is pay dirt for the ingenious electrostatic engineer.

In the past decade or two the small particles have very often been carbon-impregnated resins carried on glass beads, all chemically charge-controlled. The coated beads cascade across a surface where an optical image has been focused on a layer of charged photoconductor. The dark regions remain charged but the lighted areas lose their charge. The carbon particles leave their carrier beads and adhere to the charge pattern. Properly spray-charged paper picks up the tiny dark particles from the surface, and another of those billions of copies rolls out of the machine.

Copying is not all. Dust is charged by a corona discharge and is thus separated from the exhaust gas that leaves the smokestack. The process goes back 60 years and more and is responsible for the collection of most power-station fly ash, 20 million tons of particulate air contaminants per year.

The computer can produce bits at an inordinate rate. Sometimes it is desired to record all that output. There are mechanical printers, printing one line at a time, that press inked type on paper at the rate of 2,000 lines per minute, say a page per second. Faster yet are ingenious new means of nonimpact printing, described here in some detail. One elegant scheme fires a beam of 100-micron charged droplets of ink at the paper. Variations in charging voltage and in a deflecting field en route control the droplet's landing point, and a dot matrix builds up a letter. The text comes out at the demoniacal rate of 30,000 lines per minute, under easy computer control. (The reading machine is not so well developed!) Most refrigerators and most golf balls are electrostatically sprayed with liquid paint droplets; riding the guiding fields, the charged droplets never miss. Compared with air spray, the process coats two or three times as much surface per gallon of paint. Sandpaper is coated by the mile with grit particles in the same way. Here the advantage is even greater: the asymmetric solid particles dive into the gluey paper head first, all oriented by the field lines with their sharp tails pointing out.

This book reflects the variety and the heterodoxy connected with electrostatics. Some of that flavor is personal, but some inheres in the mysteries that still accompany the phenomena, which are surface-sensitive, depend on small-scale properties and therefore vary with the weather, and are clearly cousin to the

INDEX OF ADVERTISERS

NOVEMBER 1973

lightning bolt. All together this book about a neglected technology, half-science and half-art, partly ancient and partly modern, is an invitation to the inventor, the sophisticate of process design and the general reader alike.

RIVERS IN THE CITY, by Roy Mann. Praeger Publishers ($20). PANORAMA OF LONDON: 1749. From an original engraving made by the Buck Brothers, with contemporary notes by John Wellsman. Dover Publications, Inc. ($3.50). THE GRAND PANORAMA OF LONDON FROM THE THAMES, made for *The Pictorial Times* in 1844. Introduction by Asa Briggs. Dover Publications, Inc. ($3.50). In the year 1502 Florence was at war with Pisa downstream. The science adviser to the ruler proposed an audacious plan: to steal the river Arno from Pisa by cutting a canal link to the sea, at once eliminating Pisa as a port and giving Florence that status. The career of Duke Cesare Borgia could not accommodate so lengthy a project, however; Leonardo da Vinci left his service in a couple of years, and his plan remained on paper, a lasting contribution to the attitudes and techniques of our technology. The Arno is a short river, but it drains the steep Apennines. Since 1269 written records have been kept of its flooding of Florence; no flood has been so severe as that of November, 1966. Florence lives well with the river; the open landscape corridor of the Arno has been framed by the Medicis, whose two palaces on both banks (their residence in the Pitti within its cool, high gardens, and the bureaucratic Uffizi) are even linked by an enclosed "safe and secret" passageway, crossing the river just under the tiled roof of the Ponte Vecchio. But the Florentines need more than that wonderful, intimate, romantic framing of the river with elegant bridges and palaces, or the fine long cobbled quays. Their only green open space lies on the old Medici dairy farm just downstream. "Visitors romanticize the city's density; Florentines enjoy escaping it." The modern plan is to keep the Arno banks secure for green parkland and public recreation both upstream and down from the compact city.

The attractive and unusual *Rivers in the City*, basically a volume of annotated maps and photographs, is the fruit of five years' study by a landscape architect and planner of Cambridge, Mass. His book "explores fifteen rivers and riverine systems within their own environmental, urban/regional, and historical contexts." He gives no statistics on thermal or bacterial pollution; his interest is not in water itself but in the land, the structures and the people on the banks. He looks at riverside problems and answers in western Europe and eastern America, each case seen in its own setting, carefully documented for the eye, analyzed with insight and concern in a calm rather than polemic tone.

