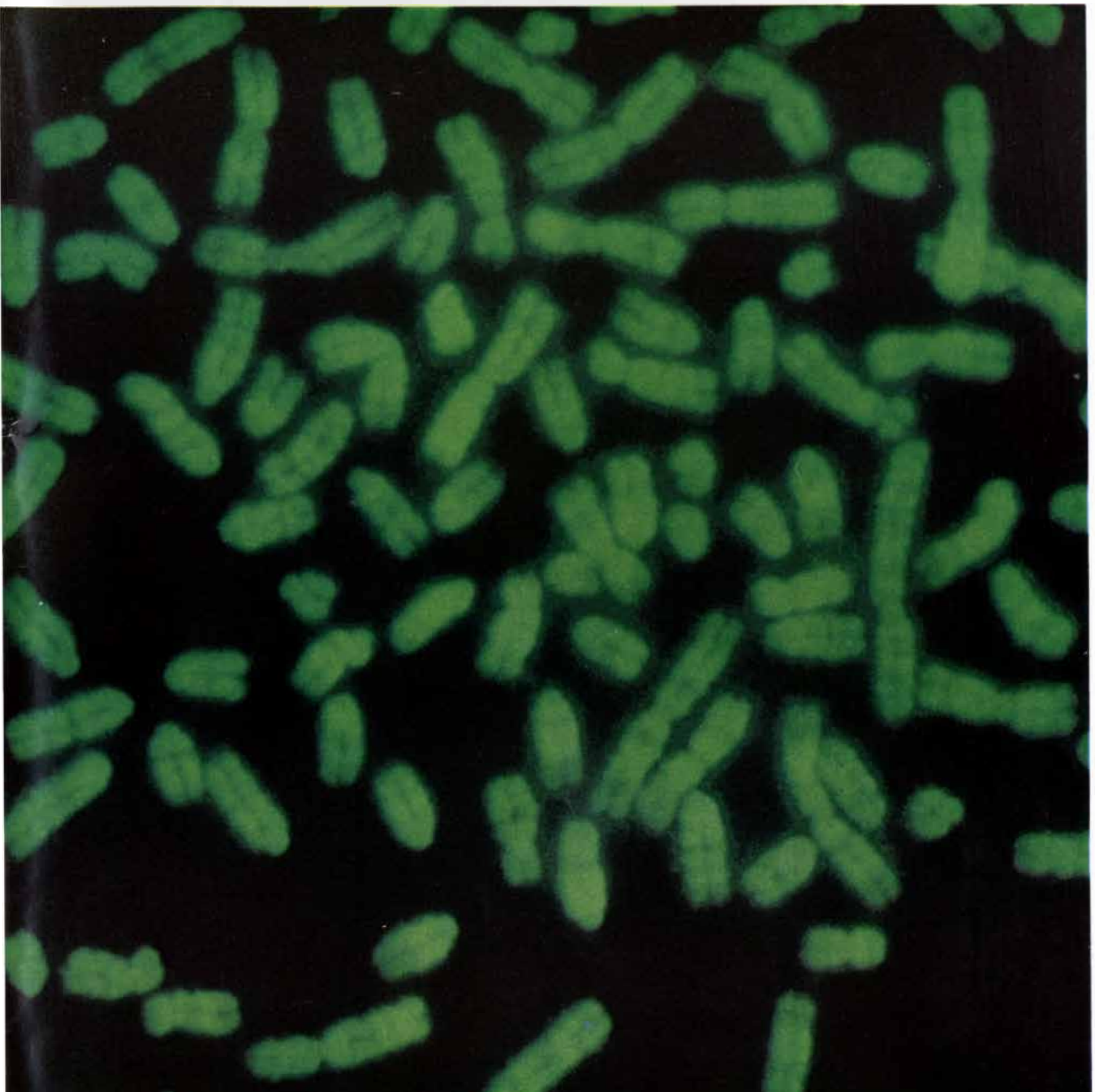
SCIENTIFIC AMERICAN



FLUORESCENT CHROMOSOMES

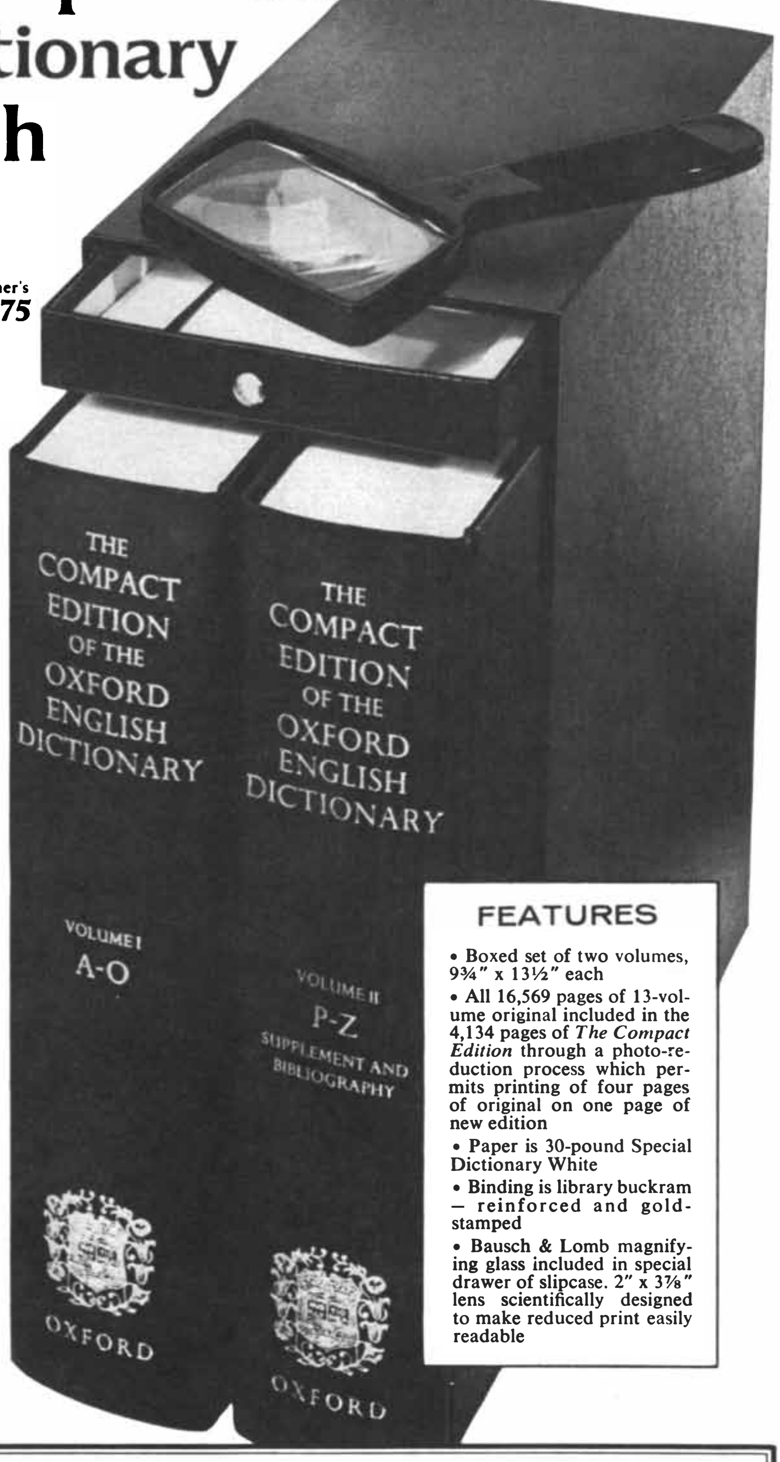
ONE DOLLAR

July 1974


PRODUCE OF FRANCE
BOTTLED IN FRANCE
Triple Orange
Triple Orange
Grand Marnier
Liqueur
PARIS — FRANCE
STRENGTH 80° PROOF — NET CONTENTS 23/32 QUART
It even makes the late evening news easier to digest.
Grand Marnier is more than a liqueur. It's a form of civilization in a relatively uncivilized world. To find out what more it can do, and what more you can do with it, write for our free gourmet booklet.
Carillon Importers, Ltd., 745 Fifth Avenue, New York, N.Y. 10022. Product of France. Made with fine cognac brandy. 80 proof.

As your introduction to membership in the BOOK-OF-THE-MONTH CLUB®

The most complete and most scholarly dictionary of the English language for only $15 Publisher's price $75



THE SUGGESTED TRIAL: You simply agree to buy four Club choices within a year at substantial savings on most books you choose

THE OXFORD ENGLISH DICTIONARY is generally regarded as the final arbiter of the meaning, origin, history and usage of words in the English language. Until now, it has been available only as a thirteen-volume set, priced at $350. Now, through the combination of an ingenious method of micrographic reproduction and a fine Bausch & Lomb optical lens, every single one of its 16,569 pages, fifty million words and close to two million illustrative quotations appears, in easily readable form, in the two volumes of *The Compact Edition.*

The *New York Times* book critic Christopher Lehmann-Haupt has said of this edition: "It is something of a miracle. . . . *The Compact Edition* is easier to work with than the original with its 13 separate volumes. . . . Even at $75, the set is an extraordinary bargain."

Even more extraordinary, as a trial member of the Book-of-the-Month Club you may obtain *The Compact Edition of the Oxford English Dictionary* for only $15. And as long as you remain a member, you will receive the Book-of-the-Month Club *News,* a literary magazine announcing the coming Selection and describing other important books, most of which are available at substantial discounts — up to **40%** on more expensive volumes. All of these books **are identical to the publishers' editions in content, format, size and quality.**

If you continue after your trial membership, you will earn at least one Book-Dividend® Credit for every Selection or Alternate you buy. These Credits entitle you to obtain a wide variety of books, called Book-Dividends, at astonishing savings — *at least 70% of publishers' list prices.* Under this unique system, more than 100 Book-Dividends will be available to choose from every year, enabling you to acquire a well-rounded library at a small fraction of what you would otherwise have to pay.

CONVENIENCES OF MEMBERSHIP

- You will receive the Book-of-the-Month Club *News,* a literary magazine published by the Club fifteen times a year. The *News* describes the coming Selection and scores of Alternates, and will be sent to you approximately every three and a half weeks.
- If you wish to purchase the Selection, do nothing and it will be shipped to you automatically.
- If you do not want the Selection — or you would like 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.
- If, because of late mail delivery of the *News,* you should receive a Selection without having had 10 days to decide whether you want it, that Selection may be returned at Club expense.

FEATURES

- Boxed set of two volumes, 9¾" x 13½" each
- All 16,569 pages of 13-volume original included in the 4,134 pages of *The Compact Edition* through a photo-reduction process which permits printing of four pages of original on one page of new edition
- Paper is 30-pound Special Dictionary White
- Binding is library buckram — reinforced and gold-stamped
- Bausch & Lomb magnifying glass included in special drawer of slipcase. 2" x 3⅞" lens scientifically designed to make reduced print easily readable

BOOK-OF-THE-MONTH CLUB, INC.
280 Park Ave., New York, N.Y. 10017

4-A170-7

Please enroll me as a member of the Book-of-the-Month Club and send me *The Compact Edition of the Oxford English Dictionary,* billing me $15 (in Canada $17.50 — publisher's price $82.50). I agree to buy at least four Selections or Alternates during the first year I am a member, paying in most cases special members' prices. My membership is cancelable any time after I buy these four books. A shipping charge is added to all shipments.

MR. / MRS. / MISS ....
(Please print plainly)

Address..54

City..

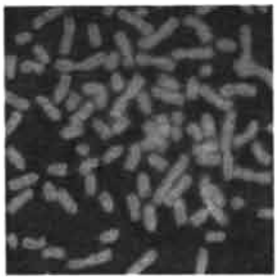
State...................................*Zip*...................

Established 1845 **SCIENTIFIC AMERICAN** July 1974 Volume 231 Number 1

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. 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: $12 PER YEAR, U.S., POSSESSIONS AND CANADA; $15 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O PUBLICATIONS DISTRIBUTION SERVICE, ST. KATELIJNEVEST 14, B2000 ANTWERP, BELGIË.



THE COVER

The photomicrograph on the cover shows the chromosomes of a hybrid cell, the result of the fusion of a mouse tumor cell and a human fibroblast, or connective-tissue cell. Examination of the chromosomes and the enzymes and other products of such cells is proving to be a valuable tool with which to map human genes (see "Hybrid Cells and Human Genes," page 36). The chromosomes have been stained with quinacrine, a dye that fluoresces under ultraviolet radiation, revealing bands that form a distinctive pattern characteristic of each different chromosome. The banding pattern makes it possible to distinguish human chromosomes from mouse chromosomes and each of the 24 different human chromosomes from one another.

THE ILLUSTRATIONS

Cover photograph by Raju S. Kucherlapati, Yale University

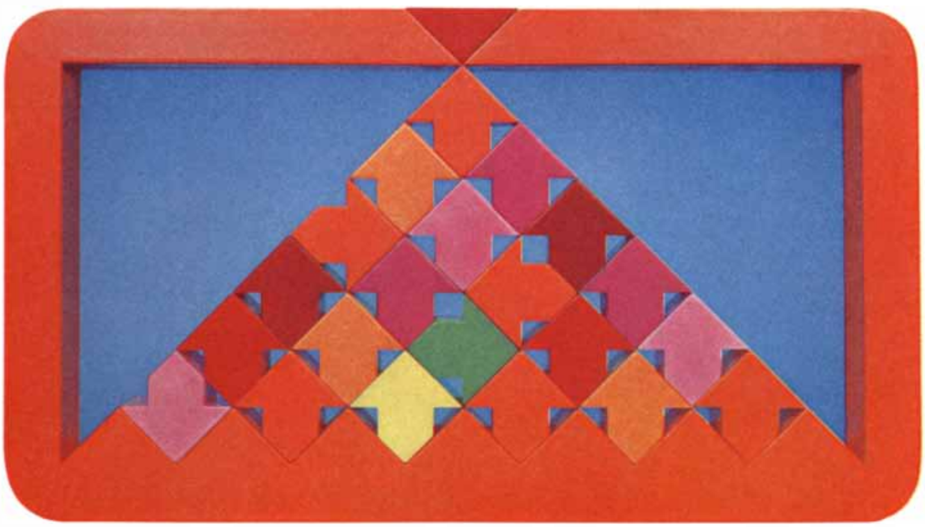
Page	Source
22	Ilil Arbel
23	Armand Dione
24–26	Ilil Arbel
27	Howard W. Emmons, Harvard University
29	K.-E. Kaissling, Max Planck Institute for Behavioral Physiology
30	Dietrich Schneider, Max Planck Institute for Behavioral Physiology (*top*); Lorelle M. Raboni (*bottom*)
31–32	Lorelle M. Raboni
33	R. A. Steinbrecht, Max Planck Institute for Behavioral Physiology
34	E. Priesner, Max Planck Institute for Behavioral Physiology
35	Lorelle M. Raboni
37	Frank H. Ruddle and Raju S. Kucherlapati, Yale University
38–44	Bunji Tagawa
50	Argonne National Laboratory, (*top*), Ilil Arbel (*bottom*)
52–59	Gabor Kiss
61	Fritz Goro (*top*); John C. Brandt, National Aeronautics and Space Administration (*bottom*)
62–64	Dan Todd
65	Howard C. Bryant, University of New Mexico; Nelson Jarmie, Los Alamos Scientific Laboratory
66	Howard C. Bryant, University of New Mexico; Nelson Jarmie, Los Alamos Scientific Laboratory (*top left*); Dan Todd (*top right and bottom*)
70	Dan Todd
74–85	Irving Geis
86	Graphic Presentation Services, Inc.
90	Edgar Rubin
91–100	Maurits C. Escher
101	Maurits C. Escher (*top*), Giovanni Battista Piranesi (*bottom*)
102	Ilil Arbel (*top*), Maurits C. Escher (*bottom*)
103	Ilil Arbel (*top*), Maurits C. Escher (*middle*), Josef Albers (*bottom*)
104	D. M. MacKay (*top*), Maurits C. Escher (*bottom*)
106–109	Tom Prentiss
110–112	Graphic Presentation Services, Inc.
113–114	Tom Prentiss
116–120	Ilil Arbel
123	Ben Rose
124–126	Jerome Kuhl

photographer: marty evans

The real

Our election laws and traditions reflect the country as it was decades ago. The time has come to ask some basic questions.

How long should elected officials serve?
Is there still a need for the electoral college?
Should taxpayers foot the bill for election campaigns?
Does our present system truly provide one vote for each person?
Is our political structure so big that officials are too far from those who elect them?



artist: per vollquartz "government: triangle and people"

The ideal

A nation in which those elected truly represent the will of those who elected them– for the good of all.

AtlanticRichfieldCompany


GLAD
Prestone

LETTERS

Sirs:

Since the purpose of Professor Condit's article ["The Wind Bracing of Buildings," by Carl W. Condit; SCIENTIFIC AMERICAN, February] was to trace the structural development of building forms tall enough for wind loading to become a principal design consideration, it is perhaps understandable that he omitted reference to the tremendous development of thin-shell structures.

Although most of this development has occurred in response to the need for structural forms capable of providing large clear-span roof enclosures with a minimum of weight of materials, within the past decade the technique has been applied to tall structures for which wind loading resistance was a major consideration. The magnificent hyperbolic paraboloid roof of the new cathedral in San Francisco, designed by Pier Luigi Nervi, rises over 200 feet. Tall shells of hyperboloid-of-revolution form have recently been built as cooling towers for nuclear power plants. In both of these cases the shells provide both the lateral and vertical support of the structures. So far as I am aware, the controversial National Gettysburg Battlefield Tower illustrates the first use of a shell structure in a lattice framing form to provide pure wind bracing. (The vertical or gravitational loads are all carried by a core structure.)

A second point about Professor Condit's article is worth mentioning. He suggests that the dead load of the structure must be sufficient to prevent the structure from overturning. On the assumption that most readers would not consider the earth or rock mass below a structure as being a part of the structure it should be noted that as our structures get both taller and lighter, their superstructures alone are not enough to provide the required stability in the face of the wind loads involved. To get the required stability they depend on massive foundation blocks or a variety of earth and rock anchorage systems to supply the needed "hold-down" forces not supplied by the superstructure dead load itself. The Gettysburg Tower is an example of a structure so light for its height that it requires anchorage for stability. It is anchored to a massive diabase formation below its ringwall base, and so is being held down by a substantial portion of Adams County, Pa.

In addition to their remarkable strength for their light weight, shell structures are also remarkably stiff for their light weight. We can expect more shell principles to be applied to tall structures to help reduce their lateral deflections due to wind loading, particularly as we develop more refined abilities to make them lighter.

JOEL H. ROSENBLATT, P.E.

Silver Spring, Md.

Scientific American, July, 1974; Vol. 231, No. 1. 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, $12 ☐ 2 years, $22 ☐ 3 years, $30
All other countries:
☐ 1 year, $15 ☐ 2 years, $27 ☐ 3 years, $37

NAME

NEW ADDRESS

OLD ADDRESS

Sirs:

I agree entirely with Mr. Rosenblatt's assertions about other forms of wind resistance and the need to tie down certain kinds of light or very slender structures. One might generalize from the material in Mr. Rosenblatt's letter and point out that any closed surface, such as a cylinder, has a built-in resistance to deformation by horizontal forces regardless of the direction in which they are acting, and one-sheet hyperboloids have an additional resistance by virtue of their two-way curvature. A conspicuous early example of tying down a high, slender skyscraper is the Lincoln Tower in Chicago (1927–1928).

CARL W. CONDIT

Northwestern University
Evanston, Ill.

Sirs:

I found "The Delivery of Medical Care in China," by Victor W. Sidel and Ruth Sidel [SCIENTIFIC AMERICAN, April], to be both fascinating and informative. I feel that they may have misled us, however, in their comment on breast feeding and birth control ("... Nursing [is] a period when, it is mistakenly believed by some, a woman cannot conceive."). Although breast feeding is not as effective as many other forms of birth control, and it is possible for a woman to conceive while nursing, it has been shown in a number of studies that conception rates are lower for nursing women than for non-nursing ones. Delays in the return of the menstrual and ovulatory cycles have been shown to be correlated with the degree of breast feeding. Mechanisms for postpartum lactation amenorrhea have not been explained, and work in this area may lead to new forms of birth control.

MARK J. POMERANTZ

University of California
San Diego

Sirs:

We reiterate our statement that the folk belief that a woman cannot conceive while nursing is false. A nursing woman is indeed statistically less likely to conceive than is a non-nursing woman, as Dr. Pomerantz notes, but since other methods of contraception are more effective, it would be unfortunate both for the society and for the woman herself for her to rely on nursing as her sole form of contraception when one or more of the other methods are available to her. This is particularly important because ovulation and conception can occur during the time between delivery and the first post-delivery menstrual flow, a time in which she may mistakenly believe herself "safe."

VICTOR W. SIDEL, M.D.

RUTH SIDEL, M.S.W.

Montefiore Hospital and Medical Center
Bronx, N.Y.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1924: "The now well-known Army intelligence tests, originally applied in order to differentiate men for different Army posts, disclosed the unwelcome fact that large numbers of the population ranked very low in innate intelligence. Professor P. E. Davidson, however, argues that if such conclusions are to be considered sound, three assumptions must be granted: that the Army draft was truly representative of the American population, that the tests were really tests of native ability and not of educational advantages, and that the native intellect in question is so general as to condition social success of any significant kind. There are reasons for disputing each of these assumptions, he maintains, and gives a salutary and timely check to the enthusiasts who would impute to tests more than they can legitimately bear."

"Some 30,000 craters are visible on the face of the moon that is presented to us, and there is every reason to believe that there are just as many on the other face, on which the human eye has never looked. It is now assumed that these craters are either volcanic or are due to the impact of extra-lunar masses of matter. The evidence, once fully appreciated, is all in favor of the impact theory. An interesting feature common to many of the craters is a central projection rising out of the floor, which has sometimes been said to indicate a volcanic origin. To appreciate the true cause of these nipple-like hills, which are in or very near the exact center of the craters, one need only study a large raindrop falling into a still pool of water. There is first a surging outward of the impacted water, then a surging backward toward the center, until a conical hill is formed in the center of the disturbance."

"Quartz, or silicon dioxide, is the commonest constituent of the earth's crust. We have always known its properties, but we have not been able to manufacture it save at inordinate expense. Today a way of treating it has been found, however, so that it can be made by fusion in an electric furnace. The outstanding merits of 'clear fused quartz,' as the new material will be called, are its vanishing coefficient of expansion and its transparency. A quartz rod one foot long, heated through 1,800 degrees Fahrenheit, will lengthen only seven thousandths of an inch and will transmit the entire spectral range, from the longest infra-red to the shortest ultraviolet. The place where these two properties may be applied to the best advantage is in the astronomical observatory, for the large mirrors, lenses and prisms."

SCIENTIFIC AMERICAN

JULY, 1874: "Work on the great suspension bridge between Brooklyn and New York, which had been temporarily suspended, is now resumed. The Brooklyn tower has reached an elevation of 222 feet above high-water mark, leaving 40 feet of masonry yet to be laid. The workmen are engaged on the arches, several courses of which are in position. The keystones will weigh 10 tons each and will constitute the heaviest blocks in the structure, the ordinary stones weighing some three tons."

"A new invention in telegraphy by George B. Prescott and Thomas A. Edison has lately been successfully tested at the main office of the Western Union Company in New York. The new invention is a process of multiple transmission by which two messages can be sent simultaneously in the same direction over the same wire, and either message can be dropped at any way station on the circuit. The old duplex system can be applied to the new invention, and by the combination four messages can be sent simultaneously over the same wire between any two terminal points."

"An Austrian inventor has recently constructed a petroleum engine, the principle of which is analogous to that of the single acting steam engine, with the difference, however, that the expansive force of steam in the latter is replaced by the explosion of the finely divided oil. The *Revue Industrielle* says that the invention has been applied to sewing machines with considerable success."

"Some months ago Darwin wrote to his disciple Fritz Müller, now in Brazil, directing his attention to the habits of the leaf-cutting ants, recently described by Mr. Belt. The reply contains a confirmation of Mr. Belt's observations to the effect that these ants do not feed on the leaves they gather in such vast quantities but on the fungus which grows on the leaves in their underground chambers. On examining the stomachs of these ants Mr. Müller found no trace of vegetable tissue that might have been derived from the leaves, but only a colorless substance showing under the microscope some minute globules, 'probably the spores of the fungus.' "

"Some time ago the International Metric Commission in Paris fixed upon an alloy of iridium and platinum for the standards to be used as the basis of the metrical system of weights and measures. Quite recently the ingot from which the standards are to be made was cast. The platinum and iridium were melted in quantities of 22 pounds. The ingots thus formed were cooled, cut in pieces and again melted, 176 pounds at a time. These masses were once more cut up and finally run into a single block, which will be cut and formed to the proper standards by mathematical measurements."

"Subcutaneous or hypodermic injection of medicines by very fine pointed syringes form a brand new feature of medical practice. Until recently these injections of liquids into the cellular tissue have been confined to narcotics or anesthetics, but at this time trials of many new articles are being held in the hope that the medicines will accomplish their object better than by the old style of dosing or inhaling."

"Some interesting experiments have been conducted with the electric light of Mr. Ladygin and Mr. Kosloff. The lamps each had two carbon rods, either of which could be placed in connection with the current of electricity. The first experiment consisted in burning a carbon rod in contact with the atmosphere, the rod being consumed in a few minutes. The current was then turned on the other rod and a brilliant and steady light was produced, which improved as the current was increased in intensity. The first rod was burned so that it might consume the oxygen in the lamp."

"A recent amendment to the building laws of New York city provides that every dwelling unit occupied by more than one family above the first floor, including all hotels and boarding houses, shall be provided with fire escapes, doors and alarms. Stores, warehouses and other buildings shall be provided, above the first story, with fireproof shutters."

I am you.

My name is Masakazu Fukushima, and here I am trumping some neighbors at contract bridge. I play every chance I get. It beats watching television.

Not that I've got anything against watching television, mind you. Watching is fine. What really bothers me is the bulk of the sets themselves. You get a big-screen color TV in your living room and it's like you parked a small car in there. An ugly small car.

The fact is I work for a company that makes a lot of TVs . . . Hitachi. I do electronics research, so I thought I'd try and do something to make color TVs easier to live with.

All people really want from a color TV is a good picture. So we figured the set should be *shaped* like a picture. Flat. To be hung on the wall.

We figured it out, then went into the lab and put it together. We came up with what we wanted . . . the world's first working prototype for a flat-profile color TV. It came out even better than we thought it would.

Not only is this new TV flat, the way it's designed is such that the screen can be just about however big you want it. There could be whole TV walls, say, in schools.

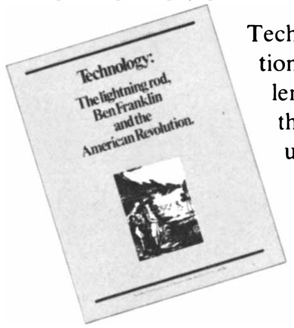
It opens up a lot of new, better possibilities for television, and that makes me happy. By doing myself a favor—by helping to invent something I'd enjoy—I've made something that can benefit an enormous number of people.

I am you.

HITACHI
Resourceful people

Hitachi America, Ltd.
Head Office: 437 Madison Avenue, New York, N.Y. 10022
Tel: (212) 758-5420
Branch Offices: San Francisco, Chicago, Atlanta.

● Power Generation & Transmission Equipment ● Industrial Machinery ● Transportation Equipment ● Communications Equipment ● Electronics & Home Appliances


Technology:
The lightning rod,
Ben Franklin
and the
American Revolution.

rod, Ben Franklin and the

ton gin and the mechanical reaper were American inventions that started the great movement toward mechanized agriculture. This allowed a given piece of land to produce more and more food. Thus, many Americans were freed from farming to engage in industrial occupations which led to additional technological advances.

When the Civil War began, the prediction of Frederick the Great met its ultimate test. And it was technology that saved the Union. The Civil War was the first technological war of the Industrial Age. The telegraph, for example, made it possible for the North to control its armies from afar. A more advanced railroad network in the North made Union armies more mobile and easier to supply. And it was the iron-clad warship, *Monitor,* steaming South just in time to neutralize the *Merrimac,* which prevented the South from breaking the Union blockade and, perhaps, breaking the Union.

The South, fighting with great bravery under perhaps the best generals America has ever produced, could not have been beaten but for the North's superior industry and technology.

Technology helped turn the United States from a wilderness into a great nation. Technology will continue being our best hope for the future. Overcoming lightning, the vastness of the early wilderness, and the drudgery of manual farming must surely have seemed insurmountable problems to early Americans. Today we have new problems that threaten to impede America's growth. Energy shortages, pollution and inflation for example. But considering what technology has accomplished before, and the far greater resources available today, we should feel confident that solutions are not out of reach.

Science and technology can solve many problems. If they don't, what else will?

GOULD
the proud inventors

THE AUTHORS

HOWARD W. EMMONS ("Fire and Fire Protection") is Gordon McKay Professor of Mechanical Engineering and the Abbott and James Lawrence Professor of Engineering at Harvard University. He was graduated from the Stevens Institute of Technology in 1933, obtaining his master's degree there in 1935 and his Sc.D. at Harvard in 1938. He joined the Harvard faculty in 1940 after several years as an engineer with Westinghouse Electric Corporation. In addition to his work on combustion he has been interested in supersonic aerodynamics and research on jet engines. "Diversion," he writes, "has been supplied over the years by tennis, the maintenance of an old New England home and the small-town problems of the school committee and selectman."

DIETRICH SCHNEIDER ("The Sex-Attractant Receptor of Moths") is a director of the Max Planck Institute for Behavioral Physiology near Munich and honorary professor of zoology at the University of Munich. He studied biological sciences at the University of Berlin and the University of Göttingen, receiving his doctorate in 1949. Before taking his present post in 1965 he had held appointments at the Max Planck Institute for Biology in Tübingen and at the Max Planck Institute for Psychiatry in Munich. He writes that he has recently become "deeply engaged in the development of the International Center of Insect Physiology and Ecology in Nairobi," serving as a member of the governing board and as a research director.

FRANK H. RUDDLE and RAJU S. KUCHERLAPATI ("Hybrid Cells and Human Genes") are at Yale University; Ruddle is professor of biology and human genetics and Kucherlapati is a Damon Runyon cancer research fellow. Ruddle obtained his bachelor's and master's degrees at Wayne State University in 1953 and 1956 respectively and his Ph.D. at the University of California at Berkeley in 1960. He joined the Yale faculty in 1961 as assistant professor of zoology. Kucherlapati received his M.Sc. from Andhra University in India in 1962 and his Ph.D. from the University of Illinois in 1972.

STEVEN WEINBERG ("Unified Theories of Elementary-Particle Interaction") is Higgins Professor of Physics at Harvard University and Senior Scientist at the Smithsonian Astrophysical Observatory. Before joining the Harvard faculty recently he had taught at the Massachusetts Institute of Technology, the University of California at Berkeley and Columbia University. He is the author of *Gravitation and Cosmology,* a book on general relativity. He won the Oppenheimer Prize in 1973 and was named Richtmeyer Memorial Lecturer for 1974; his article is based on his Richtmeyer lecture, which he presented in February at a joint meeting of the American Physical Society and the American Association of Physics Teachers. His extracurricular interests include medieval history and arms control; he has served as a consultant to the U.S. Arms Control and Disarmament Agency and is a member of the Council on Foreign Relations. He also sits on the Council of the American Physical Society and is a member of the National Academy of Sciences. He wishes to thank Sidney R. Coleman, George Wald and Victor F. Weisskopf for their counsel on the present article.

HOWARD C. BRYANT and NELSON JARMIE ("The Glory") are respectively professor of physics at the University of New Mexico and member of the staff of the Los Alamos Scientific Laboratory. Bryant writes: "I was graduated from the University of California at Berkeley in 1955 and spent the following summer as a 'summer student' at Los Alamos. I liked New Mexico so well that when I finished my Ph.D. work at the University of Michigan in 1960, I returned to the University of New Mexico to teach. I have been here ever since except for a two-year stay at the Stanford Linear Accelerator Center from 1967 to 1969. I am semiserious about painting watercolors and recommend it as a way of finding an occasional patch of serenity in life. I like camping, hiking and swimming, and I ride a bicycle a great deal." Jarmie was graduated from the California Institute of Technology in 1948 and obtained his Ph.D. at the University of California at Berkeley in 1953. "Apart from my work in nuclear physics," he writes, "I've had an abiding interest in various disciplines, both practical and esoteric, concerning the phenomena of consciousness and awareness. I'm also a frustrated teacher and can usually be talked into giving classes on any subject that I have sufficient skill in, from tennis to T'ai Chi Ch'uan."

ROBERT M. STROUD ("A Family of Protein-cutting Proteins") is assistant professor of chemistry at the California Institute of Technology. He obtained his bachelor's degree at the University of Cambridge in 1964 and his Ph.D. at the University of London in 1968. From 1965 to 1968 he was the molecular-biology reporter for *Medical News* in Britain. Apart from his studies of enzyme mechanisms he is currently working on the structure of neuroreceptor membranes and snake-venom toxins that specifically block those receptors. At Cambridge he was president of the swimming and water polo clubs, winning four Cambridge blues. Now he spends a considerable amount of time in Mediterranean and Oriental cooking and plays a number of fretted string instruments. He enjoys gliding and makes radio-controlled model gliders.

MARIANNE L. TEUBER ("Sources of Ambiguity in the Prints of Maurits C. Escher") is an art historian. "My university studies in art history were divided between Europe and America," she writes. "After four years at the University of Basel (where I also won the Swiss universities' championship in high jumping) I came to the U.S. on a foreign-student scholarship awarded by Vassar College. I received my master's degree from Vassar and then continued graduate work at Harvard University. Several years of community work followed while our two sons were young. When they went off to college, I returned to the history of art, but now with a new interest in modern art and its historical roots. One aspect fascinated me in particular, the artists whose work seemed closely tied to experiments on vision and visual perception. I have an unfair advantage in uncovering these links, because my husband is Hans-Lukas Teuber, head of the department of psychology at the Massachusetts Institute of Technology."

RALPH L. HOLLOWAY ("The Casts of Fossil Hominid Brains") is professor of anthropology at Columbia University. He writes that he "came to physical anthropology through a somewhat tortured route of metallurgical engineering, geology (B.S. from the University of New Mexico in 1959) and the recession of 1959." Holloway obtained his Ph.D. at the University of California at Berkeley in 1964 and has since been teaching at Columbia. "My wife and our two children and I are keen on traveling, tennis, swimming and beachcombing as joint hobbies," he says. "When I am not involved in these activities or academic duties, I can sometimes be heard playing the trumpet to maintain my sanity."

The Sinclair Scientific. Logs, trig and arithmetic. $99.95

(Now you can afford to get a little scientific.)

Display:
5-digit mantissa
2-digit exponent
(both signable)

Exponent:
200-decade range,
from 10^{-99} to 10^{+99}

actual size

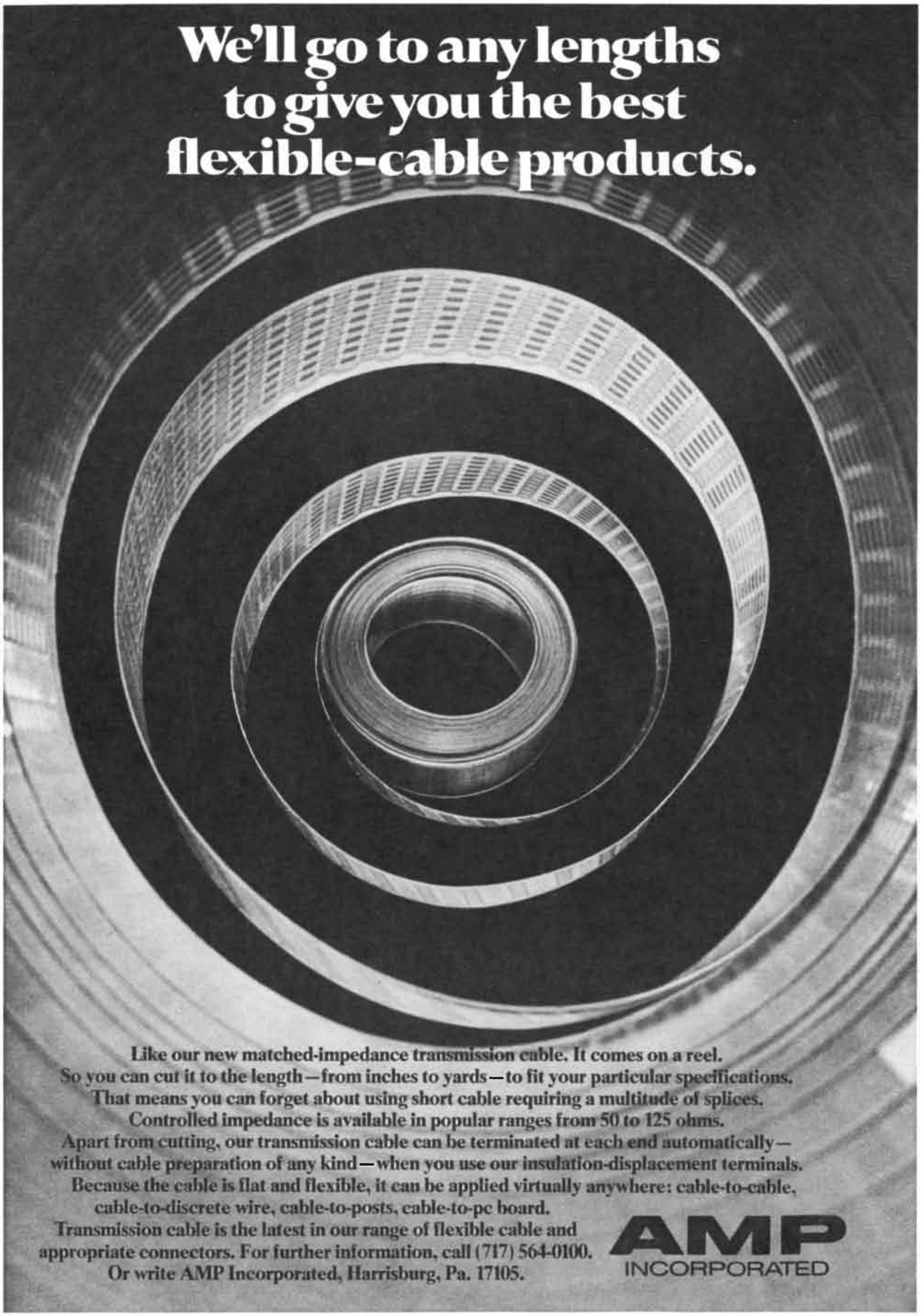
Functions:
4 arithmetic
2 logarithmic
6 trigonometric

Keyboard:
18 key format with
4 "triple-action"
function keys using
standard, upper and
lower case operation.

Logic:
Reverse Polish, with
post-fixed operators
for full flow chain
calculations.

Size:
4-3/8″ high; 2″ wide;
11/16″ thick.
Weight:
3-3/4 oz.
Warranty:
1 year.

Power Source:
Battery operated
with 4 inexpensive
AAA penlight
batteries, providing
over 25 hours of use.


We'll go to any lengths to give you the best flexible-cable products.
Like our new matched-impedance transmission cable. It comes on a reel. So you can cut it to the length—from inches to yards—to fit your particular specifications. That means you can forget about using short cable requiring a multitude of splices. Controlled impedance is available in popular ranges from 50 to 125 ohms.
Apart from cutting, our transmission cable can be terminated at each end automatically—without cable preparation of any kind—when you use our insulation-displacement terminals.
Because the cable is flat and flexible, it can be applied virtually anywhere: cable-to-cable, cable-to-discrete wire, cable-to-posts, cable-to-pc board.
Transmission cable is the latest in our range of flexible cable and appropriate connectors. For further information, call (717) 564-0100. Or write AMP Incorporated, Harrisburg, Pa. 17105.
AMP
INCORPORATED

Established 1845 SCIENTIFIC AMERICAN July 1974 Volume 231 Number 1

Fire and Fire Protection

Fire losses in the U.S. are the highest in the world. The reason is apparently the nation's high standard of living. Better protection calls basically for a better understanding of the mechanism of fire

by Howard W. Emmons

Taking into account only damage to property, the loss from fires in the U.S. last year was $3.1 billion, or more than $15 per capita. Although the U.S. is generally regarded as being a technologically advanced country, the loss was more than twice as much (on a per capita basis) as what was suffered in most other countries. In Japan, for example, which one tends to think of as having a large number of highly flammable structures packed close together, the per capita loss in 1971 was $2.60.

These figures do not reflect the loss of life in fires, the injuries from burns or the anguish associated with the deaths, the injuries and the economic disruptions caused by fires. Last year the President's Commission on Fire Protection and Control presented in its final report an estimate of the total social cost of unwanted fire. It found that in 1972 the number of Americans who died in fires was 12,000 and that the number sustaining serious burn injuries was 300,000. Moreover, taking into account not only property loss but also such factors as loss of productivity and the cost of operating fire departments, providing insurance and treating burns, the commission concluded that in 1972 the cost to society of fires in the U.S. was some $11.4 billion, or more than 1 percent of the gross national product. Still another consideration is that the yearly loss of forests due to fire is 4.8 million acres, an area approximately the size of New Jersey.

One is moved to ask why the U.S. has the world's highest loss from urban fire. Certainly the country has an ample network of fire departments trained in modern methods of fire protection and control; that network is the equal of any other country's. The reason for the high loss figures in the U.S. is the nation's high standard of living, which has given the American people a large array of appliances and conveniences that are potential sources of fire. The fire services have not made corresponding technological advances.

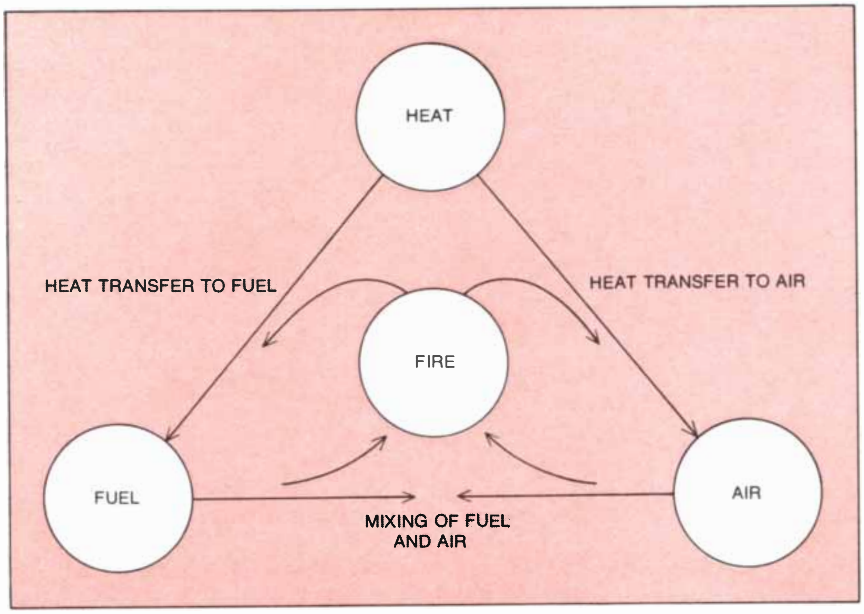
Surely the record can be improved by applications of science and technology. The problems to be attacked come under two headings. First, the physics of combustion in unwanted fires is poorly understood. Second, many of the techniques of fire protection and fire fighting now employed are clearly inadequate. Indeed, although the data on fires over the centuries are not very reliable, they do suggest that the success of fire-control measures is not significantly greater now than it was in earlier times.

The problems in this area are suggested by the remarkable discrepancy from country to country in judgments on what is flammable. For example, in the early 1960's each of six nations undertook (in cooperation with the International Organization for Standardization) to rate 24 wall-covering materials in order of their flammability according to that nation's standard test. The results disagreed widely [*see bottom illustration on next page*]. The serious nature of the disagreement between different "standards" is shown by material No. 18, a phenolic-foam wallboard; it was the safest of all 24 materials according to the standard test in Germany and the most hazardous of all 24 according to Denmark's test. On the other hand, material No. 7, an acrylic-sheet wallboard, was the third safest material by the Danish test and the third most flammable by the German.

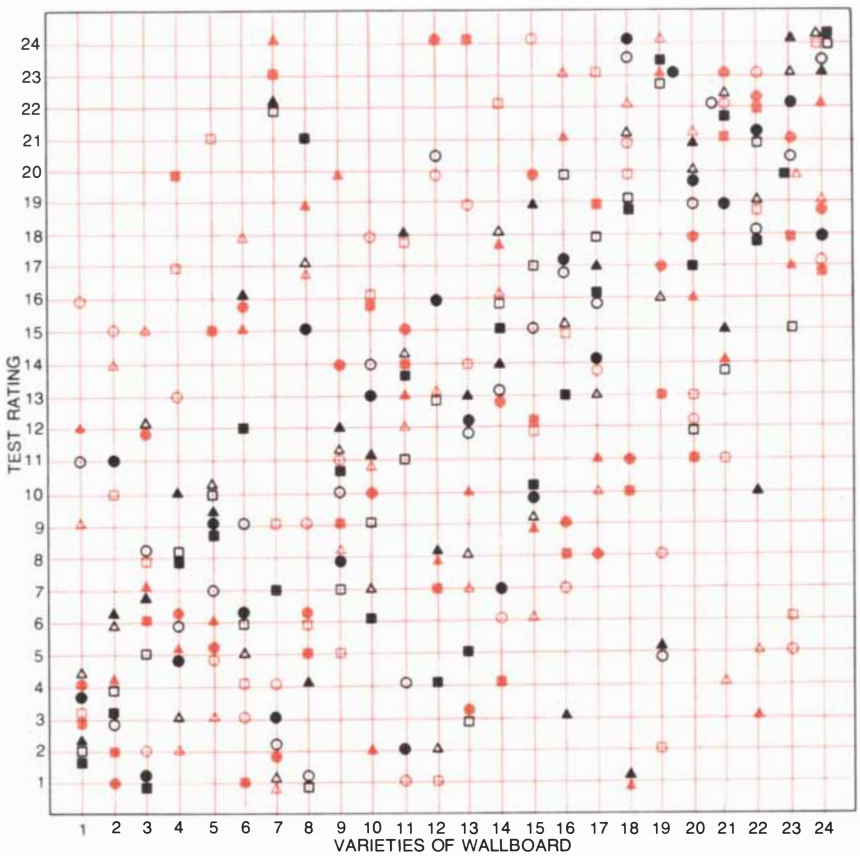
To demonstrate how bad these results really are I constructed six sets of meaningless fire test data by putting 24 small numbered cards in a box and taking them out at random six times. I then averaged and plotted the data as I had averaged and plotted the results of the flammability tests by the six nations. The scatter of the meaningless random data is only slightly wider than the scatter of the results from the real flammability test.

What the scatter of the test results means is that no one knows what characteristics a material should have to be safe in a fire. Each of the national tests was based on a best judgment rather than on rigorously established facts, and so each test measured something different under the heading of flammability. It was inevitable that the results would diverge.

The development of an improved approach requires that the basic nature of fire in a building be understood well enough to devise a test that measures the right characteristics. In fact, a little thought shows that the present approach



ELEMENTS OF FIRE are fuel, air and heat, which have the relation depicted in this triangle. Fuel and air are intimately mixed almost everywhere, but there is no fire until heat is added. Once a fire has started it continues to burn by means of a return to the fuel and the air of enough heat to gasify additional fuel, mix it with air and ignite the mixture.



WEAKNESS OF FIRE DATA is indicated by comparison of flammability tests (*black*) by six nations and random card tests (*color*) by the author. Each of the six nations tested 24 types of wallboard according to its own standard of flammability. The author drew 24 cards at random from a box six times. The scatter of the data is similar for both tests, indicating that the level of fire protection afforded by present national assessments of flammability is rather low.