A river is a unique urban corridor, a unified setting for the view of the city and a place of repair for its people. To cut it away from those people with highway and warehouse, to break its long sight lines with huge, out-of-scale power plants is to incur a heavy cost that is not yet entered on the balance sheets. Pipes can carry cooling water more easily than they can people; container ports need not sprawl into great, dead semitrailer parking lots when other designs exist. The trick is to govern well and to invest even where profit is not easily seen. To build their freeways along the landings of the Seine as that busy working port declined in the 1950's was easy for the city fathers; the land was already theirs and no complaints would come from dispossessed owners. But the loss of Paris itself is latent in such decisions. It is not too late; the long-range plan for the Paris basin is already much wiser. Dutch planners are far ahead in their claims on city rivers for leisure: mass transit, deep-bore tunneling, underground construction, compactness of design, integral planning and regulation of the scale of structures anywhere near the river's visual corridor are techniques that hold promise. Where railroads and highways have already encroached, they can be decked over, at least in part. The bleakness even of Newcastle's Tyneside, a legacy of antique industry centered on coal and *laissez faire,* can be moderated. There the plan includes a mix of landscape and recreational park development with new industrial parks. The river can become again a living and working citizen of the city. "The root of environmental abuse is not ideologic. It is economic." A new use of our resources, a new distribution of work and of incomes, a new policy of taxation, in short a new deal all around, may be a way to fine new riverbanks, from Washington to Venice.

In his book Mann reproduces two engraved pages of London in 1616, viewed from the Thames at London Bridge. The two small *Panorama of London* books, originally published in London by Sidgwick & Jackson, present similar but much more extensive engravings, showing in each case the city from the river from Westminster to the Tower. The 1749 vista is about 10 feet long, accordion-folded between its covers; the Victorian view unfolds to more than 12 feet. It is clear from the difference between Mann's Jacobean London and this Georgian panorama that the city was in rapid growth. The skyline remains dominated by the church towers. By Victorian times smoke from coal fires on land and on shipboard gives panache to Wren's old city of spires. Either little book is hard for anyone to resist who loves London or cities in general.

CONTEMPORARY ARCHAEOLOGY: A GUIDE TO THEORY AND CONTRIBUTIONS, edited by Mark P. Leone. Southern Illinois University Press ($15). For about a decade American archaeology has been in revolution. The cause was internal, not external. Year by year the general public has delighted in evidence of the antiquity of man, of the universality of his past as hunter and gatherer, of the widespread brilliance of his ancient art and artifice, in signs of lengthy migrations and apparently independent discoveries, of the roots of abstract thought in early grammar and mathematics and astronomy, in objective dating and in the details of human life revealed by pollen and bone. The literate unspecialized audience did not complain, but something was very wrong within. It is strange—even for an outsider strongly impressed by unfamiliar sculpture and careful plans for a dozen bewildering settlement layers—to encounter at the end of the thick volume of some report two diffident pages of conclusions, mainly psychoanalytic. Why all that care?

Paul Martin of Arizona, no newcomer but a convert to new ways, puts it pungently: "I used to be a virtuoso of pottery types. Given almost any sherd from the southwestern United States, I could place it spatially and temporally. But I was unable to tell you a thing about the interrelationship of shapes, designs, types, and functions. I had not 'seen' that a given pottery type might have been used almost exclusively for ritual or burial purposes . . . that the patterning of human behaviour might be explained by the variability in the archaeological record."

This change too has a history. Economics as a whole, Marxian materialism in particular, informed such distinguished workers as Gordon Childe a generation ago. Walter Taylor of Southern Illinois, who has a paper in this anthology, was close to the present trend in 1948 in a wholly American context. But it is younger workers such as Lewis Binford and Kent Flannery and James Deetz who have made the sharpest comments.

Those arguments are here; the volume has 33 papers, eight of them first published now, on the issues. About half of them are substantive pieces, presented as examples of the new trend. The rest represent prelude or the critique of method. The questions are too rich to outline in a brief review, but it may be allowable to present an impression. A descriptive paper, taxonomic in nature, is at best a part of science. Archaeology can fulfill its promise only as it sets itself a sterner task: "good archaeology" will normally include a hypothesis to explain some observations, a hypothesis that can in the end be tested by independent data. "The site" is not given by God. How are digging sites selected? In 37 reports Binford found not one statement of the criteria for site selection. Sampling theory, studies not only of types but also of their variation, a search for and recording of details (Were streams nearby? In which direction did the bones lie?) plainly belong alongside of meaningful hypotheses. Not one author disagrees: "For the last ten years, the pace has changed. . . . No author regrets the change, and none condemns it."