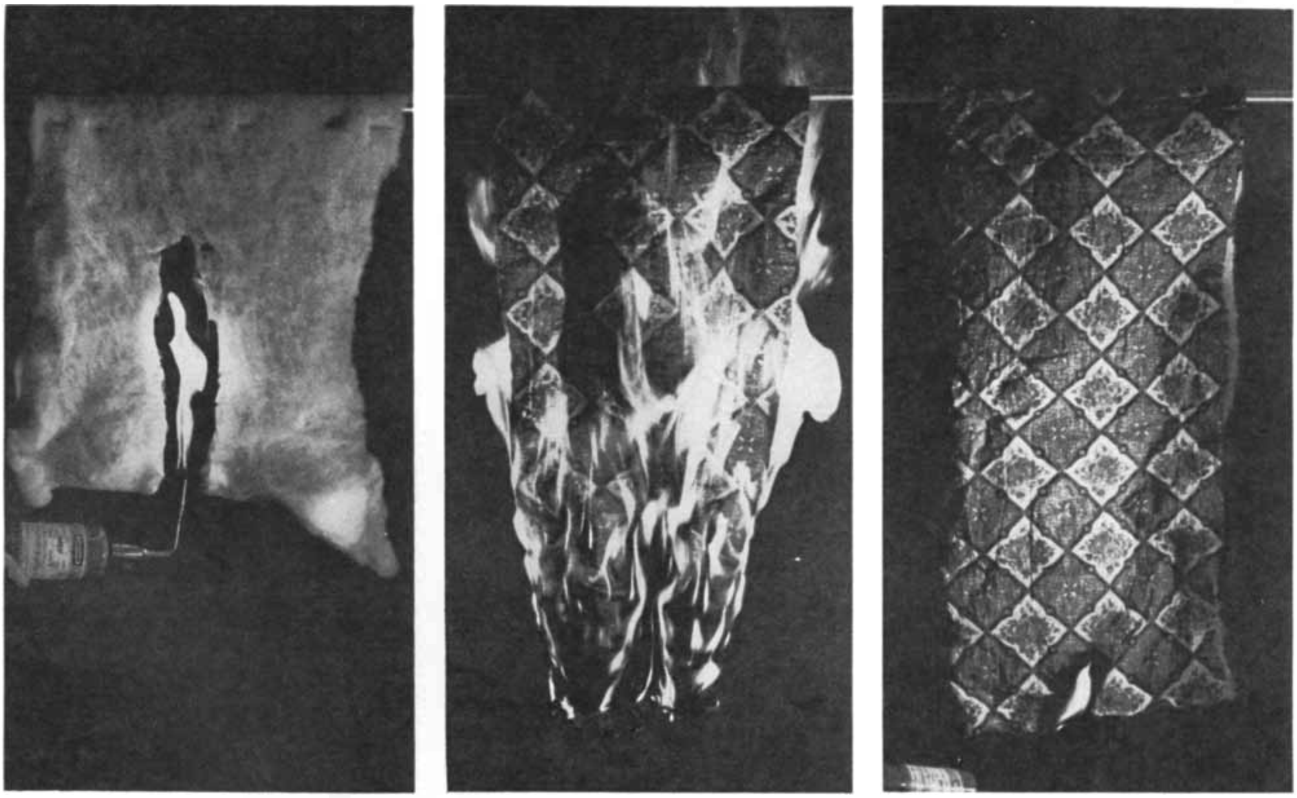
of assigning a single value to indicate the flammability of a material is basically wrong. As every Boy Scout knows, one cannot make a campfire with a single log; it takes at least two and preferably more. A single log does not have an inherent fire-safety measure. Only the entire system involved in a potential fire can be rated.

Eventually building codes will have to be rewritten to rate rooms and buildings rather than single materials. I say eventually because at present not enough is known to provide a basis for rating a room. Before such a rating can be made the materials themselves will have to be subjected to a number of separate tests that measure the ease of ignition, the rate of spread of fire, the production of smoke and toxic gas and other characteristics, which can then be combined for a given room into a measure of its flammability and hence of its acceptability in a home or a public building.

The elements of a fire are fuel, air and heat. Fuel and air are intimately mixed almost everywhere in the environment, but a fire does not start unless a spark or friction ignites a fuel-oxidizer system. A fire spreads because the heat feeds into new fuel and raises it to the ignition point. To extinguish a fire one removes one of the elements: by cooling the hot fuel or by separating the fuel and the air.

These simple ideas on the nature of fire are correct and useful, but they require quantification in order to supply the understanding needed to avoid inconsistencies of the kind represented by the various national standards on the flammability of materials. A start has been made toward a quantitative understanding of fire. In what follows I shall describe what has been learned so far and outline the problems that remain in the areas of ignition, pyrolysis, fire retardants, smoke and toxic gas, detection and convection.

On ignition the questions one wants answered are why fuel and air remain in contact for years (even centuries) without any apparent reaction and why heat applied locally upsets the equilibrium. It is characteristic of chemical reactions that they proceed faster as the temperature rises until they reach a maximum set by the required energy or the physical processes that bring the fuel and air together. On the other hand, whenever any material has a temperature higher than its surroundings, the conduction, convection and radiation of heat start to cool it off. These relations are portrayed in the



EFFECT OF RETARDANTS appears in these photographs of tests made with a commercially available comforter. At left a strip of the plastic-fiber fill from the comforter appears safe because it does not burn. The reason it fails to burn, however, is that the fibers melt away from the flame and drip off. A strip of the same fibers in cotton covers, which is the way the comforter is sold, burns fiercely (*center*) because the melted fibers stick to the burning covers. If the covers are treated with a fire retardant, the comforter is much more difficult to ignite (*right*). In this test by the author the retardant applied to the covers was monobasic ammonium phosphate.

accompanying illustration [*top of next page*], which indicates three points of equilibrium: the environmental or room temperature *R*, the fire point *F* and the ignition point *I*. They are the points at which the heat produced in the chemical reactions between fuel and air exactly balances the heat lost. Points *R* and *F* represent dynamically stable conditions, but point *I* represents an unstable equilibrium.

If at point *I* the balance is disturbed even slightly, so that the temperature of the fuel rises above *I*, then the heat released in the reaction exceeds the heat lost, and the temperature of the fuel continues to rise to point *F* without further external help. A fire is in progress. If, on the other hand, the balance is disturbed by a slight drop in temperature, the fuel will cool off to the room temperature, *R*. Point *I* is the unstable ignition point above which the temperature must be raised (as by a match) if a fire is to result.

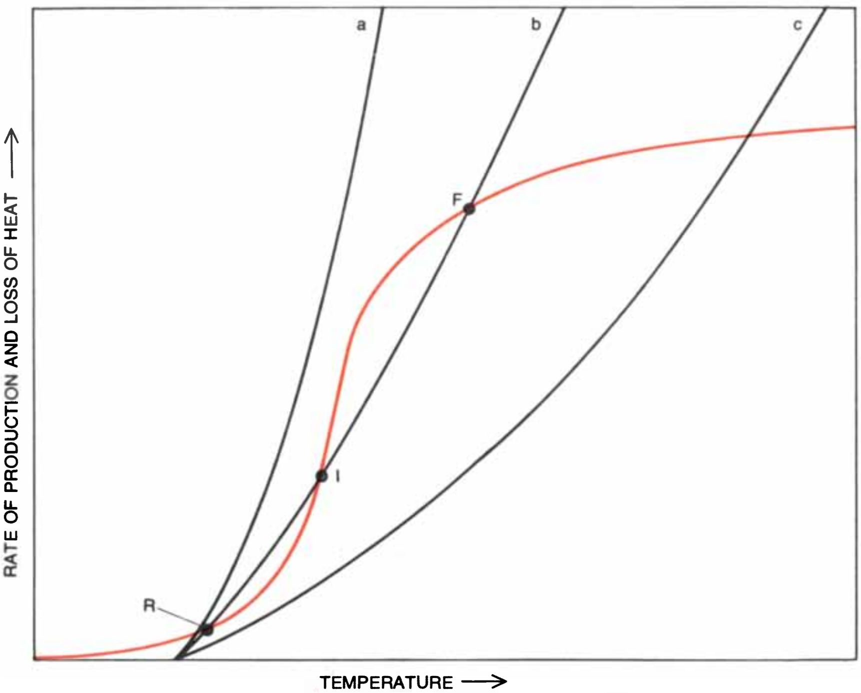
At point *F* the high temperature causes a high rate of release of energy and an equally high rate of loss of heat. If the temperature rises higher, the increased rate of reaction will release heat faster but the rate of loss of heat will rise faster still, so that the fuel will cool off until the production and loss of heat are again in balance at *F*. The fire is at a point of stable dynamic equilibrium.

A fire in a corner of a room heats everything in the room. As the temperature of a material rises, its atoms and molecules acquire increased kinetic energy. When the atoms of a solid vibrate too violently, the chemical bonds between them are broken and smaller molecules are evolved. This process, whereby a complex solid is thermally decomposed into simpler solids or liquids and ultimately into gases, is pyrolysis. Pyrolysis of a material is the essential first step in the ignition and burning of the materials from which buildings are made.

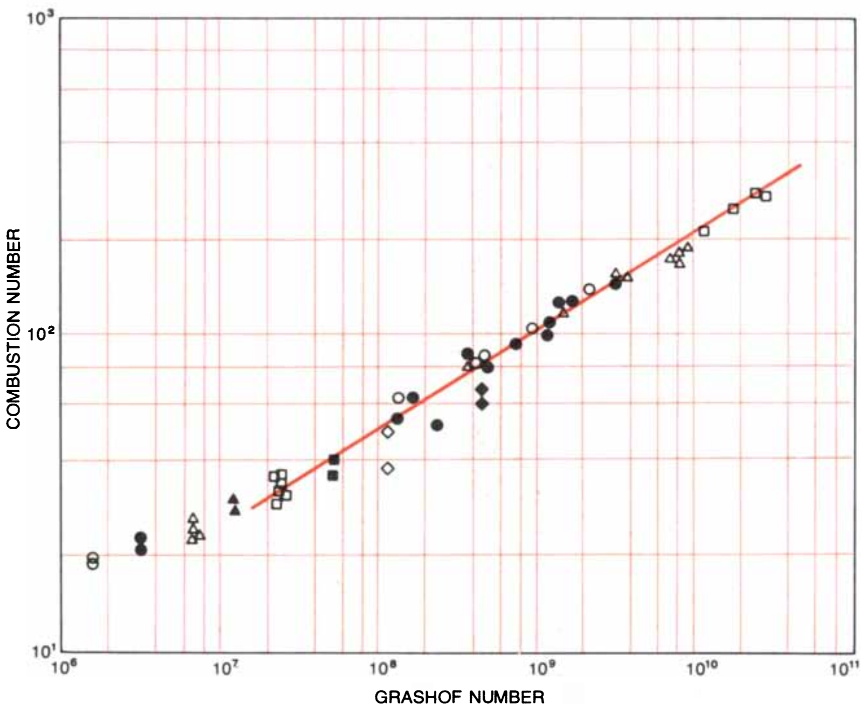
In order to select materials that are fire-safe from the point of view of ignition, it would be desirable to devise a test for the processes of chemical reaction and heat loss I have described. The processes of heat loss are adequately understood, but measuring the rate of chemical reaction proves to be difficult because of the complex chemical and physical steps involved in the pyrolysis and burning of most materials. For flammable liquids pyrolysis is replaced by evaporation, and a simple test has been found that is reasonably satisfactory. A standard cup of the test liquid is exposed to a small pilot flame. At low temperature nothing happens. At a higher temperature the vapor above the liquid burns in a flash and goes out. At a still higher temperature the liquid supplies vapor fast enough for continuous burning. Two temperatures are thus defined: a flash point and a fire point. A liquid with a low flash point, such as gasoline, must be handled with care, whereas lubricating oil, with a significantly higher flash point, is much safer.

What about wood and the various other materials of construction? No correspondingly simple test that is fully satisfactory has been found. Furthermore, the chemical processes responsible for ignition are exceedingly complex.

Anyone who has observed wood burning has seen it darken, has seen and heard gases issue from cracks and burn as flames and has seen a layer of glowing



IGNITION OF FIRE requires certain physical conditions. The colored curve portrays the release of heat by the reaction of fuel and air in a fire. *R* represents room temperature, a normal equilibrium point. *F* is the stable fire point at which a fire will continue to burn until its fuel or air is depleted. *I* is an unstable ignition point. The black curves show how heat is lost. With excessive cooling (*a*), as from a water spray, the rate of loss of heat is high and a fire is impossible. Normal heat loss is indicated by curve *b*. With a low heat loss, resulting from good insulation or a highly reactive fuel, ignition can be spontaneous (*c*).



RATE OF BURNING in a fire is shown to depend on the Grashof number, which is derived from the cube of the size of the specimen, the square of the air pressure and fuel-vapor properties. In the experiments reflected by this curve the specimens were Plexiglas cylinders from two inches to two feet long. They were burned at pressures of from one to 40 atmospheres. The significance of the curve is that rate of burning is controlled mainly by convection and processes of mass transfer rather than by chemical and radiative effects. Hence it is possible to test certain effects of combustion by burning models in a laboratory.

char accumulate on the surface. What cannot be seen are the processes, chemical and physical, by which the large molecules of cellulose and lignin break up to form smaller molecules ot gas and the manner in which the gases either move deeper into the wood and partly condense or accumulate locally to raise the pressure and crack the wood, perhaps even blowing off a piece of glowing char. The char itself shrinks as it forms, thereby producing additional cracks.

No one of these processes would be too difficult to evaluate quantitatively if it were not for the contradictions that arise when they all operate at once. The measured chemical-reaction rates and the heat of reaction depend on the size and shape of the piece of wood and are highly sensitive to small amounts of mineral impurities. Even the products of the wood's own decomposition affect the burning processes. The heat released as cellulose pyrolyzes varies from 88 calories per gram endothermic (heat absorbed by the process) to 400 calories per gram exothermic (heat released). Although explanations for these effects have been proposed in terms of chemical decomposition rates and self-catalysis, none of them provides the quantitative predictive capacity required to devise a reliable test of flammability.

The sensitivity of pyrolysis to small amounts of contaminants provides a basis for work with fire retardants. Flammable materials are treated with substances that decrease their susceptibility to fire. The complexity of the chemical processes involved has so far prevented the prediction of which chemical will retard the combustion of a given fuel. It is known, however, that some (not all) compounds of nitrogen, phosphorus, the halogens (fluorine, chlorine, bromine and iodine), antimony and potassium retard the ignition or the burning (sometimes both) of various materials.

By directing pyrolysis away from the production of large amounts of flammable gas and toward the production of char and water vapor, retardants exert a profound influence on the hazardousness of cellulose and plastic material. Demonstrations can easily be made of the effects of retardants on the flammability of fabrics and other materials [*see illustration on preceding page*]. Since it is a relatively simple matter to greatly improve the fire safety of clothing, bedding, drapes and the like, there is a need for further legislation barring from the marketplace all such products that have not been adequately treated with retardants.

Smoke and toxic gases result either because in a fire many materials are heated without sufficient air to burn completely or because the products of pyrolysis do not encounter a source of ignition and therefore fail to burn. In either case the smoke (consisting of small, condensed drops of liquid or of solid particles) and the gases move away from the fire. They burn if they are ignited elsewhere, or they settle on and damage walls, floors and furniture, or they prevent the escape of occupants. In fact, few people are burned to death; most of the victims of fire die from inhaling smoke and hot gas before flames reach them.

The complex processes of pyrolysis give rise to a large number of chemical compounds that appear in the smoke. Chromatographic analysis of smoke from cellulose has revealed the presence of some 175 different organic compounds, of which about 150 have been identified. The products formed in any particular fire depend on the configuration of the fuel and the fire and on the chemical nature of the fuel and the retardants and other impurities it contains. Each new material employed for construction or furnishings brings a new set of smoke hazards. Much work remains to be done on identifying the toxic products of fire and on ascertaining their effects on people. It is equally important to learn how to control the production of smoke and to select building materials that produce less harmful substances in a fire.

If all else fails to prevent a fire, the best hope is to detect it soon after it starts. The detection devices now available operate by hot gas, by a high rate of rise in temperature, by the flicker of flame, by obscuration of light by smoke or by the presence of smoke particles. None of these methods is entirely satisfactory. All the devices have too high a false-alarm rate and are too expensive for widespread installation.

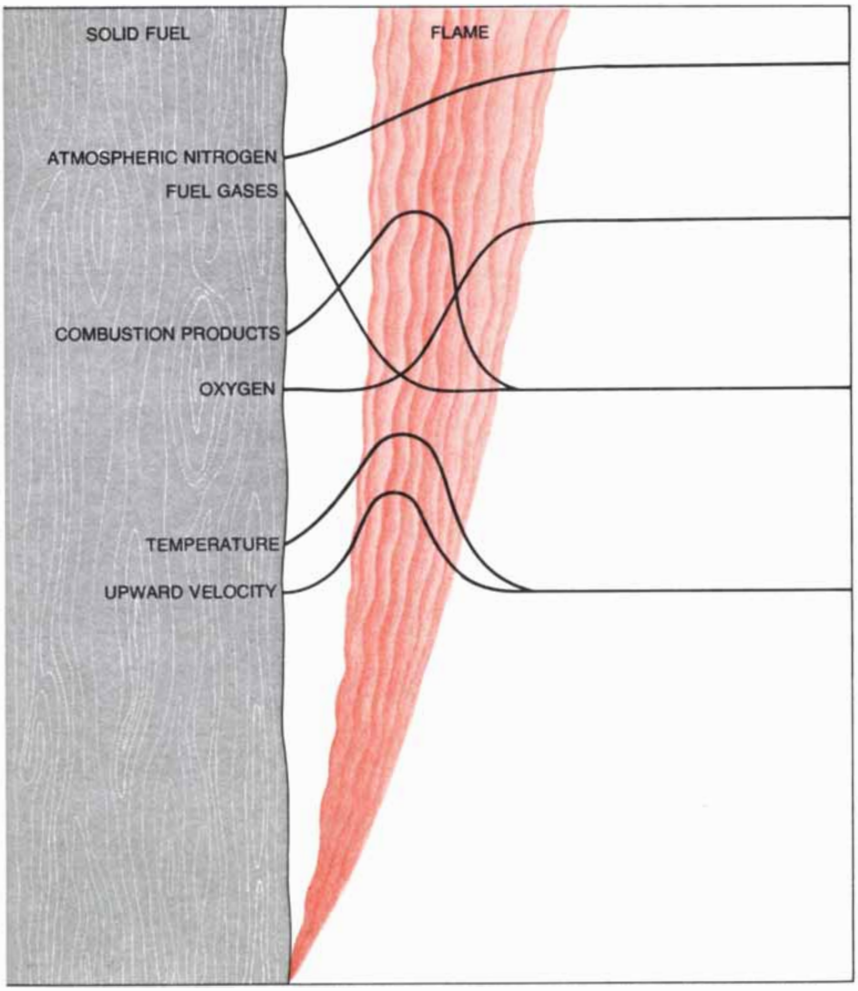
In almost every area of measurement man's instruments are superior to his senses, but not in fire detection. No practical detector now made can distinguish reliably between tobacco smoke, wood smoke and smoke from an electrical appliance with anything like the acuity of the human nose. One can anticipate, however, that advances in solid-state electronics will eventually result in detectors that are sensitive, reliable and inexpensive. Before that can be done much more needs to be learned about the chemical compounds produced in the pyrolysis of materials commonly found in buildings.

The heat released in a fire expands the gas next to it and the gas rises. It is this buoyant convection that draws fresh air to the fire, enabling it to spread. Convection also carries smoke and heat from the fire to a detector, be it an instrument or the human nose. In addition it carries smoke and toxic gases that endanger occupants and destroy property in other parts of the building.

Since convection is a relatively simple dynamic process, the discipline of fluid mechanics (with the aid of computers) makes it possible to calculate the movements of the gas. The success of such calculations is demonstrated by the similarity of predicted and photographed streamlines for the simple case of the convection produced by a hot spot in the floor of a cylindrical room. The close agreement may suggest that this part of the fire problem, at least, is understood. Unfortunately, although progress is being made, we are still far from being able to predict the behavior of real fires.

The difficulties that remain are of two kinds. First, the calculations of the simplest two-dimensional convection problems take a modest amount of computer time and hence are obtained at a modest cost, but a real fire is three-dimensional and constantly changing. Adding these effects makes the calculations long and prohibitively expensive.

Second, convection in a real fire is not just a simple buoyant flow from a single hot spot. Many chemical reactions proceed in the flames; the many flames add heat throughout the gas; the walls of the room gradually heat up and begin to burn, and the motion of the gas, instead of being smooth (laminar flow), consists



PROCESSES OF COMBUSTION are portrayed. In a vertical fuel surface the combustion processes are all in a thin boundary layer near the surface. The curves indicate by their height where the various processes are most intense. Heat from the flames is conducted to the surface, producing fuel vapors that diffuse away from the surface. They meet oxygen that is diffusing inward and react to sustain the flame. Buoyancy of hot gases moves flame and combustion products upward and maintains a fresh supply of air near the surface.

of many nonsteady eddies (turbulent flow).

These effects add so much complication to the predictions that an exact computer solution cannot be obtained with the computers now available. Present work in the calculation of convection in fire is directed toward finding simpler ways to add the complex effects in the right place and time so that the computation of sufficiently accurate answers will become practical. In addition the numerical methods by which the computer solves convection problems are under study in an effort to find ways of speeding up the required calculations.

The rate of burning of the fire has not yet been incorporated in the general convection problem. It has, however, been carefully studied under certain simpler circumstances. If a vertical wall is burning steadily, the rate of burning can be calculated. In such a wall the flame covers the surface but does not in general touch the surface. The fuel heated by the flame pyrolyzes; the gases thus generated can burn only after coming out of the fuel and mixing with oxygen. The oxygen in the air is consumed by the flame, so that there is no oxygen next to the surface. What happens is that the vapors from the fuel blow the flame away from the surface just far enough so that the heat transferred from the flame back to the surface of the fuel is sufficient to produce the fuel vapors by pyrolysis. The balance between heat transfer to the surface and fuel transfer to the flame regulates the rate of burning independently of the chemical-reaction rates.

In the flame, fuel and oxygen react to generate hot products of combustion. Being hot, they are less dense (lighter) than air and so are carried upward by convection. The flames generally extend above the top of the burning surface (or along the ceiling if the fuel is the wall of the room). Therefore one finds a variation of temperature and velocity through the flame.

Since all these effects are found in a relatively thin layer next to the wall, the entire process can be calculated by what is known in fluid mechanics as boundary-layer theory. The calculated rate of burning agrees well with the measured one.

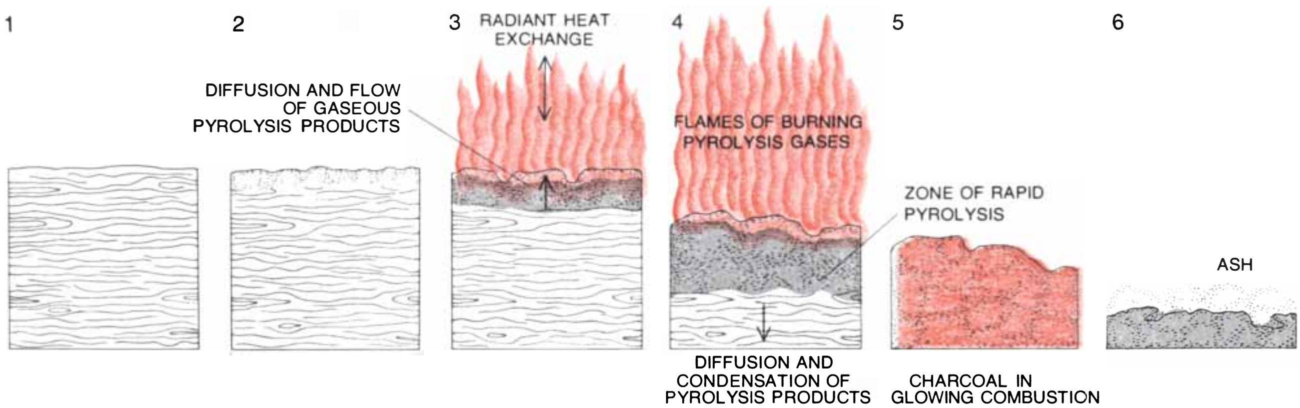
When the surface is as large as the wall of a room, the burning layer becomes highly turbulent, and exact calculations are no longer possible. The simpler calculations, however, show that the rates of flow and mixing control the process, and therefore the burning rates should depend only on the dynamic processes of diffusion and velocity that mix the fuel and air. This concept works well for vertical cylinders of Plexiglas burning in air [*see bottom illustration on page 24*]. The burning rates of many different sizes of cylinder all fall on the same line when the data are plotted against the Grashof number (a measure of the importance of the convection of gas produced by the buoyancy of the gas; it is named for Franz Grashof, a 19th-century German engineer). With curves like this one for various materials, it is possible to begin to estimate how fast a fire will burn and hence how fast a burning building will be destroyed.

Another important development in fire technology is illustrated by the simple relation between the burning-rate variable (mass per unit width of surface divided by the viscosity of the gas) and the Grashof number. The Grashof number incorporates the cube of the length of the specimen multiplied by the square of the gas pressure; its value depends only on this product and not on the length of the specimen or the pressure alone. With this fact one can investigate how a large object burns by burning a small one inside a pressure tank. It is cheaper and quicker to test on a laboratory scale than to burn houses. Nonetheless, the testing of models always has limitations. The limitations of the model tests I have described are now being evaluated in work being carried out at the Factory Mutual Research Corporation in a joint project with Harvard University.

Model testing and the scientific understanding of the way in which the laws of nature conspire to produce a violent unwanted fire must finally be validated by full-scale tests. As I have mentioned, flammability is a property of an entire room or building, not of a single material. Full-scale fires must be studied to find out what is really important.

My colleagues and I have conducted such a test with a bedroom [*see illustrations on opposite page*]. We chose a bedroom because fires caused by people who fall asleep while smoking in bed are a major cause of loss of life in fires. Usually a fire that starts in a bed flashes over in about eight minutes, that is, it spreads rapidly throughout the room, the smoke level drops to the floor and flames come out of the door (or window) with a roar.

In our test the flashover came much more slowly. Instead the fire on the bed receded and then spread slowly to the bedstead and the rug under the bed.