Many hopes arise from these pages. General systems theory, location theory, game theory, multivariate typology, sensible random sampling, numerical studies of data clearly are large schemes open to men and women who no longer need to concentrate on knowing all the shards by space and time. Some of these techniques will bear fruit; the reader can wish the workers well.

Still, there are dangers. Samples remain stubbornly small when experiments cannot be repeated. Not every worker learns that lesson soon enough. Nor can broad judgments be the subject of easily tested hypotheses. One must indeed hope for a science more balanced between data and inference, a subject disciplined by ideas and processes, not by technique and collection alone. But one must also look beyond balance into the depth of the strategy; it was a wry poet who prayed God that he not be allowed to commit a social science. We have plenty of long questionnaires and slight conclusions, fully testable by chi square but piecemeal and worthless after the test. There is no royal—or even mathematical—road to science. Fine-grainedness is not enough, any more than is library speculation or authoritative pot-piling. Each generation can find its own errors!

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

MULTIPLE-WARHEAD MISSILES

RACE TO OBLIVION: A PARTICIPANT'S VIEW OF THE ARMS RACE. Herbert York. Simon and Schuster, 1970.

MISSILE SUBMARINES AND NATIONAL SECURITY. Herbert Scoville, Jr., in *Scientific American*, Vol. 226, No. 6, pages 15–27; June, 1972.

WORLD ARMAMENTS AND DISARMAMENT: SIPRI YEARBOOK 1973. Stockholm International Peace Research Institute. Almqvist & Wiksell, 1973.

COMMUNICATION BY OPTICAL FIBER

LASER COMMUNICATION SYSTEMS. W. K. Pratt. John Wiley & Sons, Inc., 1969.

LIGHT TRANSMISSION OPTICS. Dietrich Marcuse. Van Nostrand Reinhold Company, 1972.

GLASS FIBERS FOR OPTICAL COMMUNICATIONS. Robert D. Maurer in *Proceedings of the IEEE*, Vol. 61, No. 4, pages 452–462; April, 1973.

PROTON INTERACTIONS AT HIGH ENERGIES

THE ENERGY DEPENDENCE OF THE PROTON-PROTON TOTAL CROSS-SECTION FOR CENTRE-OF-MASS ENERGIES BETWEEN 23 AND 53 GEV. U. Amaldi, R. Biancastelli, C. Bosio, G. Matthiae, J. V. Allaby, W. Bartel, G. Cocconi, A. N. Diddens, R. W. Dobinson and A. M. Wetherell in *Physics Letters*, Vol. 44B, No. 1, pages 112–118; April 2, 1973.

MEASUREMENT OF THE TOTAL PROTON-PROTON CROSS-SECTION OF THE ISR. S. R. Amendolia, G. Bellettini, P. L. Braccini, C. Bradaschia, R. Castaldi, V. Cavasinni, C. Cerri, T. Del Prete, L. Foà, P. Giromini, P. Laurelli, A. Menzione, L. Ristori, G. Sanguinetti, M. Valdata, G. Finocchiaro, P. Grannis, D. Green, R. Mustard and R. Thun in *Physics Letters*, Vol. 44B, No. 1, pages 119–124; April 2, 1973.

THE COMPLEMENT SYSTEM

DEVELOPMENT OF THE ONE-HIT THEORY OF IMMUNE HEMOLYSIS. Manfred M. Mayer in *Immunochemical Approaches to Problems in Microbiology: A Symposium Held at the Institute of Microbiology, Rutgers University, September 6–8, 1960*, edited by M. Heidelberger and Otto G. Prescia. Rutgers University Press, 1961.

THE MOLECULAR BASIS OF THE BIOLOGICAL ACTIVITIES OF COMPLEMENT. Hans J. Müller-Eberhard in *The Harvey Lectures*, Series 66. Academic Press, Inc., 1970/71.

MECHANISM OF CYTOLYSIS BY COMPLEMENT. Manfred M. Mayer in *Proceedings of the National Academy of Sciences of the U.S.A.*, Vol. 69, No. 10, pages 2954–2958; October, 1972.