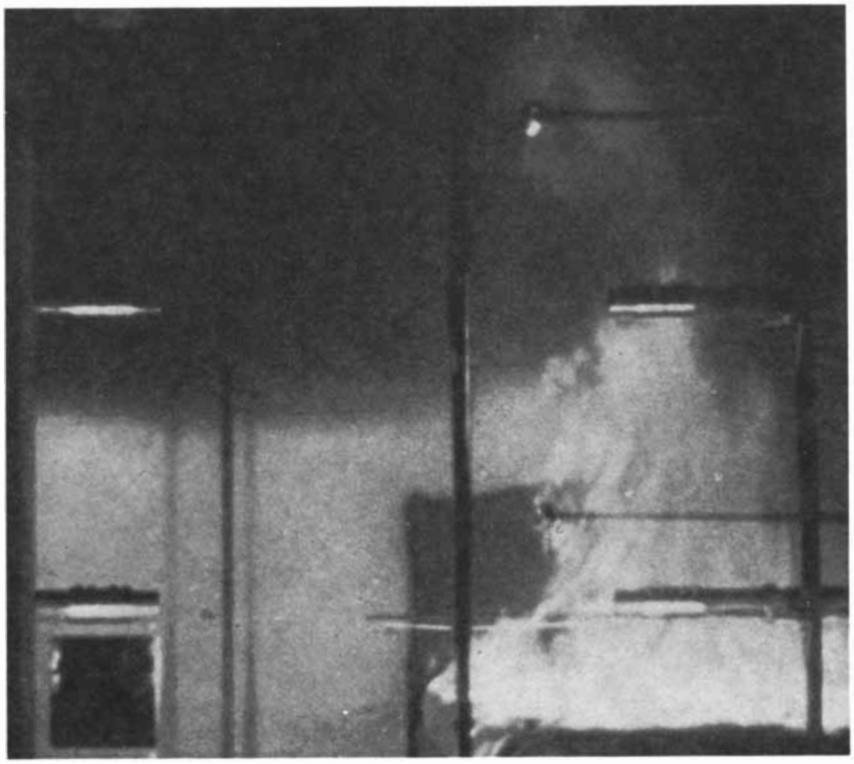
STAGES OF BURNING of a piece of wood are depicted, starting with the original surface (*1*). The first step in the destruction of the wood is the blistering of the surface (*2*). Then active pyrolysis begins (*3*) and moves slowly through the wood, with a radiant heat exchange between the wood, the flames and the environment. Heat from the surface is conducted through the char layer to maintain active pyrolysis in the interior of the wood. Some of the pyrolysis products move farther into the interior and condense, and some continue to pyrolyze into simpler gases while diffusing and flowing outward through pores and cracks in the char to mix with air and burn (*4*). After the wood has been pyrolyzed the charcoal continues to glow (*5*). A small content of minerals remains as ash (*6*).

When these items became fully involved, 17 minutes and 35 seconds after ignition, the fire flashed over.

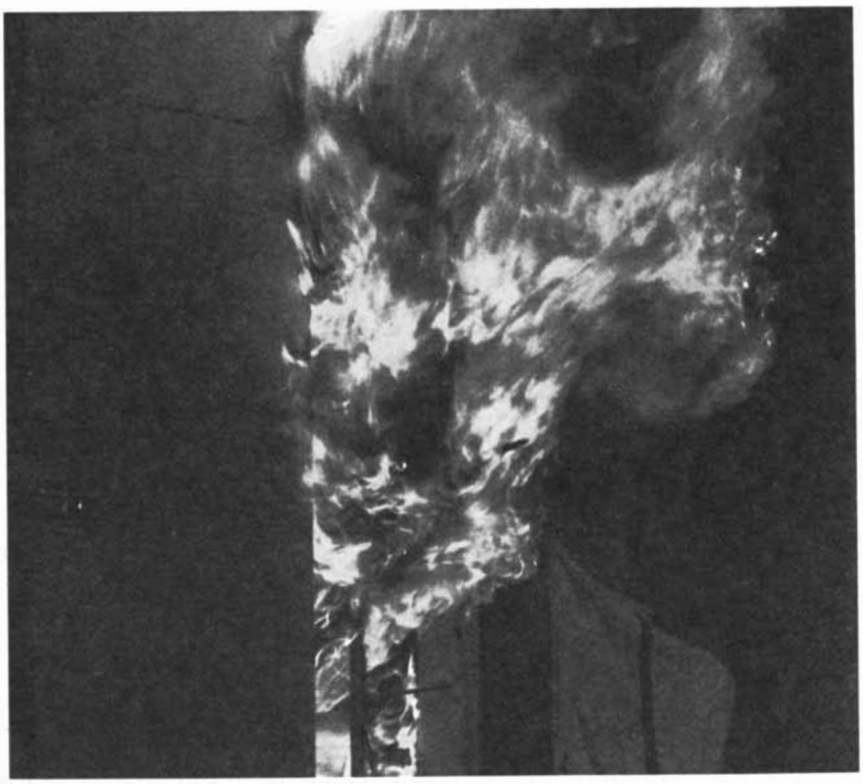
The time of flashover is critical. By then the value of essentially everything in the room is destroyed. Until then a person in the room can crawl to safety. To stand up would be fatal, since the top half of the room contains hot, smoke-filled air. One breath of that air does serious physiological damage, although it is not fully known exactly what happens or what antidotes might be applied. Moreover, in our test the change from the situation in which the fire was confined to a corner of the room and the smoke to the upper half of the room to the flashover situation occurred within five seconds and without warning. Because of this hazard one should never return to a burning room no matter how safe it looks.

We are not sure what accounted for the delay in flashover in our test. Only further study of the mechanisms by which fire grows can supply a definite answer. Nonetheless, the observations we made during the test offer a hint. Pieces of paper at various places in the room did ignite at about eight minutes. At six and a half minutes, however, the plastic-fiber curtains beside the bed melted and fell to the floor. If they had remained in place, they too would have started to burn at about eight minutes. The heat of combustion from them, added to the heat already in the room, probably would have caused a general flashover. If such a simple change as replacing flammable curtains with fire-resistant ones can cause such a significant delay in the time of flashover, it could save many lives and much property, since the occupants of a room thus prepared would have twice as much time to escape and firemen would have twice as much time to respond effectively.

Fire as a wasteful, destructive force in human affairs presents a host of problems. They include financing fire departments, promulgating codes and standards, training firemen, devising strategy and tactics for fire fighting, providing insurance and obtaining the technical and scientific facts that underlie many of the decisions that must be made. Since fire strikes without warning, the fireman and the fire-protection engineer must do the best they can with the knowledge available at the moment. It will take many specialists in many fields, working many years, to provide all the knowledge and data required to make the urban environment fire-safe.



FIRE TEST was conducted in a bedroom to obtain quantitative information on what happens in a full-scale fire. The scene here shows the room eight minutes after ignition. Usually at about that time the fire would flash over, that is, spread rapidly throughout the room. In this case it receded temporarily and did not flash over until almost 10 minutes later.



FLASHOVER IN BEDROOM started 17 minutes and 35 seconds after ignition. The delay in flashover, which would have given occupants more time to escape and firemen more time to respond, was thought to be due in part to the fact that the plastic-fiber drapes melted and fell to the floor in six minutes. Retardants in drapes might similarly delay flashover.

The Sex-Attractant Receptor of Moths

The sex attractant of the female silk moth is detected by an array of receptors on the feathery antennae of the male. A nerve impulse in a receptor cell can be triggered by one molecule of attractant

by Dietrich Schneider

All sensory systems, each with its specific peripheral receptor cells and its integrating neurons in the brain, are designed to form a biologically relevant image of the outer environment of the organism. In many animals chemical signals play a major and often decisive role as a means of communication. The domesticated silk moth, for example, cannot fly and therefore the male cannot readily scout the terrain in search of a mate. It does, however, have two featherlike antennae that are finely tuned to detect a certain chemical compound: the scent emitted by the female. So sensitive are the male's antennae that one molecule of the vaporous sex attractant will trigger a nerve impulse in a receptor cell. When approximately 200 nerve impulses have been generated in the span of a second, a message is received in the moth's brain and it moves upwind to claim its mate.

Chemical compounds such as the sex attractant of the silk moth, which are secreted by an animal and elicit a specific kind of behavior in animals of the same species, are called pheromones [see "Pheromones," by Edward O. Wilson; Scientific American, May, 1963]. A major difficulty in studying the olfactory system in most animals is the enormous variety of chemical compounds that elicit a response. The situation is different in the olfactory pheromone-receptor system of the silk moth. Here we have found a rather simple olfactory system where many receptor cells respond identically to only one compound.

In the past 20 years interest in pheromones has grown steadily, and the number of identified pheromones has rapidly increased. This holds particularly for the pheromones of insects because here the investigator's curiosity is augmented by the hope of putting pheromones to work as lures in the control of insect pests [see "Insect Attractants," by Martin Jacobson and Morton Beroza; Scientific American, August, 1964]. In most cases sex-attractant pheromones are produced and emitted by the female in order to lure a mate. Sex attractants are widespread in the insect world, but they are also found in other animal classes, as is well known to every owner of a female cat or dog in heat.

The impressive attractiveness of a virgin female moth to its male partners was described as early as the 18th century by naturalists such as René Antoine Ferchault de Réaumur, F. Ch. Lesser and August Johann Rösel von Rosenhof. At the beginning of the 20th century the psychiatrist and entomologist Auguste Forel reported that when some wild European female silk moths emerged from the pupa in his studio in Lausanne, they attracted large numbers of male moths to the windows (along with a large number of *Gassenbuben,* or street urchins, who were attracted by the spectacular assembly of moths).

Although most of the early observers of the sexual attraction of male moths to females agreed that an odorous signal was involved, some were in doubt because of certain puzzling facts. Not only was the human nose unable to detect the alluring odor but also it was hardly conceivable that the amount of odorant that could be produced by the female would be able to lure the males from a distance of at least a kilometer. The critical experiment demonstrating that the attraction was definitely based on odor was described in 1879 by Jean Henri Fabre in his famous *Souvenirs Entomologiques.* When Fabre picked up a female moth and put her under a glass hood, male moths that flew into his house paid no attention to her but went to the place where she had been sitting a short time earlier. Although Fabre realized that the female's scent was guiding the male at short range, he still believed unknown radiations from the female lured and guided the male from greater distances.

Other experiments with male moths clearly showed that the presumed olfactory faculty of these insects is localized in the antennae, since the males did not react to the "calling" female after their antennae had been removed or covered with varnish. This was the state of knowledge about these moth sex attractants until about 20 years ago. Then advances in chemical, physiological and histological methods led to the identification of the chemical nature of the luring substance and to the definite proof that the receptors are on the male moth's antennae.

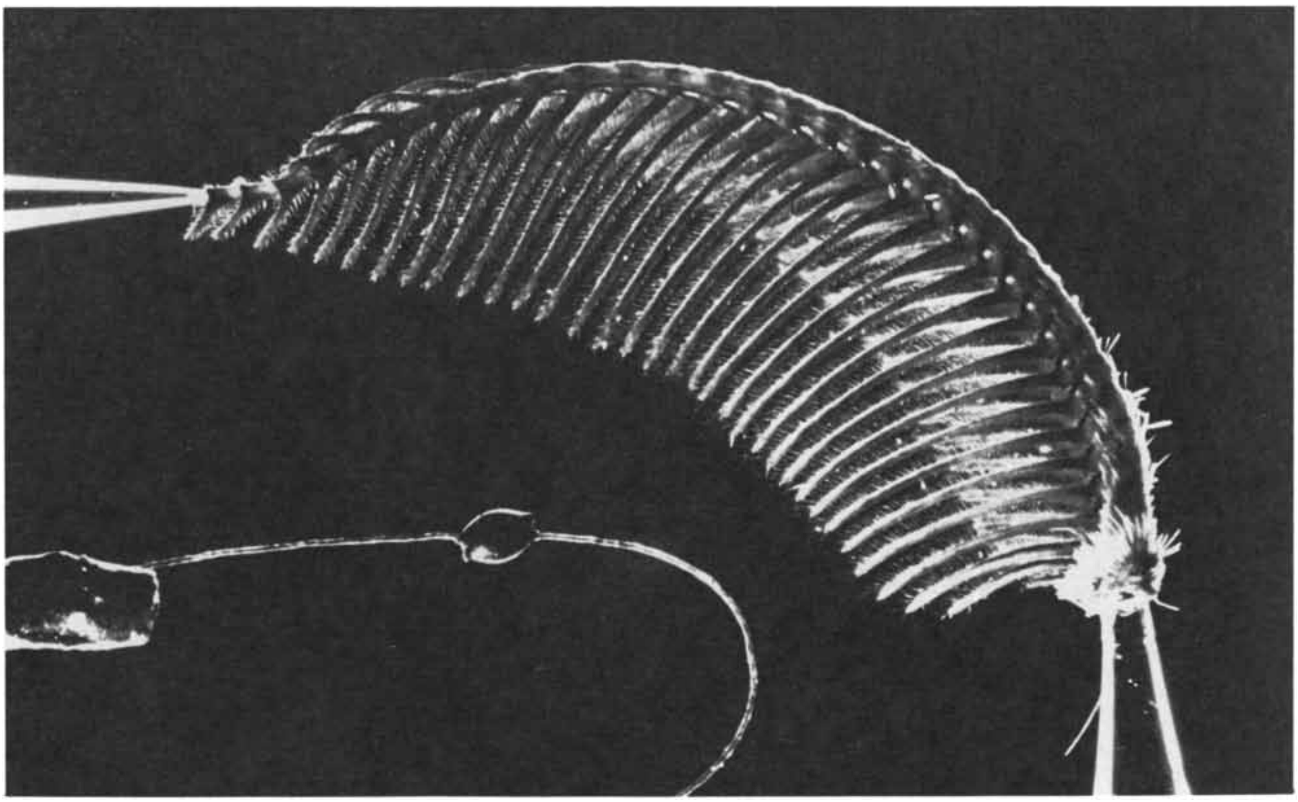
The first sex-attractant pheromone to be chemically identified was the lure substance of the female of the commercial silk moth *Bombyx mori.* Adolf F. J. Butenandt and his co-workers at the Max Planck Institute for Biochemistry in Tübingen (and later in Munich) reported in 1959 that the attractant has a chain of 16 carbon atoms and is a doubly unsaturated fatty alcohol (later identified as *trans*-10-*cis*-12-hexadecadien-1-ol). They named the compound bombykol. Their success was the outcome of long years of difficult analytical work, and their choice of *Bombyx mori* was a wise one. This moth is bred for silk production in many parts of the world, so that it is possible to obtain large quantities of the female glands that manufacture the compound. In order to extract 12 milligrams of pure bombykol the biochemists needed the glands of half a million moths.

In the early 1950's, when only enriched extracts from female *Bombyx* glands were available, I met with my biochemical colleagues in Tübingen and was challenged by the problems of olfactory perception in the silk moth. I thought that insight into the highly specific olfactory function in this animal might lead to a better understanding of the still unknown mechanisms of olfactory perception in general. My research began as a one-man enterprise but later involved my students and associates. Members of the silk moth research team in my laboratory at the Max Planck Institute for Behavioral Physiology in Seewiesen are the biologists K.-E. Kaissling, E. Kramer, E. Priesner and R. A. Steinbrecht and the chemist G. Kasang. My recent studies of the gypsy moth and the nun moth were done in collaboration with the biophysicist W. A. Kafka. In our research we hoped to approach an answer to questions about the threshold and the dose-response functions of the odor receptors, the specificity of the receptors, the mechanism of odor-molecule capture, the mechanism of stimulus transduction and the fate of the odor molecule after it has transferred its information.

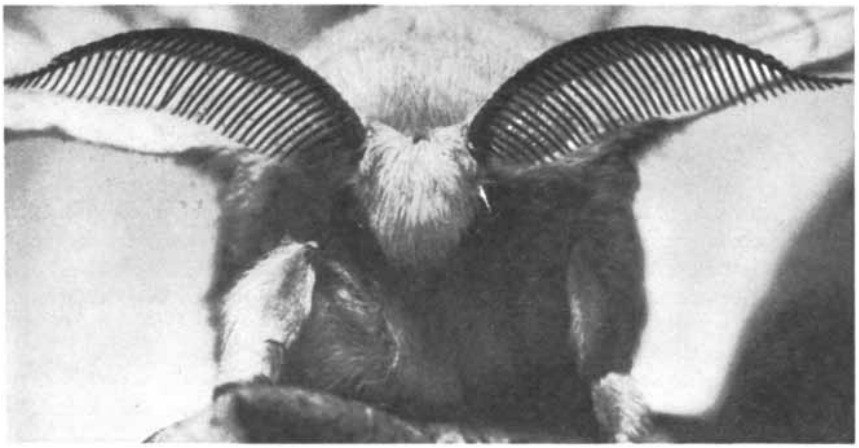
The peripheral part of any sensory cell reacts to an adequate stimulus (chemicals, light, mechanical displacement or temperature) with a temporary change in the electric charge of its membrane. This response is called the receptor potential. It can be recorded from whole sense organs, provided that the sensory cells are lined up rather like an array of interconnected electric batteries. In a silk-moth antenna simultaneous receptor potentials of many olfactory cells can be recorded simply by mounting the antenna between two electrodes connected to an amplifier and a recording instrument [*see illustration below*]. The antenna is stimulated by putting an odor source into a glass tube and blowing air through the tube onto the antenna. We tested the response of the male silk-moth antenna to various concentrations of bombykol, of steric isomers of the compound and of homologous fatty alcohols. Although all these compounds generated a response of the olfactory cells, none was nearly as effective as the natural material. We also found that the dose-response curve of this olfactory system covers a wide range of stimulus intensities, as do the response curves of the visual and auditory systems [*see bottom illustration on page 31*]. Interestingly, the antenna of the female silk moth does not respond to bombykol.

Microelectrode probes revealed that only those receptor cells connected to the specialized long hairs of the antennae respond to bombykol and its isomers. The amplitude of the discharge of the receptor cells increases with increasing concentration of the stimulating compound. This receptor potential generates a series of nerve impulses that travel to the olfactory center of the brain. The frequency of these nerve impulses depends on the amplitude of the receptor potential.

We then asked the following key questions: How many bombykol molecules are required to elicit the male's behavior response and what is the minimum number of molecules a cell needs to generate a nerve impulse? Before we could deal with these questions, however, we had to collect information on the amount of bombykol in the stimulating airstream and on the number of molecules being adsorbed on the antenna and on the hairs. For this purpose we resynthetized the pheromone in a form that incorporated tritium, the radioactive isotope of hydrogen. The measurements we then made with the tritium-labeled bombykol yielded surprising results: more than 25 percent of the bombykol is



ANTENNA OF THE MALE SILK MOTH has some 17,000 long odor-receptor hairs on its branches. The electrophysiological response to an odor can be measured by mounting an isolated antenna between two glass-capillary electrodes (*top left and bottom right*) and passing over the antenna a stream of air containing the odor. The oscillographic record of the changes of electric potential in the antenna is called an electroantennogram. The wire loop at lower left holds a thermistor that measures the airflow past the antenna. The photograph was provided by K.-E. Kaissling of the Max Planck Institute for Behavioral Physiology in Seewiesen.

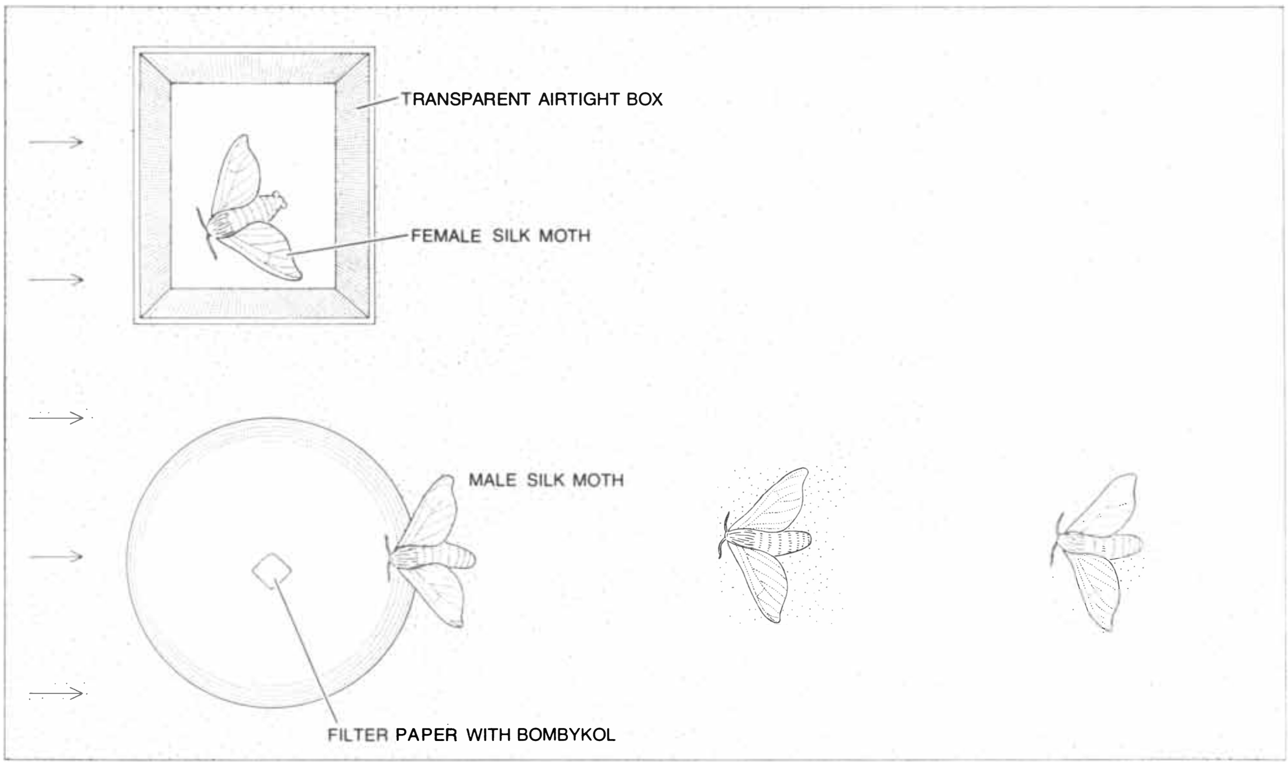


MALE SILK MOTH of the species ***Bombyx mori*** is seen from the front. The species is the commercial silk-producing one; it has been domesticated for some 4,000 years. Fifty percent of the odor-receptor cells in the male's antennae are tuned to respond to a single substance: bombykol, the sex attractant emitted by the female silk moth when it is ready to mate. The length of each of the male's antennae is six millimeters, or about a quarter of an inch.

TIP OF THE ABDOMEN of the female silk moth holds a pair of glands, the *sacculi laterales,* that contain about one microgram of the sex attractant bombykol. The glands shown here are in an expanded active state.

filtered out of the airstream when it hits the antenna.

We next conducted behavior tests and recorded electrophysiological signals from single odor-receptor cells in *Bombyx* males. When the male silk moth senses the sex attractant, it responds by fluttering its wings. A barely noticeable but nonetheless significant response is observed when the stimulating airstream contains about 1,000 bombykol molecules per cubic centimeter. Such a stimulus is produced by an odor source of only 3×10^{-6} microgram of bombykol. Within a second, which is somewhat more than the insect's reaction time, approximately 300 of the odor molecules are adsorbed on the 17,000 sensory hairs of the antenna. Each hair is innervated by the dendrites, or fiber endings, of two bombykol receptor cells. In this situation Poisson probability statistics, which mathematicians use to distinguish between different kinds of random events, enable us to predict that during that one second only two of the hairs receive double hits of bombykol molecules. The rest



EXPERIMENTAL ARRANGEMENT for testing the sexual attractiveness of an odor is depicted. Male silk moths are placed downwind from a fan; the direction of the airflow is indicated by the arrows. The male moth cannot fly and is always sexually responsive. A female silk moth is placed in an airtight transparent box. A small piece of filter paper soaked with bombykol is put on a glass dish near the female moth. When the males sense the sex attractant, they move upwind to the source of the odor and not to the visible female. Bending of the abdomen of the male moth is a copulatory movement regularly observed with strong stimulation.

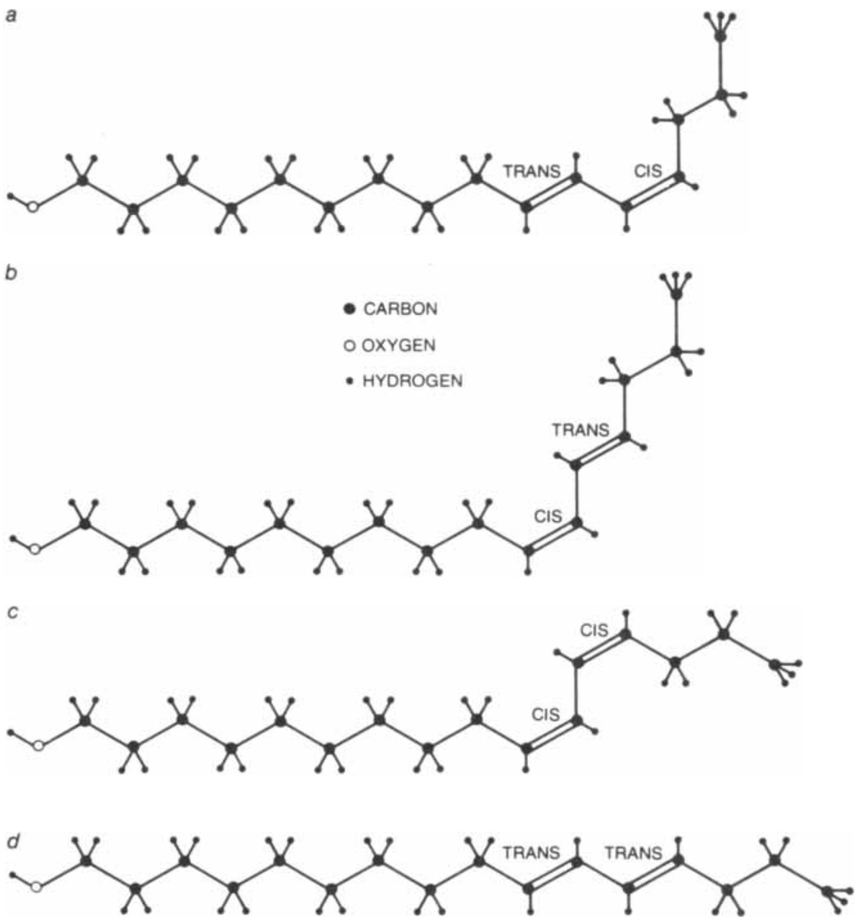
receive single hits. This observation already made it highly probable that the receptor cells are activated by single bombykol molecules.

We now recorded olfactory-nerve impulses from hundreds of single bombykol receptor cells. The stimulating technique was essentially the same as the one we had used during the behavior tests, which enabled us to compare the two types of experiment. With a bombykol-source load of 10^{-4} microgram, one in three hairs adsorbs somewhat more than one molecule per second and fires about one impulse. Both of these values are averages.

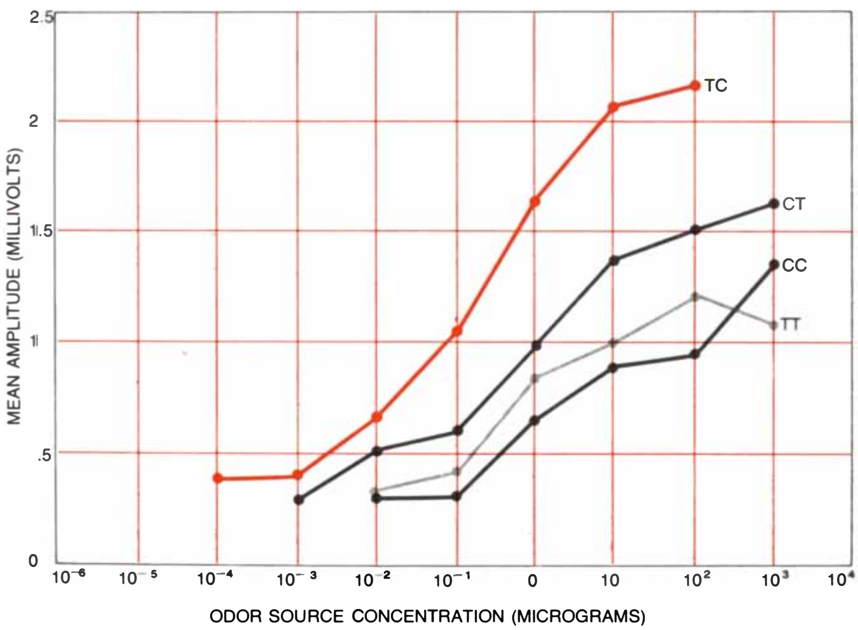
How are the impulses distributed? After subtracting the spontaneous activity of the cells we plotted the experimental data against the Poisson curves. The responses were separately plotted for the cells that reacted with one impulse or more, two impulses or more and 10 impulses or more. The one-impulse curve and the two-impulse curve exactly fitted the "one hit" and "two hit" Poisson probability curves, whereas the 10-impulse curve did not correspond at all to the 10-hit probability curve [*see illustration on page 35*]. This shows that the small impulse numbers near the response threshold are randomly distributed, as one would expect to be the case with the stimulating molecules also. Behavior and single-cell responses therefore allow us to state that a single nerve impulse is generated when one molecule of bombykol hits a receptor and that two impulses are generated when two molecules hit. The transduction of hits into impulses is obviously more complex with higher hit rates per second.

One difficulty with the radioactivity measurements on which the calibrations of these experiments are based was the limited sensitivity of even the best measuring instruments available. The detection limit for tritium-labeled bombykol with these devices is 3×10^{-8} microgram, or about 10^{8} molecules. That is enough for the measurement of the bombykol sources, but much smaller amounts are adsorbed on the antenna. Extrapolation is necessary to determine what the threshold amounts are. Our extrapolation is based on the assumption that the constant relation between the odor-source load and bombykol adsorption that is found in the range of the stronger stimuli holds true for the low rates near threshold.

We are now fairly certain that near the threshold level the receptors actually count the stimulating molecules. The cell can be said to be so sensitive



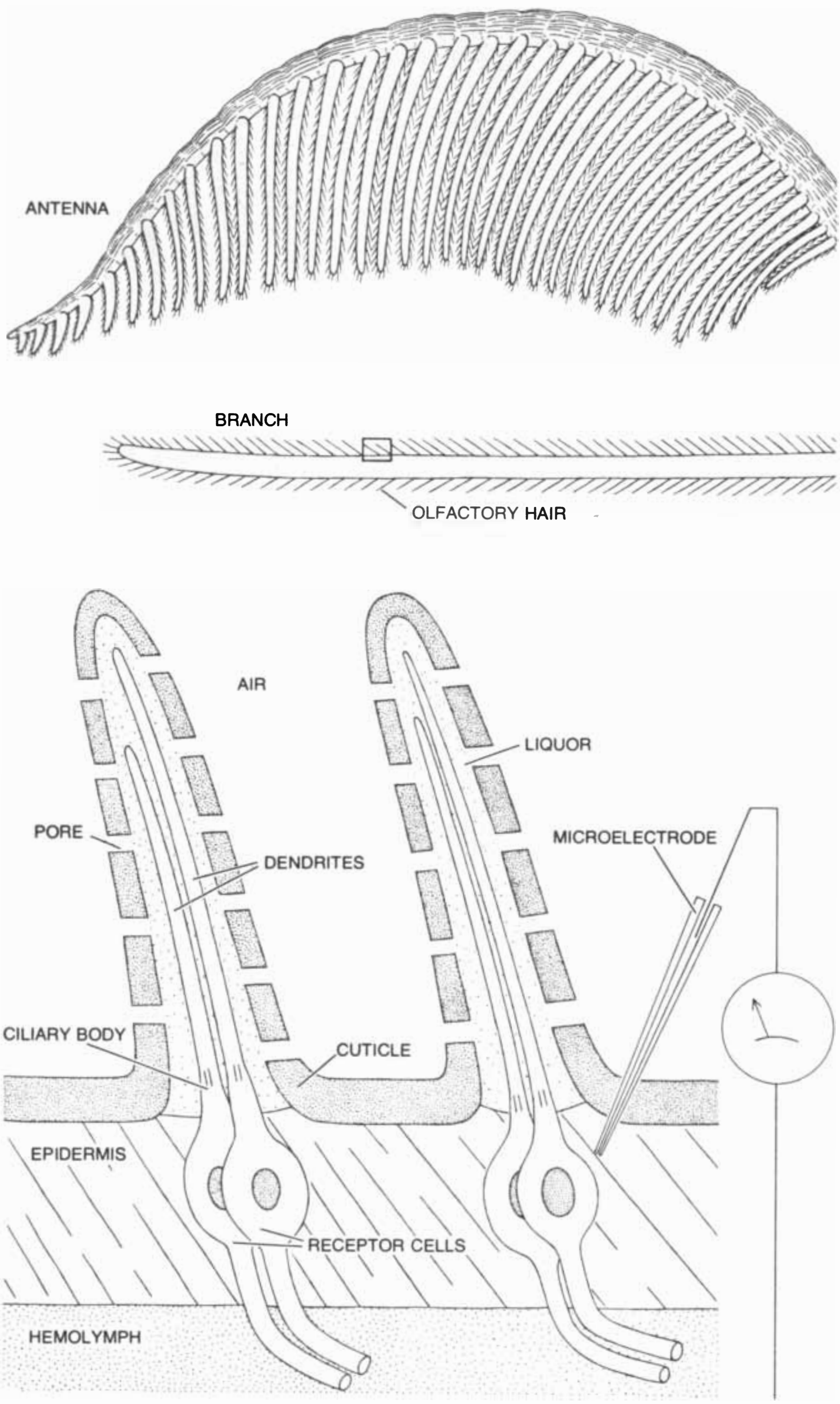
BOMBYKOL is an unsaturated fatty alcohol with the chemical designation *trans*-10-*cis*-12-hexadecadien-1-ol (*a*). It is the *trans-cis* isomer of the compound. The three other geometrical isomers, the *cis-trans* (*b*), the *cis-cis* (*c*) and the *trans-trans* (*d*) are much less effective.



ELECTROANTENNOGRAM RESPONSE of the male silk moth to different concentrations of bombykol and its isomers is shown. When the concentration of the odor is less than .01 microgram, there is little or no difference in the antennal response to each isomer. At higher concentrations bombykol (*TC*) gives rise to a much greater response than the *cis-trans* (*CT*), *cis-cis* (*CC*) and *trans-trans* (*TT*) isomers do. In all cases the responses to very low odor concentrations are not significantly different from antennal responses to pure air.

that it reacts to single quanta of odor. A comparably high sensitivity has been found in the rod cells of the vertebrate retina, which respond to a single quantum of light.

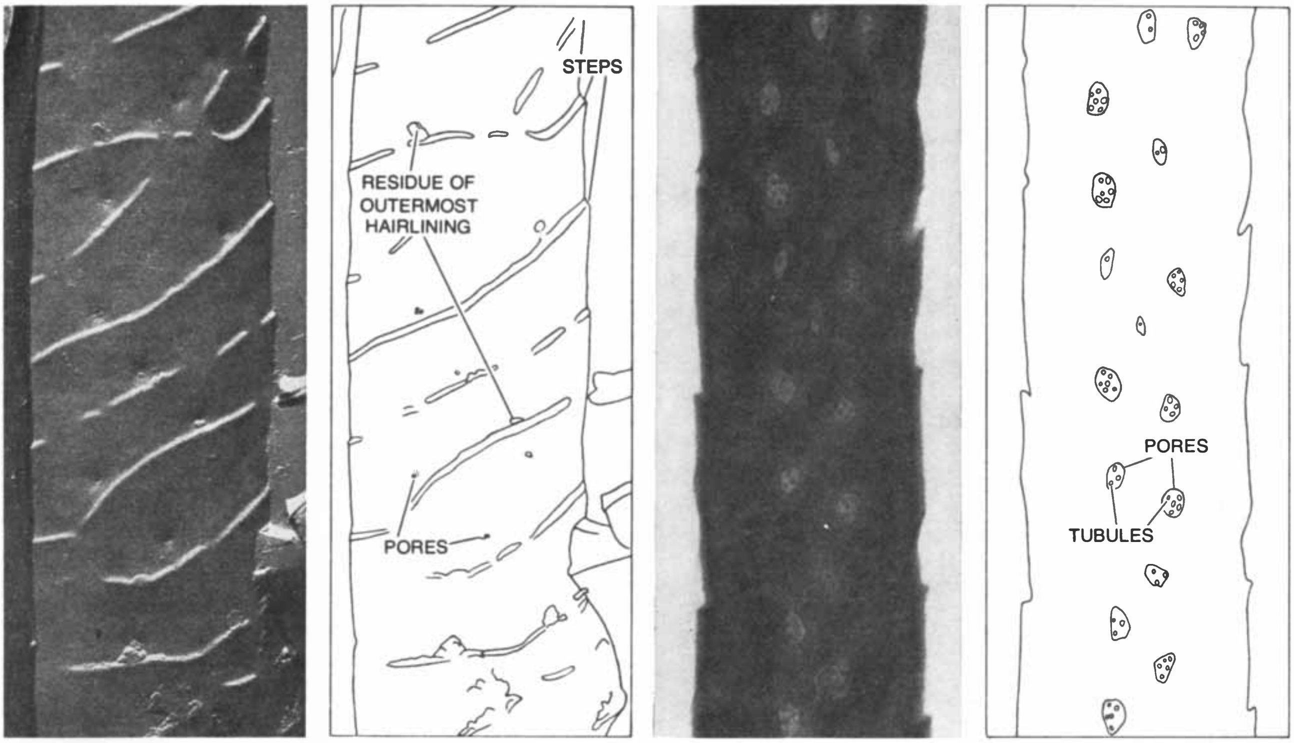
If the receptor cells are already responding to a single pheromone molecule, why must 200 cells be activated to make the male moth respond with its typical vibrations? The answer is that all the bombykol receptor cells of the antenna spontaneously fire about 1,600 impulses per second in the resting state, and the threshold signal of 200 impulses is necessary to overcome this background noise. Information theory requires that a detectable signal be greater than three times the square root of the noise. Since here the noise is about 1,600 impulses per second, a meaningful signal must be greater than $3 \times \sqrt{1,600}$, which is equal to 120. The threshold signal of 200 impulses is well above the required minimum, so that the signal is sufficient to tell the moth's brain: "There is bombykol in the air."



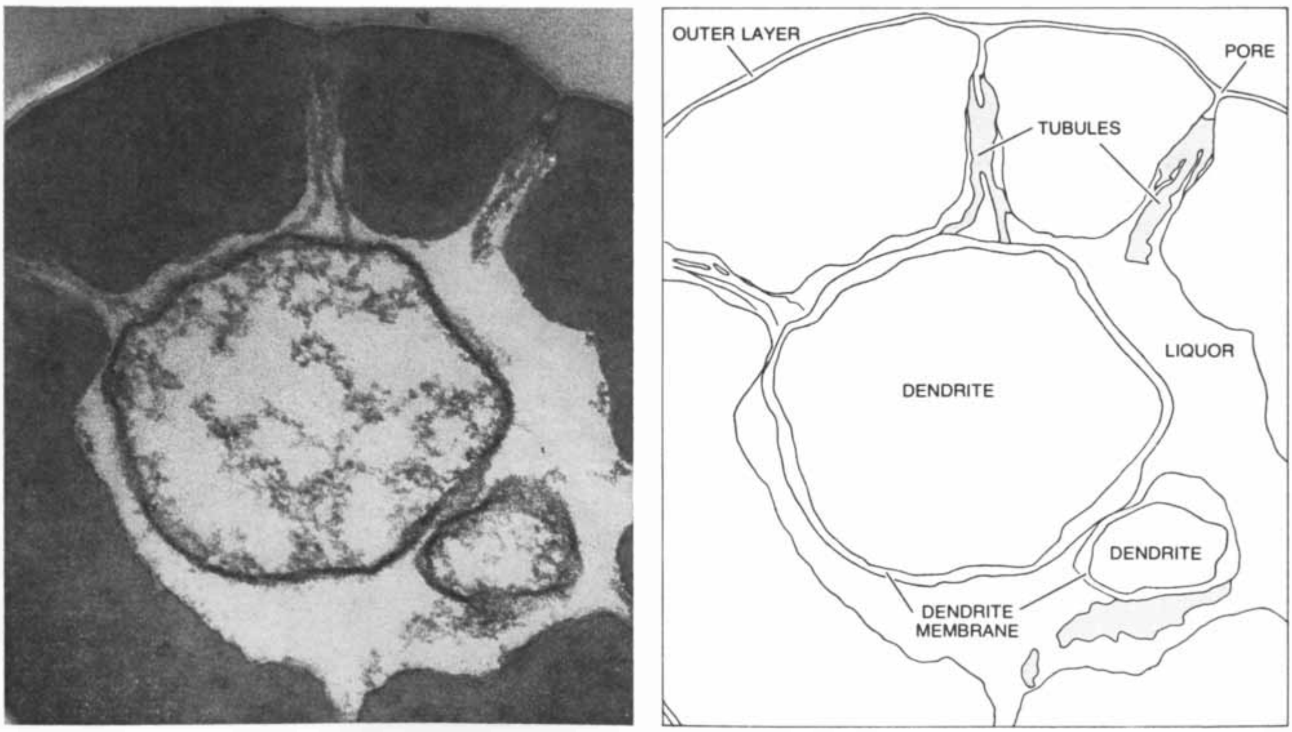
OLFACTORY RECEPTOR SYSTEM of the male silk moth is shown at increasing magnifications, first the entire antenna (*top*), then a single branch of the antenna (*middle*) and finally a schematic longitudinal section of two olfactory hairs (*bottom*). Each hair is innervated by dendrites of two receptor cells. Molecules of bombykol diffuse through the pore openings in the hair and give rise to an electrical change in the membrane of the receptor cell. Nerve impulses are recorded from a microelectrode inserted into the base of the hair.

The next questions we asked were: Why is the *Bombyx* antenna so effective in filtering the bombykol molecules out of the air? How do the molecules, after adsorption on the hair surface, find their way to the receptor-cell dendrite? Our thinking and experimentation on these lines have been strongly influenced and guided by Gerold Adam and Max Delbrück of the California Institute of Technology, who clearly outlined for us the physical principles that must govern these processes. Adam and Delbrück predicted that the antenna must be an optimal sieve for molecules because of the spacing and arrangement of the receptor hairs. The width of the mesh represented by the hairs is small enough so that the molecules of an odorant, because of their fast thermal movements, cannot pass through the hairs without coming in contact with them and being preferentially adsorbed. By measuring the adsorption of bombykol by the whole antenna and the adsorption by individual hairs that had been shaved from the antenna, we found that more than 80 percent of the molecules are adsorbed on the hairs. This result was significant because the total surface area of the hairs is less than 13 percent of the total surface area of the antenna.

If a receptor cell on a hair is to be activated by one or two molecules of bombykol, the molecules presumably need to diffuse to the receptor-cell dendrite from the surface of the hair. In order to demonstrate the validity of the assumption that there is such a two-dimensional diffusion of bombykol on the surface of the olfactory hair, it was necessary to analyze the structure of the hair. Such a hair, along with its olfactory receptor cells and auxiliary cells, is called a sensillum. The hair is part of the cuticle, which is the tough outer lining of the moth's body. The striking feature of the hairs is that they are perforated by pores connected to fine tubules. In some cases these tubules extend right down to the surface membrane of one of the receptor-cell dendrites. The pores and the tubules can be invaded by test



SURFACE OF RECEPTOR HAIR of a male silk moth is shown in replica after freeze etching in the electron micrograph at left. The tracing at right identifies the primary features that are visible. A tangential section of a receptor hair appears in the second micrograph. The pores are shown in cross section. Inside the pores tubules that extend to the dendrite of the bombykol receptor cell can be seen. The electron micrographs were made by R. A. Steinbrecht of the Max Planck Institute for Behavioral Physiology.



CROSS SECTION OF BOMBYKOL RECEPTOR HAIR is shown in this electron micrograph made by Steinbrecht. The cross section is through the apex of the hair. The tracing at right identifies certain of the hair's features. Some of the tubules in the pores reach to the outer membrane of the dendrite of the bombykol receptor cell. The outer pore openings are covered by several layers of a different electron density. Although the chemical composition of these layers is unknown, they are probably lipophilic, or fat-loving. Layers of such a composition would allow the fatty-alcohol molecules of bombykol to diffuse into the pore openings more readily.

substances from the outside, as K. D. Ernst has shown in our laboratory, but the content of the tubules is unknown [*see illustrations on preceding page*].

These observations enabled us to construct the following model: The odor molecules are adsorbed on the hair surface, diffuse to the pores and from there through the tubules to the receptor-cell dendrite, where they elicit the receptor potential. We calculated the diffusion time of bombykol on the hairs and found it to be well within the electrophysiologically determined response time. Although the adsorption of bombykol on the hair surface is an established fact and diffusion along the surface of the hair is highly probable, we have not yet been able to follow the pheromone into the pores and tubules. The spatial resolution of the currently available autoradiographic methods is not high enough to locate the bombykol molecule in these conduits.

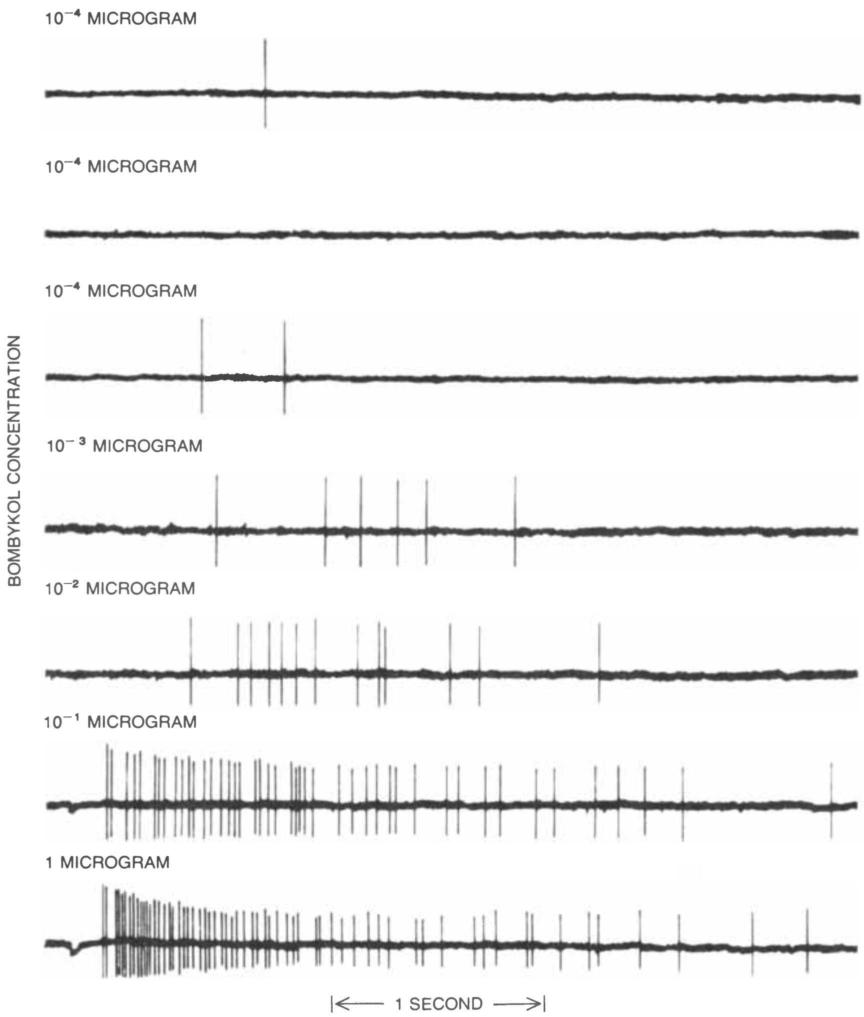
The process by which stimulus energy is transduced into receptor excitation is only partly understood for any sensory receptor cell, but with some reasonable assumptions we can at least outline it for the *Bombyx* odor-receptor cell. Our working hypothesis is that the bombykol molecule interacts with an acceptor in the membrane of the receptor-cell dendrite. The acceptor could function as a gating device, controlling the flow of ions through the membrane and thus the distribution of electric charges inside and outside the cell. The gating might be achieved by conformational changes in the molecular structure of the acceptor when it adsorbs the bombykol molecule.

Although the gating mechanism of the acceptor is definitely speculative, we can make some predictions about the properties of the active site of the acceptor. Fortunately we have quantitative information on a number of compounds that are effective in stimulating odor-receptor cells in other insects. When we looked into the physical properties of those molecules that activate a given type of cell, we were led to the conclusion that what happens in the process is not chemical bonding but weak physical interaction. We assume that the binding site and its reaction partner, the odor molecule, are complementary. On this basis the molecular specificity of the binding site can be deduced from the relative effectiveness of various stimulating molecules. Our observation that even the synthetic steric isomers of bombykol were from 100 to 1,000 times less effective than the natural pheromone indicates that the selective binding capacity of the bombykol acceptors is very high. The answer to the often-raised question "What makes a molecule an odor?" is on this functional level "The binding properties of the acceptor."

What is the fate of the odor molecule after it is bonded to the acceptor site in the receptor-cell membrane? With such a sensitive system it is unlikely that after the information transfer the odor molecule has a chance to activate the cell again. On the other hand, the odor molecule should not stay on the receptor site for too long, because the cell must be freed to be able to respond to another stimulus. We do not know as yet how the odor molecule is removed. One possible mechanism would be that it is immediately metabolized after the interaction. We have found that there is such a mechanism available, but it is neither specific enough nor fast enough to be directly involved in the transduction process.