THE RECOGNITION OF FACES

SOME ASPECTS OF RECOGNITION OF HUMAN FACES. L. D. Harmon in *Pattern Recognition in Biological and Technical Systems: Proceedings of the Fourth Congress of the Deutsche Gesellschaft für Kybernetik*, edited by Otto-Joachim Grüsser and Rainer Klinke. Springer-Verlag, 1971.

IDENTIFICATION OF HUMAN FACES. A. Jay Goldstein, Leon D. Harmon and Ann B. Lesk in *Proceedings of the IEEE*, Vol. 59, No. 5, pages 748–760; May, 1971.

MAN-MACHINE INTERACTION IN HUMAN-FACE IDENTIFICATION. A. J. Goldstein, L. D. Harmon and A. B. Lesk in *The Bell System Technical Journal*, Vol. 51, No. 2, pages 399–427; February, 1972.

MASKING IN VISUAL RECOGNITION: EFFECTS OF TWO-DIMENSIONAL FILTERED NOISE. Leon D. Harmon and Bela Julesz in *Science*, Vol. 180, No. 4091, pages 1194–1197; June 15, 1973.

HILBERT'S 10TH PROBLEM

COMPUTABILITY & UNSOLVABILITY. Martin Davis. McGraw-Hill Book Company, 1958.

GAMES, LOGIC AND COMPUTERS. Hao Wang in *Scientific American*, Vol. 213, No. 5, pages 98–106; November, 1965.

HILBERT. Constance Reid. Springer-Verlag, 1970.

HILBERT'S TENTH PROBLEM IS UNSOLVABLE. Martin Davis in *American Mathematical Monthly*, Vol. 80, No. 3, pages 233–269; March, 1973.

THE FLYING LEAP OF THE FLEA

POWER IN FLAPPING FLIGHT. T. Weis-Fogh in *The Cell and the Organism: Essays Presented to Sir James Gray*, edited by J. A. Ramsay and V. B. Wigglesworth. Cambridge University Press, 1961.

THE JUMP OF THE FLEA: A STUDY OF THE ENERGETICS AND A MODEL OF THE MECHANISM. H. C. Bennet-Clark and E. C. A. Lucey in *Journal of Experimental Biology*, Vol. 47, No. 1, pages 59–76; August, 1967.

ANIMAL LOCOMOTION. Sir James Gray. W. W. Norton & Company, Inc., 1968.

ANIMAL MECHANICS. R. McNeill Alexander. University of Washington Press, 1969.

THE EVOLUTION OF THE PACIFIC

200,000,000 YEARS BENEATH THE SEA: THE STORY OF THE GLOMAR CHALLENGER. Peter Briggs. Holt, Rinehart and Winston, Inc., 1968.

DEEP OCEAN DRILLING WITH GLOMAR CHALLENGER. M. N. A. Peterson and N. T. Edgar in *Oceans*, Vol. 1, No. 5, pages 17–32; June, 1969.

DIACHRONOUS DEPOSITS: A KINEMATIC INTERPRETATION OF POST JURASSIC SEDIMENTARY SEQUENCE ON THE PACIFIC PLATE. B. C. Heezen, I. D. MacGregor, H. P. Foreman, G. Forristal, H. Hekel, R. Hesse, R. H. Hoskins, E. J. W. Jones, A. Kaneps, V. A. Krasheninnikov, H. Okada and M. H. Ruef in *Nature*, Vol. 241, No. 5384, pages 25–32; January 5, 1973.

MATHEMATICAL GAMES

TAXICAB GEOMETRY—A NON-EUCLIDEAN GEOMETRY OF LATTICE POINTS. Donald R. Byrkit in *The Mathematics Teacher*, Vol. 64, No. 5, pages 418–422; May, 1971.

SPIROLATERALS. Frank C. Odds in *The Mathematics Teacher*, Vol. 66, No. 2, pages 121–124; February, 1973.

PATERSON'S WORM: ARTIFICIAL INTELLIGENCE MEMO NO. 290. Michael Beeler. Massachusetts Institute of Technology, June, 1973.

THE AMATEUR SCIENTIST

ADVANCES IN GEOPHYSICS: VOL. II. Hugo Benioff in *Earthquake Seismographs and Associated Instruments*. Academic Press, Inc., 1955.

ELEMENTARY SEISMOLOGY. Charles F. Richter. W. H. Freeman and Company, 1958.

SCIENTIFIC AMERICAN BOOK OF PROJECTS FOR THE AMATEUR SCIENTIST. C. L. Stong. Simon and Schuster, 1960.