Three years ago Morton Beroza of the U.S. Department of Agriculture asked us if we would be interested in extending our investigations to the gypsy moth, *Porthetria dispar*. His group had just successfully identified the lure pheromone of the female of this insect. It is *cis*-7,8-epoxy-2-methyloctadecane, dubbed "disparlure." As a result of the work of Beroza and his colleagues the compound was available in synthetic form. They had also succeeded in synthesizing 50 related epoxides, some with a different carbon-chain length, a shifted epoxy bridge (an oxygen attached to two carbons in the chain) and/or a shifted methyl group (CH_3). Starting a fruitful collaboration, we first repeated the measurement of the antennal response that had been made by our American colleagues and the experiments they had conducted with gypsy moths in the field. We also recorded the responses of single pheromone-receptor cells.

Our experiments clearly showed that disparlure is a more potent sex lure than



NERVE IMPULSES generated in a bombykol receptor cell increase in frequency as the concentration of odor increases. A concentration of 10^{-4} microgram of bombykol on odor source gives rise to one or two impulses or none. Recordings were made by E. Priesner.

any of the related compounds. The two compounds that elicited the next-largest responses are 20 and 100 times less effective. The male antennae of the gypsy moth and those of the silk moth are very much alike, so that there is little doubt that in principle the mechanisms by which the odor molecules are captured, are transferred and elicit electrical responses are identical in the two species. The details, however, have yet to be worked out.

The gypsy moth is a destructive forest pest in large parts of the eastern U.S. and in some parts of Europe (whence it was introduced to New England in 1869). The closest kin of this species is the nun moth, *Porthetria monacha,* which is found in the forests of central and northern Europe. We extended our studies to the nun moth, whose female sex attractant is not yet known. In our electrophysiological recordings we found that the nun moth had the same preference pattern for disparlure and related compounds as the gypsy moth has. The parallelism in their responses strongly suggests that the two species use the same compound as a sex attractant. Such a lack of species specificity in a sex attractant seems not to be an exception with species of the same genus but rather the rule.

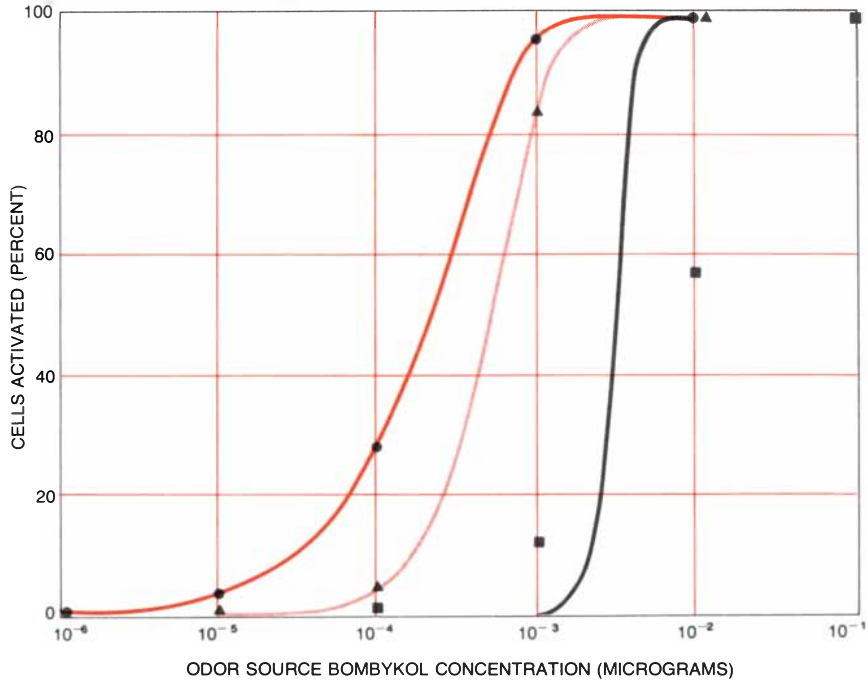
Field experiments conducted earlier by H. Schönherr of the University of Freiburg, in which disparlure was successfully used to trap nun moths, also suggest that disparlure is probably the attractant for this species. In collaboration with R. Lange and F. Schwarz of the same institution, we have continued these studies by comparing disparlure as a bait with some of its related compounds. Again disparlure was by far the most effective attractant.

For two species to have the same attracting pheromone would not present any problem if the species were widely separated ecologically. For the nun moth and the gypsy moth, which live in close proximity in some parts of Europe, it could cause confusion. Possible mechanisms that would prevent uneconomical cross-attraction or even hybridization are differences in the rhythm of daily activity of the species, differences in general and sexual behavior and even a morphological incompatibility for copulation.

When Butenandt started the chemical analysis of the silk moth's sex attractant in the 1930's, he was not just interested in the composition of this enigmatic compound. He was already thinking of the possibility that such substances could be synthesized and used to lure pest insects into traps with high specificity. He and others hoped that this stratagem could be used to avoid the dangerous side effects of the generalized chemical insecticides. Although the need for species-specific biological pest controls is now even more pressing than it was 40 years ago, the trapping of insects (or perhaps the confusion of insect sexual behavior) with pheromones does not appear to be the panacea. Nonetheless, in a number of cases the luring of pest insects with pheromone traps has proved to be useful in predicting the outbreak of a large infestation and thus in timing, calibrating and eventually reducing the application of generalized insecticides. The U.S. Department of Agriculture still hopes to be able to use the gypsy moth's sex attractant to prevent or even halt the insect's dangerous rate of expansion.

Successful pheromone research and promising field trapping has also been conducted with leaf-roller moths by Wendell L. Roelofs and his colleagues at the State Agricultural Experiment Station in Geneva, N.Y. They found that a mixture of pheromonal components will specifically attract several of these fruit tree pests. Another pheromone control method has been examined by H. H. Shorey and his colleagues at the University of California at Riverside. They evaporated a synthetic sex attractant of the cabbage looper in cabbage fields and the attractant of the pink bollworm moth in cotton fields. Under favorable conditions and with high doses of the attractant they produced an impressive degree of confusion among the males and a high percentage of unfertilized females.

In the future pheromones will probably become a component of a concerted system of biological control methods against pests that threaten crops and human health. Biologists the world over are searching for the Achilles' heel in the life of many pest species. My own view is that in many of these investigations a detailed analysis of the physiology, behavior and ecology of the insect involved is neglected. To be sure, behavior and ecology are complex and not easily studied expressions of the phenomenon called life. They nonetheless deserve the best effort of investigators not only for economic reasons such as the control of insect pests but also for a more fundamental reason: to better learn how we humans can survive in an ecologically balanced world together with our fellow organisms.



NUMBER OF CELLS that responded with one or more impulses (*circles*), two or more impulses (*triangles*) and 10 or more impulses (*squares*) were plotted against the concentration of bombykol. The data for the one-impulse response exactly fitted the Poisson probability curve for a random one-hit process (*darker color*). The data for the two-impulse responses fitted the two-hit curve (*lighter color*). The data for 10 or more impulses did not fit the 10-hit probability curve (*black*). These results, together with measurements of the adsorption of radioactively labeled bombykol on antennae, indicate that single impulses are generated by single molecules of bombykol and double impulses by two molecules.

HYBRID CELLS AND HUMAN GENES

The mapping of human genes and the study of how they are regulated is facilitated by an experimental substitute for sexual breeding: the fusion of human somatic cells with the cells of other mammals

by Frank H. Ruddle and Raju S. Kucherlapati

In order to recognize, understand and eventually treat human diseases that arise from the inheritance of faulty genes one must learn what genes are responsible for each disorder, where they are situated on the 24 different chromosomes in the human cell and how they function in the complex milieu of cell and organism. In the past decade the increased medical interest in human genetic analysis has been matched by the development of a major new experimental approach that makes such analysis feasible in the laboratory: the hybridization of somatic cells, which is to say the fusion of body cells (as opposed to male and female germ cells) from other mammals with those of humans in such a way that the resulting hybrid cells contain different assortments of only a few human chromosomes each. Somatic-cell hybridization experiments, facilitated by new chromosome-staining techniques and an enhanced ability to distinguish homologous gene products from different species, are beginning to fill in the human genetic map. Moreover, human genetic analysis is providing information on not only the location of genes on chromosomes but also the regulation of their expression in cells. It should therefore contribute to our understanding both of such medical problems as hereditary disease and cancer and of such fundamental biological problems as differentiation, development and evolution.

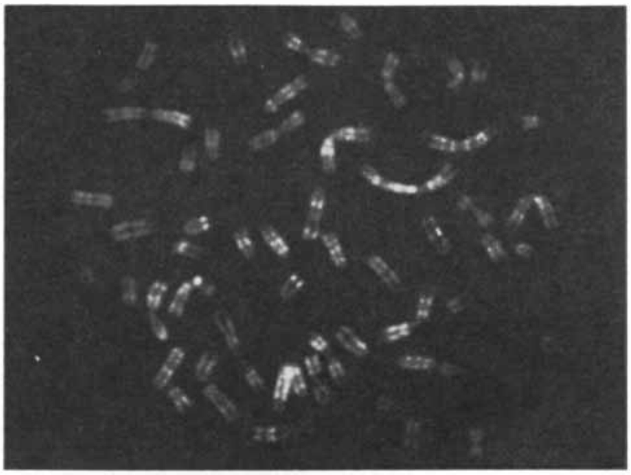
The classical method of mapping human genes has been to trace the pattern of inheritance of individual characteristics through many generations of a family and thus learn what traits are linked, or associated with one another on a single chromosome, and in some cases what that chromosome is. Such studies were extremely difficult because human families are relatively small, the generation time is long and it is impractical to carry out the controlled sexual breeding on which the classical genetic analysis of organisms higher than bacteria depends. Human geneticists developed statistical methods that eased some of these difficulties [see "The Mapping of Human Chromosomes," by Victor A. McKusick; SCIENTFIC AMERICAN, April, 1971], but the procedures are at best laborious, time-consuming and expensive. They are also limited in scope: the low frequency of rare traits in the total population makes many interesting disorders hard to study, and some genes cannot be investigated by familial analysis because they have no variant forms.

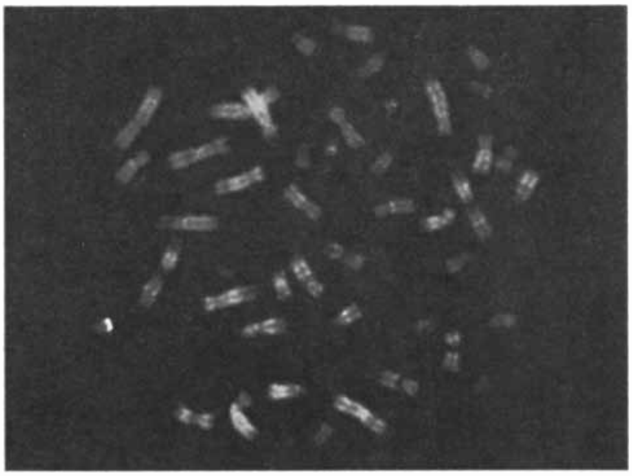
As long as 20 years ago geneticists such as Guido Pontecorvo, Curt Stern and Joshua Lederberg were suggesting that parasexual experimental systems should be developed for human genetic analysis. In 1960 George Barski and his colleagues in Paris discovered that two different mouse-cell lines would fuse in a laboratory culture, forming hybrid cells. The techniques of hybridization were improved by a number of workers, notably Boris Ephrussi, first in France and later at Case Western Reserve University, and Henry Harris of the University of Oxford. In 1967 Mary C. Weiss and Howard Green, then at the New York University School of Medicine, first demonstrated the potential of mouse-human hybrid cells for human genetic analysis [see "Hybrid Somatic Cells," by Boris Ephrussi and Mary C. Weiss; SCIENTIFIC AMERICAN, April, 1969].

In essence the hybridization procedure is simple. Mammalian cells can be treated in ways that adapt them to survival and multiplication in a suitable medium. Cells from two species—human fibroblasts (connective-tissue cells), say, and mouse cells adapted to tissue culture—are mixed in a laboratory dish and allowed to remain in contact for several hours. Spontaneous fusion between cells of the two different species occurs at a low frequency, but in practice fusion is usually enhanced by the addition of agents such as inactivated Sendai virus, which forms intercellular bridges between closely adjacent cells. When a suspension of highly dispersed parental and fused cells is plated on a culture medium, clones, or colonies, of cells arise, each stemming from the repeated division of an individual parental or fused cell and its progeny.

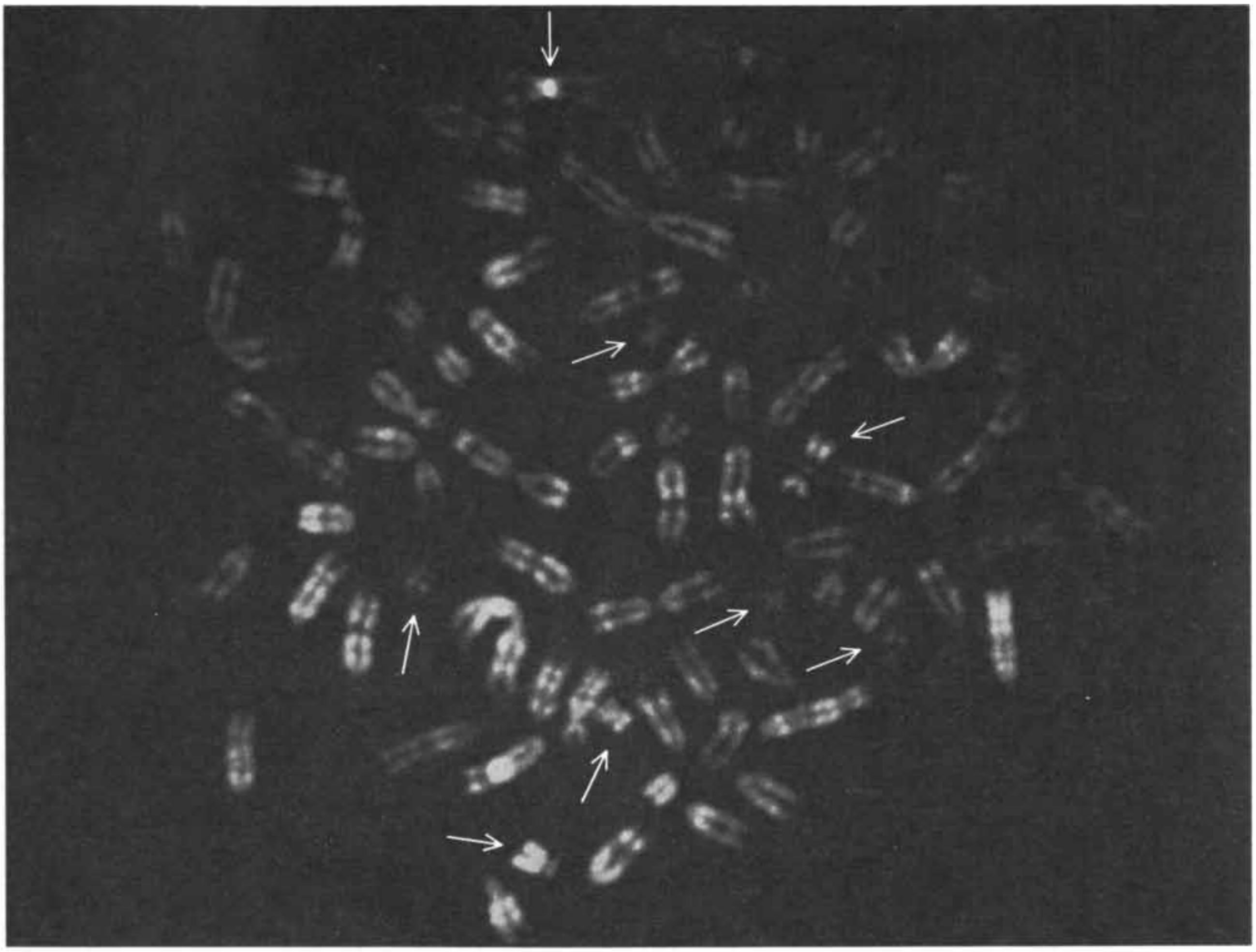
To give the hybrid-cell clones (and the investigator trying to isolate them) an advantage, the cells are usually grown in a "selective" medium. The one most commonly used is based on a selective strategy developed by Waclaw Szybalski, then at Rutgers University, and John W. Littlefield of the Harvard Medical School. It depends on the drug aminopterin, which blocks the production of folic acid, which is necessary for the normal synthesis of nucleotides, the building blocks of DNA. The nucleotides can be synthesized in the presence of aminopterin if their precursors, hypoxanthine and thymidine, are added to the medium, but only if the cells contain the "salvage pathway" enzymes hypoxanthine guanine phosphoribosyltransferase (HGPRT) and thymidine kinase (TK).

The usual practice is to fuse human fibroblasts that are capable of making both enzymes with rodent cells that are deficient in either HGPRT or TK, in a medium containing hypoxanthine, aminopterin and thymidine (HAT medium). The enzyme-deficient rodent parental cells die whereas the human parental

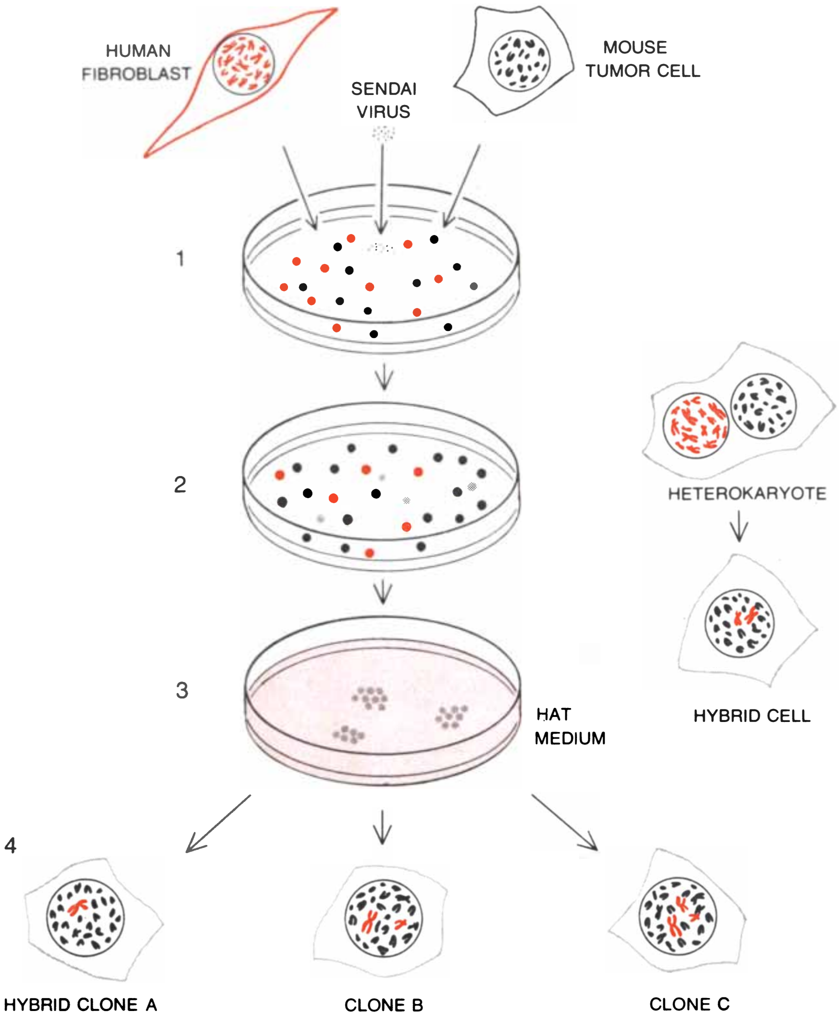




CHROMOSOMES of a mouse tumor cell (*left*) and a human fibroblast, a connective-tissue cell (*right*), were stained with the fluorescent dye quinacrine for these photomicrographs. The chromosomes are at metaphase, the stage in the cell cycle in which they are highly condensed and have replicated preparatory to cell division; the twin strands of each chromosome are joined to form *V*-shaped or *X*-shaped double chromosomes. There are 40 chromosomes in a normal mouse cell and varying larger numbers in tumor cells; there are 46 in a human cell. The fluorescent dye delineates characteristic pattern of bright and dark bands on each chromosome.



MOUSE AND HUMAN CHROMOSOMES are both present in a metaphase preparation of a hybrid cell that was formed by the fusion of two cells like those whose chromosomes are pictured at the top of the page. There are 73 chromosomes in this aberrant fused cell, only eight of them human chromosomes (*arrows*). This situation is typical of mouse-human hybrid cells, in which most of the human chromosomes are eliminated. Because different human chromosomes remain in different hybrid clones, or cell lines, it is possible to match them with the specific products of their genes and thus locate genes on chromosomes. This process is facilitated by stains, such as the quinacrine stain used here, that delineate the characteristic banding patterns. The bands tell an investigator, for example, that the arrowed chromosome at the top is human chromosome No. 3, distinguished by a particularly bright centromere.



HYBRID SOMATIC CELLS are formed by mixing human fibroblasts with mouse cells that are deficient in either of two enzymes, TK or HGPRT, and adding a fusion-enhancing agent such as Sendai virus (*1*). Some of the cells fuse (*2*), forming first heterokaryons with two nuclei and then new hybrid cells (*gray*). When cells are plated on HAT, a "selective" medium, fused cells proliferate and form colonies (*3*); the mouse parental cells cannot for lack of an enzyme and the fibroblasts do not multiply rapidly. The fused cells retain all the mouse chromosomes but only a few of 46 human chromosomes; human-chromosome loss is random, so that each hybrid clone has a different human chromosome complement (*4*).

		HYBRID CLONES				
		A	B	C	D	E
HUMAN ENZYMES	I	+	−	−	+	−
	II	−	+	−	+	−
	III	+	−	−	+	−
	IV	+	+	+	−	−
HUMAN CHROMOSOMES	1	−	+	−	+	−
	2	+	−	−	+	−
	3	−	−	−	+	+

LINKAGE AND LOCATION of genes is studied in hybrid cells by noting which gene products, usually enzymes, consistently appear together in a clone and then noting which chromosome is consistently present along with each product. In this table of hypothetical results of testing for enzymes and chromosomes, for example, enzymes I and III are seen to be linked on the same chromosome, that is, they are "syntenic." Moreover, they are assignable to chromosome No. 2. Enzyme II is seen to be assignable to chromosome No. 1.

cells and the fused cells survive. If the human cells are either fibroblasts that have been cultured repeatedly and therefore grow slowly or white blood cells, which do not normally proliferate, there will not be many human parental clones, and it is easy to isolate pure clonal populations of hybrid cells, each of which is derived from a single fusion event.

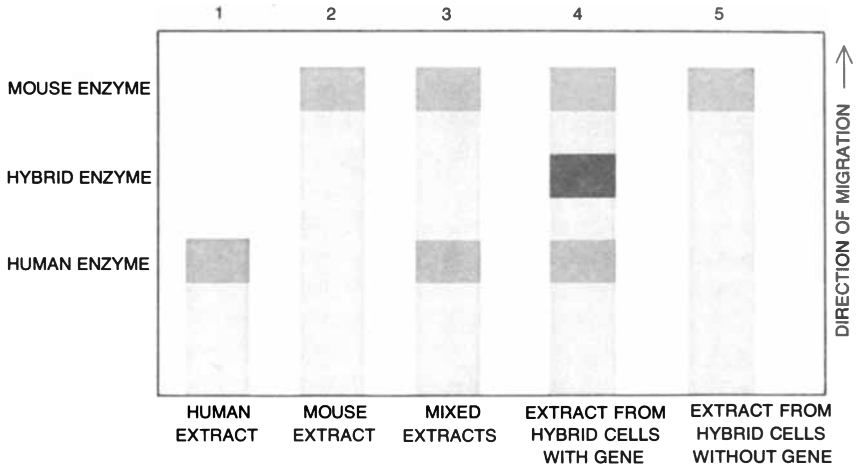
There are two important things about such hybrids. One is that the mouse genome, or genetic material, is retained essentially intact whereas the human genome is preferentially, partially and randomly lost. Normal human cells have 46 chromosomes: an *X* and a *Y* sex chromosome (in male cells) or two *X* chromosomes (in female cells) and 22 pairs of autosomes, or nonsex chromosomes. Normal mouse cells have 40 chromosomes. The hybrids always have fewer than the expected combined total of 86, with the usual range being from 41 to 55. All the mouse chromosomes are present; it is most of the human ones that are lost, and that loss is random: hybrids that are the result of different fusions retain different sets of human chromosomes. The other important thing about the hybrids is that, as Weiss and Green discovered, both mouse and human genomes are functional: both sets of genes are expressed at the same time, each coding for its appropriate proteins. The mouse chromosomes and the human chromosomes can be distinguished from one another, and the mouse and the human enzymes or other protein products can also be distinguished.

The chromosomes of the two species and some of the 24 different human chromosomes differ fairly obviously in shape, in size and in the location of the centromere: the point at which the two strands of a metaphase chromosome converge. The problem of distinguishing certain very similar human chromosomes from one another has been largely solved in the past few years by the development, primarily by Torbjörn Caspersson of the Royal Caroline Institute in Sweden, of a fluorescent staining method. When the chromosomes are stained with quinacrine, the dye binds and fluoresces in such a way that different chromosomes have different patterns of dark and bright bands. The banding patterns serve to identify the chromosomes and also provide landmarks along the chromosomal strand. As for the gene products, the evolutionary distance between mouse and man is large enough so that homologous proteins generally differ in their amino acid constitution, and these differences can be detected by electrophoresis [*see illustration at right*].

The task in mapping is first to learn whether two or more genes are on the same chromosome and then if possible to learn which chromosome they are on and just where on the chromosome individual genes are located. The human chromosomes are usually lost from hybrid cells as discrete units. Genes that are on the same chromosome will therefore usually be expressed together. In other words, two genes that are consistently either present together or lost together can be considered to be syntenic, or on the same chromosomal strand. Assaying a number of clones for various human enzymes therefore provides information on the synteny of genes. It requires only one more step to assign genes to chromosomes by essentially the same method: one simply compares the pattern of gene expression with the presence or absence of specific chromosomes [*see bottom illustration on opposite page*].

With these methods it has been possible in the past several years to assign more than 50 genes to 18 human chromosomes [*see illustration on next two pages*]. A number of the genes already mapped are implicated in human genetic diseases. For example, a deficiency in the enzyme hexosaminidase *A* is associated with Tay-Sachs disease; a deficiency in the enzyme galactose-1-phosphate uridyl transferase is associated with galactosemia. Increased knowledge of the map positions of such genes and their syntenic relations with other genes increases the ability to predict abnormal offspring by prenatal examination of the fetus. Progress in mapping has been so rapid that we can expect that in a few years at least one gene will have been assigned to each of the human chromosomes.

Although the basic method is logical and straightforward, various refinements have been developed in order to get results more quickly and with more precision. Theoretically if one had 24 rodent-human hybrid-cell lines, each retaining only one human chromosome, and if each hybrid retained a different chromosome, the gene for any testable enzyme could be assigned to the proper chromosome. In practice it is hard to get such one-chromosome hybrid lines, and so we have developed a scheme that should enable us to obtain similar information with as few as five hybrid lines. Each line has been selected for its possession of a unique subset of chromosomes. For example, consider three different lines, each of which has a different subset of the human chromosomes Nos. 1, 2, 3, 4, 5, 6, 7 and 8 [*see illustration below*]. Each of the three clones has a unique binary combination of chromosome presence or absence; to put it another way, the pattern of presence or absence in each clone is unique for each chromosome: chromosome No. 2 is present or absent in the panel of three clones in the sequence +, +, −; chromosome No. 6 has the pattern −, +, −.



HOMOLOGOUS ENZYMES are distinguished by electrophoresis. The diagram shows how the process works for glucose-6-phosphate dehydrogenase (G6PD). Cell extracts from each of five sources are placed in channels at base of a starch gel. G6PD is a dimer: a polymer with two subunits. Human cells make a human homopolymer and mouse cells a mouse homopolymer, each of which migrates at a different speed in an electric field, so that the human enzyme (*1*) does not move as far as the mouse enzyme (*2*). A mixture of extracts from both kinds of cell shows the two components (*3*). In a hybrid cell human and mouse subunits are synthesized and interact randomly, giving rise to the two homopolymers and to mouse-human heteropolymers that migrate at an intermediate speed. The enzyme from a hybrid cell that includes the gene for human G6PD therefore separates into three bands (*4*). Hybrids that do not retain the human G6PD gene produce only the homopolymer from the mouse (*5*).

The panel can be tested for any human phenotype (which is to say for any enzyme or other gene product) and will yield a specific pattern of expression that depends on the location of the responsible gene. That is, if testing the panel for the presence of human enzyme *A* yields the result −, −, +, one can say that the gene responsible for the expression of enzyme *A* must be on chromosome No. 7. Given five clones (with 2^5, or 32, combinatorial possibilities), one can cover the entire set of human chromosomes. In practice we work with from eight to 10 clones because increasing the level of redundancy helps us to avoid false interpretations based on errors that are encountered in testing for the presence or absence of an enzyme.

The methods we have described so far can pin a gene down to an individual chromosome but cannot go further; they cannot locate a gene at a precise point on a chromosome or establish the order in which several genes are arrayed along a chromosome. These things can often be done if one makes use of chromosomal aberrations such as translocations or de-

		HUMAN CHROMOSOMES							
		1	2	3	4	5	6	7	8
HYBRID CLONES	A	+	+	+	+	−	−	−	−
	B	+	+	−	−	+	+	−	−
	C	+	−	+	−	+	−	+	−

PANEL OF THREE HYBRID-CELL CLONES, each containing a unique subset of eight human chromosomes, can be tested for the presence or absence of any gene product. The pattern of presence or absence will match a pattern associated with one of the eight chromosomes. For example, if an enzyme is present only in clone *C*, its gene must be on chromosome No. 7. A larger panel of five clones, each with a unique subset of the 24 chromosomes, offers 32 combinatorial possibilities, enough to cover the 24 different human chromosomes.

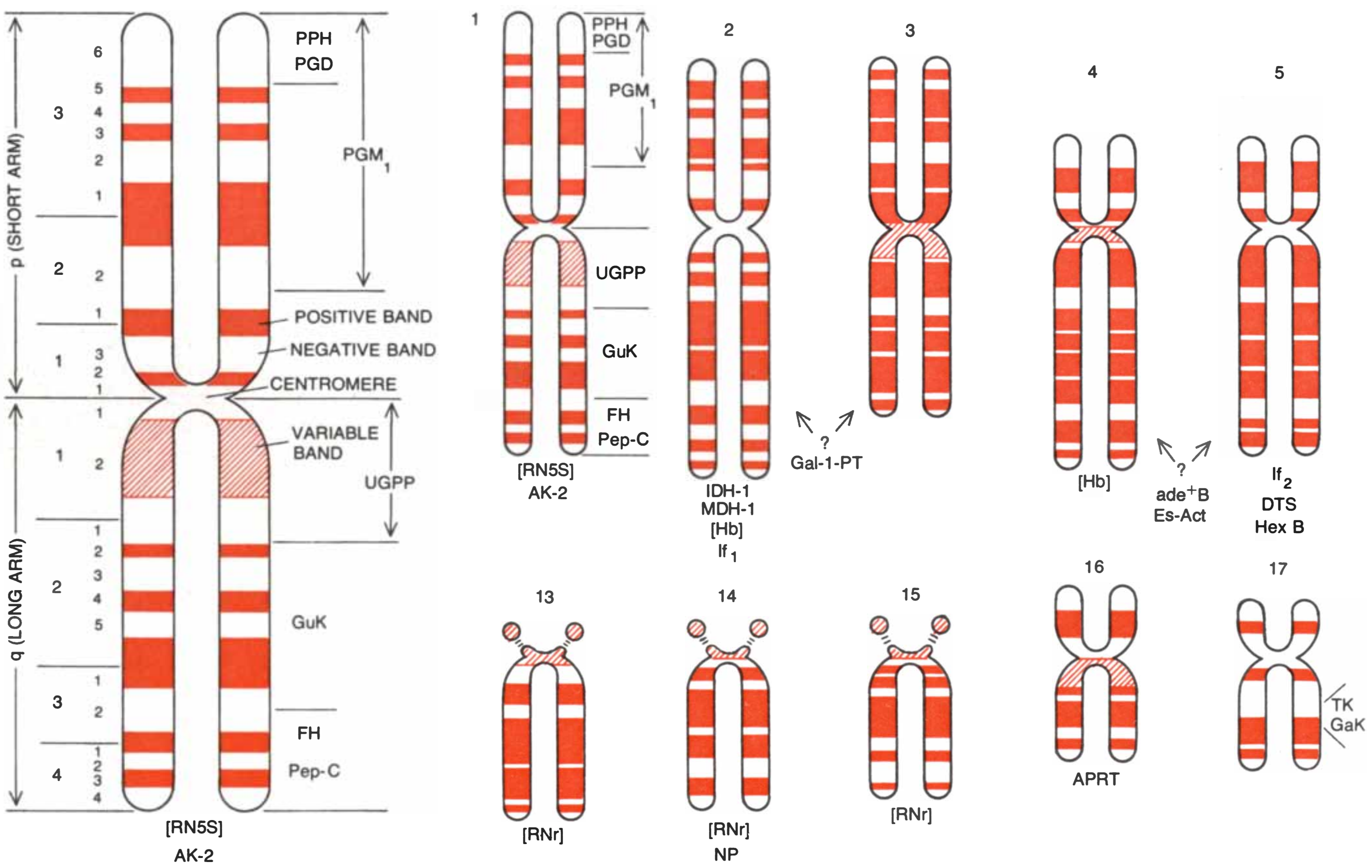
letions, which disturb the normal syntenic relations of genes.

Florence C. Ricciuti of our laboratory at Yale University capitalized on a translocation to localize three genes that were known to be on the human *X* chromosome: the genes for the enzymes HGPRT, phosphoglycerate kinase (PGK) and glucose-6-phosphate dehydrogenase (G6PD). The translocation, which appeared in one family, is such that most of the long arm of the *X* chromosome is translocated to an almost intact chromosome No. 14. The two translocation products are a short one bearing the centromere and the short arm of the *X* and a long one made up of most of the No. 14 with the long arm of the *X* attached to it; they are easily distinguished from each other and from the rest of the human chromosome complement. As for the location of the three genes, there were three possibilities: they could all be on the long arm of the *X*, they could all be on the short arm or they could be distributed on both arms.

Ricciuti hybridized human cells that carried the translocation with a mouse-cell line that was deficient in HGPRT. The hybrids were selected in HAT medium and then were analyzed for the presence of a number of human enzymes. The three human enzymes in question were present in all the hybrids. Since the ability of such hybrids to grow in HAT depended on their retention of human HGPRT, the results suggested that all three genes must be on a single translocation product, that is, all three were on the short arm of the *X* or all three were on the long arm. The enzyme screening showed, moreover, that the gene for nucleoside phosphorylase (NP), which was known to be on an autosome (a nonsex chromosome), was segregating along with the *X*-linked markers. That was strong evidence that the NP gene is on No. 14 and therefore also that the three *X*-linked enzymes are most probably on the translocated long arm of *X*. Detailed analysis of the chromosome pattern confirmed this by demonstrating a positive correlation between the long translocation product and the four enzymes: NP, PGK, HGPRT and G6PD.

Similar experiments were done by Park S. Gerald of the Children's Hospital Medical Center in Boston and by Dirk Bootsma at Erasmus University in Rotterdam. They worked with translocations that had different break points, and so their results not only confirmed Ricciuti's but also provided a unique sequence for the three genes on the *X* chromosome [*see illustration on page 42*]. At the present time several workers, including John Hamerton of the University of Manitoba and Bootsma, are doing similar analyses in an effort to localize genes on chromosome No. 1.

Another translocation has helped us, in collaboration with James K. McDougall of the University of Birmingham, to localize the gene for human thymidine kinase (TK) rather precisely. TK, the first gene mapped through somatic-cell hybridization, is on chromosome No. 17. Charlotte Boone and T. R. Chen of our laboratory examined several mouse-human cell lines in which the mouse cells had been deficient in TK. One of the hybrid lines was found to have a spontaneous translocation between a mouse chromosome and the long arm of human chromosome No. 17; the human TK was synthesized by that line, and so



BANDING PATTERNS AND KNOWN GENE LOCI are shown for the 24 different human chromosomes. Chromosome No. 1 is drawn large (*left*) to illustrate the standard nomenclature and the numbered designations of the various segments. The colored bands are the positive ones, those that are stained brightly by a fluorescent dye such as quinacrine; the white bands are the negative ones; hatched regions are variable bands that stain differently in the chromosomes of different individuals. Genes are designated by the initials of the enzymes or other products they code for: *PPH* is phosphopyruvate hydratase, for example; *RNr* is ribosomal RNA.

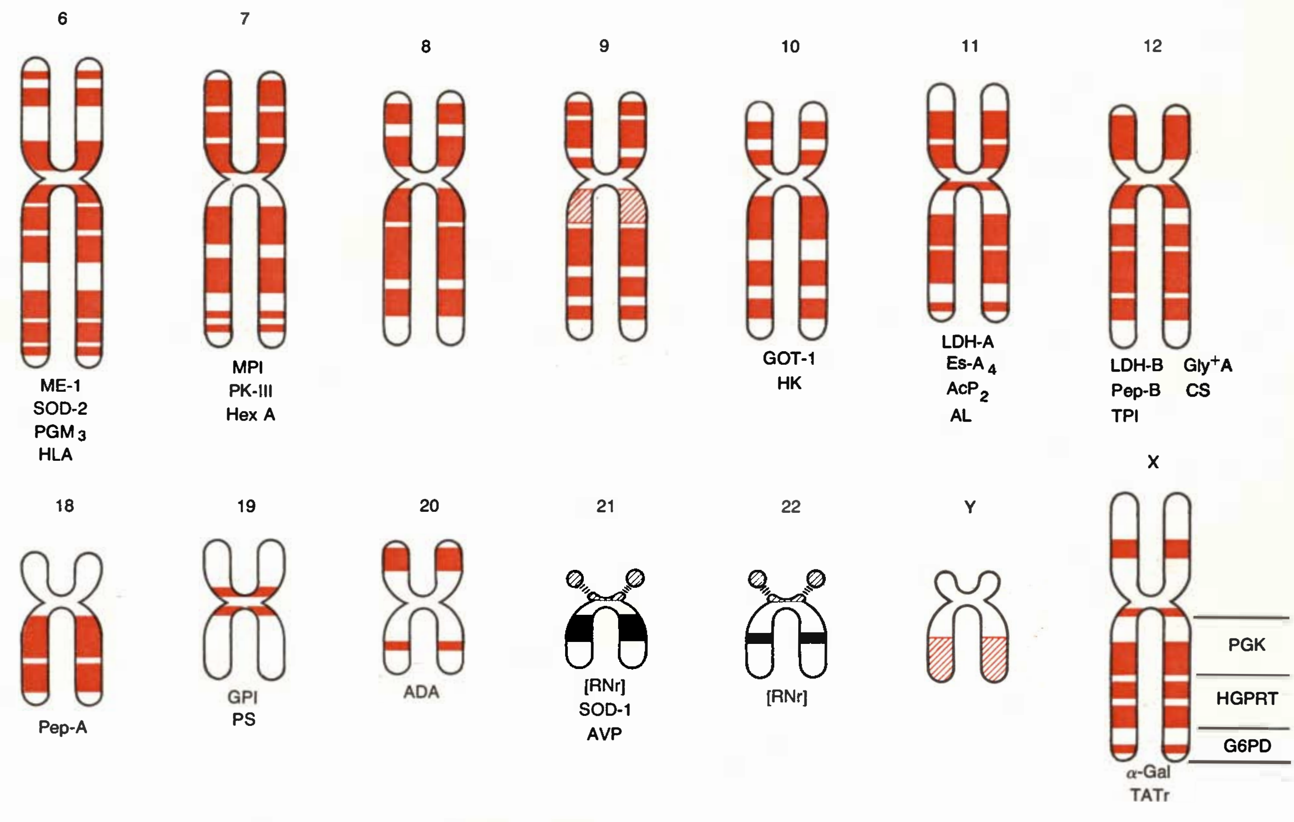
we could conclude that the TK gene is on the long arm of No. 17. (The translocation, incidentally, provided the first indication that chromosomes from different species could exchange segments.)

In order to localize the TK gene more precisely we capitalized on McDougall's discovery that human cells infected with adenovirus 12 exhibit a high frequency of breakage in chromosome No. 17. The breakage is preceded by the formation of an "uncoiler" region in the chromosome: the chromosome strand, which is ordinarily tightly coiled in these metaphase chromosomes, is uncoiled here, as it is during the stage of the cell cycle when the DNA is being transcribed into RNA and the RNA is being translated into protein. As a matter of fact, it was known that the level of host-specific TK in cells increases after adenovirus-12 infection; apparently extra TK is being synthesized for a time. We postulated therefore that the primary effect of the virus was to induce the uncoiling or to interfere with normal compaction of the chromosome and that the chromosome breaks were subsequent, secondary results of the infection. If that was the case, we expected the chromosomes might break at different points, so that some broken chromosomes would retain the TK gene and others would not. That would provide a unique opportunity for determining the exact location of the gene.

We chose hybrid cells in which the mouse-human translocation represented the only detectable human chromosome segment, treated the cells with adenovirus 12 in various concentrations, cultured the infected cells and then tested different colonies for their ability to grow in HAT medium, which would reflect their retention or loss of the human TK gene. We found that cells without the ability to make TK had lost a portion of the long arm of No. 17. There were two kinds of cells that retained TK activity: cells that had not lost any part of No. 17 and cells in which the deletion was somewhat smaller than the one that resulted in the loss of TK activity. Comparison of the breaking points for the two classes of cells gave us the location of the TK gene [*see illustration on page 43*]. The three cell lines, incidentally, constitute a small panel for the long arm of chromosome No. 17. Working with this panel, we could map an unrelated gene, for galactokinase, to a region close to TK. The concept can be extended to map genes in all other subregions of chromosomes.

The methods we have described so far involve the mapping of "constitutive" genes: genes that are expressed during the entire life of most or all types of cells. There are two other classes of genes: those that are expressed only when the cell is exposed to a specific external challenge and those that are expressed continuously but only in specialized cell types. Mapping such genes calls for somewhat different techniques, and it can also provide some particularly rewarding insights into mechanisms of gene regulation.

Mammalian cells protect themselves against attack by a virus by producing a protein, interferon, that somehow inhibits virus replication. Just what interferon does and how it works are not yet clear, although the matter has been under vigorous investigation because of interferon's obvious potential for treat-



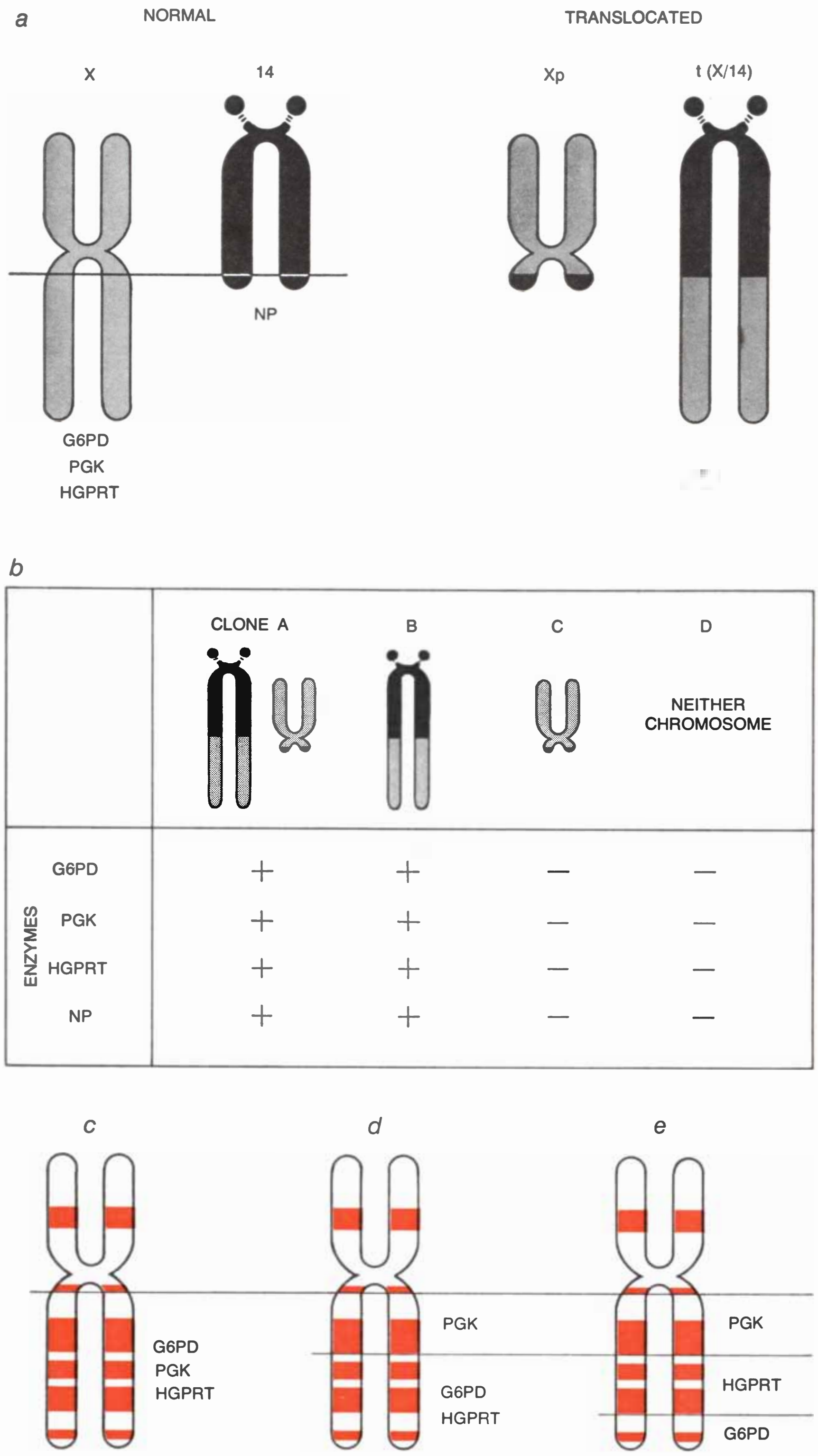
The genes are listed under the chromosome with which they have been associated or, in cases in which a location on the chromosome has been determined, are listed opposite that location. Genes that are listed with question marks may be on either of the chromosomes to which they are referred. Genes that are listed in brackets were mapped by annealing techniques involving the preparation of complementary nucleic acid; all the other genes on the map were located through the technique of somatic-cell hybridization. Still other genes have been mapped over the years through studies of genetic anomalies in families, but they are not included here.

ing virus diseases [see "The Induction of Interferon," by Maurice R. Hilleman and Alfred A. Tytell; SCIENTIFIC AMERICAN, July, 1971]. An important fact about interferon is that in many instances it is species-specific: mouse interferon is made by mouse cells and protects only mouse cells, not human ones; human cells make human interferon.

In our laboratory Y. H. Tan and Richard P. Creagan treated a number of different mouse-human hybrid cells with viruses and with chemical agents that mimic the virus in inducing cells to produce interferon. They found that two chromosomes, No. 2 and No. 5, had to be present together if interferon was to be induced. Even then, however, the antiviral response was not forthcoming. If chromosome No. 21 was present, on the other hand, and if interferon was then supplied, the cells did exhibit the antiviral response; cells with No. 21 alone, however, did not respond to the inducer by making interferon or by inhibiting virus replication.

Our interpretation is that two different genes, one on chromosome No. 2 and one on No. 5, are required for the expression of human interferon. We do not know which of the genes codes for interferon itself and which is required for some other, supportive function, and so studies to determine the nature and action of the two genes are in progress. These results constitute, among other things, the first reported case in which two genes appear to be required for the expression of a specific function in human cells.

As for chromosome No. 21, we believe it codes for still another protein, whose identity and function have not yet been determined, that may interfere directly with viral gene replication or may possibly code for a species-specific receptor site on the cell membrane that accepts interferon. The hybridization system is well suited for the genetic characterization of just such discrete membrane functions. Recently Dorothy Miller and her associates at Columbia University have shown that the poliomyelitis virus receptor site is coded by a gene on chromosome No. 19. In our laboratory Creagan and Susie Chen have shown that the diphtheria-toxin receptor site is coded by a gene on chromosome No. 5. It may soon be possible to identify gene products, coded by each of the human chromosomes, that are expressed at the cell surface. This may provide investigators with additional opportunities to regulate the human chromosomal composition of hybrid cells and will also contribute to our knowledge of the



TRANSLOCATIONS can be capitalized on to locate genes in specific regions of chromosomes. Genes for three enzymes, HGPRT, PGK and G6PD, were known to be on the human *X* chromosome; another gene, for the enzyme NP, was known to be on some autosome (nonsex chromosome). The translocation is such that most of the long arm of the *X* is exchanged with the tip of chromosome No. 14 (*a*). Cells bearing the translocation were fused with mouse cells deficient in HGPRT and hybrid cells were selected in HAT medium. Examination of clones yielded the results shown (*b*), indicating that all three of the *X*-chromosome genes must be somewhere on the translocated long arm (and that the NP gene must be on No. 14). Comparison of the *X*-chromosome results from the authors' laboratory at Yale University (*c*) with results of similar experiments, in which the translocations had different break points, in two other laboratories (*d*, *e*) mapped the sequence and the approximate loci of the three genes along the arm, as shown in illustration on the preceding two pages.

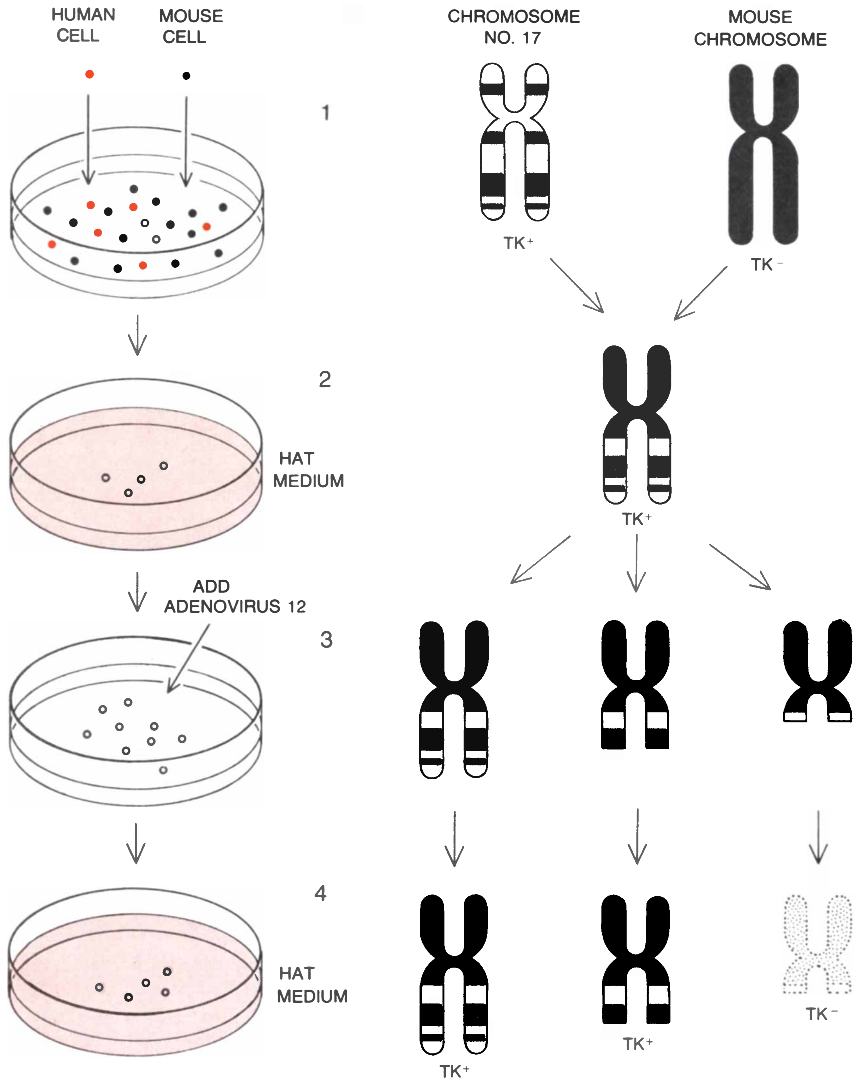
structure and function of the cell membrane.

We shall describe two examples of genes that are expressed in specialized cell types: the gene for the blood-plasma protein albumin, which is expressed in the liver, and the gene for globin (the protein of hemoglobin), which is ordinarily expressed in red blood cells and their immediate precursors. Characterizing such specialized functions as these and understanding the mechanisms whereby they are expressed to the exclusion of other functions in specialized cells should provide important information bearing on the regulation of gene activity and thus on the processes of differentiation and development.

Gretchen Darlington and Hans Peter Bernhard of our laboratory set out to study the regulation of albumin production by hybridizing liver cells with cells that do not express this phenotype, and to map the human albumin gene. Since normal liver cells do not survive well in culture, they worked with mouse hepatoma (liver cancer) cells, which are easily cultured and which synthesize and secrete albumin. They fused the hepatoma cells with human white blood cells, which do not make albumin, and examined the resulting hybrid cells for their ability to secrete albumin and for their human-enzyme and chromosome composition.

All the hybrid lines produced mouse albumin, and two of them also produced human albumin (which can be distinguished from the mouse version of the protein by immunological methods). What this meant was that white blood cells contain the hereditary information for the synthesis of albumin: the albumin gene is there but it is repressed, or inactivated, and it is somehow activated by factors present in the albumin-making mouse-liver cell. A similar result was attained by Mary Weiss and Jerry Petersen and by Stephen Malawista of the National Center for Scientific Research in France, who crossed rat hepatoma cells with mouse cells that do not normally make albumin and found that the hybrid cells produced some mouse albumin. These results provide additional evidence to support the assumption that all cells of an organism contain all its genes, that cells are genetically equivalent even if they have become differentiated to the point where they are functionally quite distinctive.

So far not enough hybrid-cell lines have produced human albumin for us to establish the syntenic relations of the



INDUCED CHROMOSOME BREAKAGE defined the location of the gene for the enzyme TK, which had previously been mapped to chromosome No. 17. The investigators fused mouse cells deficient in TK with human cells (*1*). A hybrid line in which there was a translocation of the long arm of a human No. 17 to a mouse chromosome was viable in HAT medium, showing that the TK gene was on the long arm of No. 17 (*2*). Adenovirus 12 was added to induce chromosome breakage at several different points, and cells with different translocation products were grown in nonselective mediums (*3*). When they were tested in HAT medium (*4*), it was found that the cells with translocation products retaining the whole long arm of No. 17 or most of the arm still had the TK gene; the cells that had lost more of the arm did not have the gene. Comparison of the three break points located the TK gene.

human albumin gene with any other human markers or to associate the gene with a chromosome; we hope that will be possible after study of a larger series of hybrid cells. Quite apart from the mapping, however, the techniques developed in the albumin experiments provide a new way to get at the complex and fascinating subject of gene repression and derepression, which is currently under intensive study. And they should also be applicable to the prenatal detection of hereditary diseases affecting such specialized tissues as the brain, the liver and the kidneys.

Prenatal detection is now confined to diseases involving gross chromosomal abnormalities or enzymes that happen to be produced and expressed in fibroblasts, the only cells that are easily recovered from the amniotic fluid. Perhaps fibroblasts (like the white blood cells in the albumin experiment) could be made to reveal the properties of their entire genetic complement (as the white blood cells were made to reveal their albumin gene) through fusion with established cell lines that do express specific genes that are medically significant. Then many more defects in the fetal genome could presumably be diagnosed.

Rather than activating a repressed

gene, hybridization can have the opposite result: it can extinguish a tissue-specific, differentiated function that was expressed by one of the parental cells. Arthur Skoultchi, now at the Albert Einstein College of Medicine in New York, developed a system in our laboratory that makes it possible to investigate the mechanisms of this extinction process, which are still not understood.

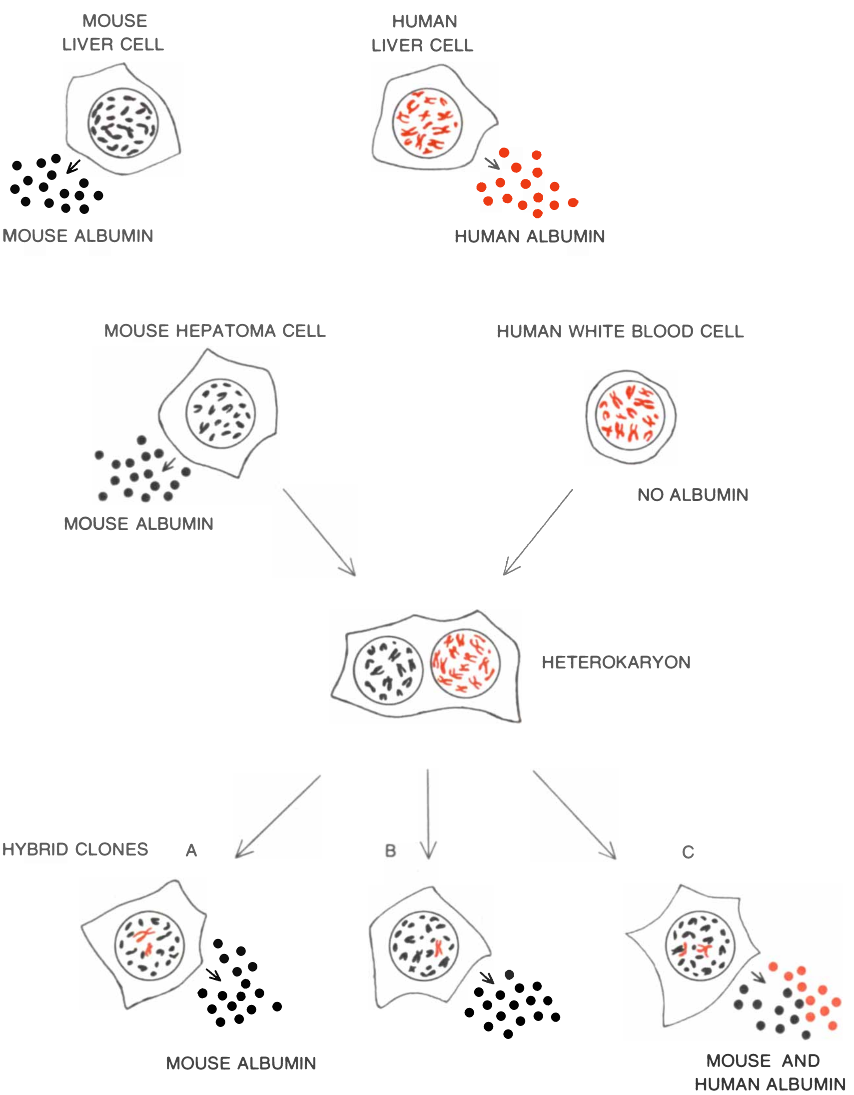
When certain strains of mice are treated with a complex of viruses known collectively as Friend leukemia virus, they develop a cancerous condition called erythroleukemia. Infected cells (precursors of red blood cells) from the spleen of such mice have the property of growing indefinitely in a culture. Charlotte Friend of the Mount Sinai School of Medicine in New York, who first isolated the cells, showed that they can also be made to produce hemoglobin if they are treated with the chemical dimethyl sulfoxide. Skoultchi found that if the erythroleukemic cells are fused with mouse fibroblasts, the hybrid cells cannot synthesize hemoglobin even after treatment with dimethyl sulfoxide.

One obvious explanation for their loss of this capability would be that they have lost the chromosome or chromosomes that carry the genes for globin, the hemoglobin protein. That explanation can be ruled out, however. The reasoning is as follows: the gene for the enzyme glucose phosphate isomerase (GPI) is known to be linked to the gene for one of the two kinds of globin; all the hybrid-cell lines made GPI in the two distinguishable forms characteristic of each type of parental cell but did not make hemoglobin; therefore the globin gene must be present even though its product is not. Such a result is consistent with gene modulation, that is, with such processes as repression and derepression.

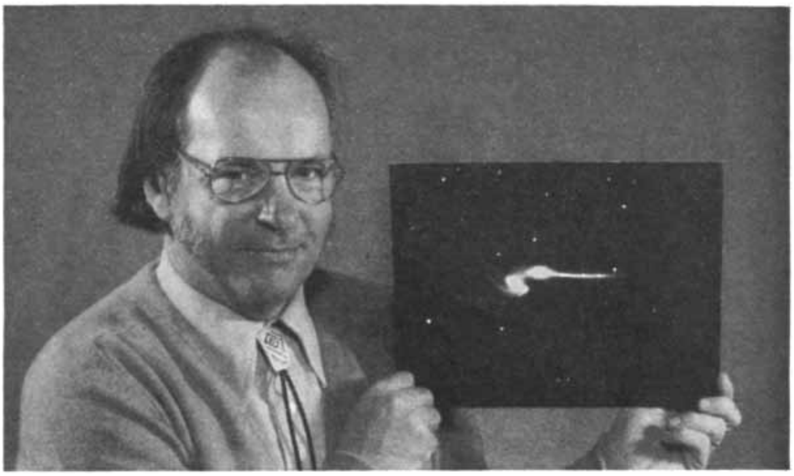
The globin system is particularly adaptable to the study of gene modulation because one can look into the cell and learn not only whether or not globin is being produced but also, if it is not, at what stage its synthesis was blocked. Globin messenger RNA, the nucleic acid that is transcribed from the DNA of the gene and is subsequently translated into the protein globin, is relatively easy to purify because it constitutes a large fraction of the RNA of any cell that makes it and because it can be separated from other RNA's by sedimentation. Once globin messenger RNA has been purified it is possible, with the help of the enzyme reverse transcriptase (also called RNA-directed DNA polymerase), to prepare DNA that is complementary to the RNA. This is DNA whose subunit sequence forms a template that in effect fits the sequence of the RNA subunits and that therefore tends, when exposed to the globin RNA, to become "annealed" to the RNA.

The complementary DNA can be used as a probe with which to measure the amount of globin RNA present in a cell. When the probing technique is applied to the hybrid cells that result from fusion of mouse fibroblasts and erythroleukemic cells, the data indicate that there is very little if any globin messenger RNA in their cytoplasm. The extinction of gene activity, at least in this instance, is apparently at the level of transcription from DNA to RNA, not at the level of translation from RNA into protein. (It is also possible that it is the transport, processing or stability of the messenger RNA that is affected.)

Somatic-cell hybridization has proved to be an efficient tool for the mapping of human genes. It has opened new doors to the understanding of human hereditary disorders and of the regulatory mechanisms that underlie both the malignant transformation of cells and the mutually interdependent functions of that most complex group of cells, the human organism.



ACTIVATION of a repressed gene coding for the protein albumin was demonstrated in hybrid cells. The albumin gene is ordinarily expressed in mouse liver or hepatoma (liver cancer) cells, as it is in human liver cells. White blood cells do not make albumin. Investigators fused mouse hepatoma cells with human white blood cells. The resulting hybrid cells all produced mouse albumin. Some of them also produced human albumin, which can be distinguished from the mouse protein by immunochemical methods. The result shows that the gene for human albumin is present in white blood cells but is not expressed. It is activated in those hybrid cells that contain its chromosome (which has not yet been surely identified), presumably by some regulatory factor that is supplied by the mouse liver cell.



Astronomer

(with The Mice galaxies)

Pride is of the highest importance as a business motivator. We are proud that A. G. Millikan, a Kodak emulsion researcher, has sufficient standing in his own right as an astronomer to be granted telescope time at a major observatory. We are pained that the admirable accomplishments of radio astronomy, x-ray astronomy, and infrared astronomy can sometimes leave a nonparticipating observer of the scientific scene thinking that the photographic emulsion has been dropped from the tool kit. Heavens, no! Heavenly endeavor without it would be tough indeed.

To broaden understanding of its present role in astronomy, Millikan has written an article "Image Detection at the Telescope" in American Scientist *for May-June, 1974. We would be proud to be asked for a reprint at Dept. 55W, Kodak, Rochester, N. Y. 14650.*

Astronomers don't need advertising to remind them that they need the photographic emulsion. They can understand that there are no huge stakes to be won in plates and films for their use. We persist for 1) Kodak tradition and pride, 2) a desire to do something useful for Science, which has done a lot for us, 3) awareness of how much such merchandise as the gentleman below promotes owes to sensitizing-dye technology built up years ago to solve purely astronomical problems.

If Kodak products for astronomy don't need advertising, perhaps certain Kodak films that offer the advantage of complete insensitivity throughout the electromagnetic spectrum do need advertising. They are manufactured in France with dyed cellulose nitrate in place of an emulsion.

In Kodak LR 115 film (type II) it is 10 μm thick; in CA 80-15, 100 μm. Protons below 100 keV leave damage tracks in it, as do α particles below 4 MeV and heavy ions. Caustic soda enlarges the tracks for visibility.

LR 115 is for individual particles, as in detecting radioactive material at very low levels. CA 80-15 is for gross imagery—of objects to be viewed solely by their energetic-particle emission amidst overpowering radiation, and for neutron radiography.

Since neutron beams are attenuated by hydrogen scattering but not by most heavy elements, they penetrate to visualize organic material. They leave no tracks, but we can supply CA 80-15 film coated with material they activate to shoot α particles, which do register on the film. Neutron radiography can resolve a few micrometers and show up 1% thickness differences, it is said.

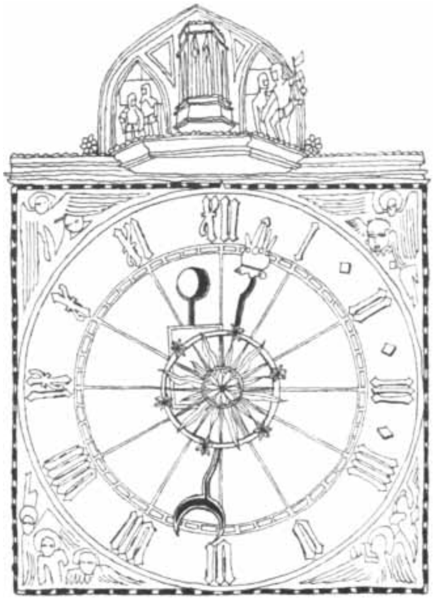
Inquiries on Kodak damage-track materials should be addressed to E. J. Hahn, Dept. 756, Kodak, Rochester, N.Y. 14650. Hahn doesn't push them. They are not the only way to register damage tracks. He merely ascertains whether they fulfill enough need to continue offering them.

Smile Saver®

Joe Ganey does indeed push the saving of smiles. Mr. Ganey is an enthusiast by profession, a strong advocate for the affirmative in the question To buy or Not to buy. He puts together a Kodak pocket Instamatic® 10 camera, a cartridge of Kodacolor II film, 3 magicubes with extender, a print frame, a little carrying case, an instruction booklet. He then directs his considerable energies to convincing as many members of the human family as possible that the joys of photography are worth ever so much more than the pittance a retailer asks for the Kodak pocket Smile Saver kit.

Without guys and gals like Ganey beating their brains out, astronomers and neutron radiographers might find their problems just a trifle more difficult than they already are.

An equal-opportunity employer f/m



SCIENCE AND THE CITIZEN

Then There Were Six

India became the sixth member of the nuclear club on May 18 by carrying out what it called a "peaceful nuclear explosion experiment using an implosion device." The explosion, which released the energy equivalent of between 10 and 15 kilotons of TNT, involved two novelties, one technical and one political. In detonating its nuclear device at a depth of "more than 100 meters" in the Rajasthan Desert, which borders Pakistan to the west of New Delhi, India became the first nation to avoid dirtying the atmosphere in its first nuclear test. The political novelty was that India declared that it had "no intention of producing nuclear weapons and reiterated its strong opposition to military uses of nuclear devices." India's Defense Minister Jagjivan Ram said his country was interested solely in employing nuclear explosives "for peaceful uses, for mining, for oil and gas prospecting, for finding underground water and for river diversion." The Indian government announced it had spent only $173 million on atomic energy in the past five years and had budgeted an expenditure of $315 million for the next five.

India's claim to be interested solely in the civil-engineering uses of nuclear explosives did not seem credible to U.S. observers who were familiar with Project Plowshare, the U.S. Atomic Energy Commission's long, costly and so far unsuccessful effort to find peaceful uses for nuclear explosives. "The Indians have put us in the Plowshare box," says one American who has spent nearly 30 years working for international renunciation of nuclear weapons. "After all of the AEC's ballyhoo about the potential peacetime uses of nuclear explosives we are in a poor position to say now that we do not believe India's declarations."

The strongest denunciation of India's nuclear test came from Canada, which helped India to build a large research reactor and two power reactors, all three of which use natural uranium as a fuel and heavy water as a moderator. The two power reactors, in addition to two U.S.-built power reactors that require enriched uranium, are subject to inspection by the Vienna-based International Atomic Energy Agency; hence they presumably were not the source of the plutonium used in the test explosion. The Canadian-built research reactor, however, which has a thermal rating of 40 megawatts, does not fall under the IAEA inspection system. The logical conclusion is that it produced the plutonium needed for the test. A 40-megawatt reactor is capable of yielding about seven kilograms per year of plutonium, assuming that the reactor is typically in service about half of the time. The amount of plutonium needed for a 10-to-15 kiloton explosion is five or six kilograms. The Canadian-built research reactor went into operation in the late 1950's. India has its own facilities for refining uranium, for fabricating natural-uranium fuel rods and for producing heavy water. It has also had in operation for about 10 years a plant for extracting plutonium from uranium fuel rods. H. N. Sethna, chairman of India's Atomic Energy Commission, told newsmen the day of the explosion: "It was a 100 percent Indian effort and the plutonium required for the explosion was produced in India."

The Indian test explosion came at a time when many nuclear geopoliticians had persuaded themselves that the proliferation of nuclear weapons they had widely predicted in the early 1960's was not going to take place after all, and that the size of the nuclear club might remain stabilized indefinitely at five (the U.S., the U.S.S.R., the United Kingdom, France and China). They reasoned that nuclear weapons were of little value without the concomitant development of delivery systems so costly that they could not be seriously contemplated by any but the largest powers. This view seemed to be sustained by the 97 signatures to the nonproliferation treaty put forward by the superpowers in 1968.

India did not, however, sign that treaty. "I would suggest," says one observer, "that the Indian bomb was made in Washington and Moscow. It may force the two powers to stop toying with arms-control schemes that are merely cosmetic and finally come to grips with the real problem of disarmament."

The U.S. and the U.S.S.R. have been engaged in what Alva Myrdal, disarmament negotiator for Sweden, has called "vertical proliferation." Since the adoption of the treaty banning atmospheric nuclear tests in 1963, the U.S. has conducted 255 underground tests at a cost of $3.5 billion and the U.S.S.R. has set off at least 90 underground blasts. The U.S. has steadfastly refused to accept a total ban on underground testing without a comprehensive inspection system, to which the U.S.S.R. has steadfastly refused to agree. Recent reports indicate that Secretary of State Henry Kissinger is currently trying to get the U.S.S.R. to accept an uninspected "threshold" ban that would limit underground tests to devices with a yield of less than 50 or 100 kilotons. The "higher politics" of this position, according to *The New York Times*, is that Kissinger could not support a total test ban lest he seem to be conspiring with the Russians in putting pressure on China to halt its weapons tests.

Slicing the Pie

In perpetuation of the economic relationship that once tied the peoples of Asia, Africa and Latin America to the colonial powers, the underdeveloped countries have over the years continued to ship mineral and agricultural raw materials to the developed countries in exchange for manufactured goods. The terms of this trade have increasingly been to the disadvantage of the underdeveloped countries throughout the period during which they were gaining their political independence. World prices for their commodities have consistently fallen behind the prices of the manufactured goods they have had to import. Now that pattern has begun to change. The oil embargo imposed last winter by the Organization of Petroleum Exporting Countries (OPEC) not only precipitated the "energy crisis" in the industrial economies; it also brought about a radical upward revision in world oil prices.

Evidence that the underdeveloped nations are determined to see the new pattern evolve was provided by the recent special session of the UN General Assembly on raw materials and development. The session was the sixth special session of the Assembly and the first to deal primarily with economic issues. (The others have all taken up political issues.) It was called on the motion of Algeria and was in effect a rally of the poor nations against the rich. The session produced two resolutions that reflect the temper of the underdeveloped nations.

One of the resolutions bears the title "Declaration on the Establishment of a New International Economic Order." It sets forth the view that "the developing world has become a powerful factor that makes its influence felt in all fields of international activity." Among the principles laid down for the new economic order are "full and effective participation on the basis of equality of all countries in the solving of world economic problems...; full permanent sovereignty of every State over its natural resources and all economic activities...; regulation and supervision of the activities of transnational corporations... in the interest of the national economies of the countries where such transnational corporations operate...; extension of active assistance to developing countries by the whole international community...."

The second resolution is a program of action aimed at bringing the new international economic order into being. Among other things, it calls for "arrangements to promote an increasing net transfer of real resources from the developed to the developing countries" and for "measures to encourage the industrialization of the developing countries."

Heavy Oil

The U.S. has underground deposits of "heavy" oil totaling more than 150 million barrels, according to the Bureau of Mines. If this virtually untapped energy source could be exploited, it could double the nation's proved petroleum reserves. Heavy oil is crude oil with a higher density and much greater viscosity than ordinary crude; technically speaking, all oils having an American Petroleum Institute "gravity number" of less than 25 are termed "heavy."

Although heavy oil can be refined by much the same methods as ordinary oil, its great viscosity precludes its being pumped through pipes. As a result the U.S. deposits have, with a few exceptions, been considered uneconomic. With the current high price of oil, however, and with the development of special recovery methods, a number of oil companies, both large and small, are moving into the heavy-oil business.

Most of the new methods are based on the injection of steam to heat the oil and thereby lower its viscosity so that it can be pumped. Some methods involve the addition of solvents or other chemicals to the steam. Collectively these methods are referred to as "tertiary recovery," since they are often used in ordinary oil fields to increase the percentage of recovered oil after waterflooding, a technique of secondary recovery, has been employed.

The development of low-cost recovery methods could have a significant impact on the world energy situation and could help to make the oil supply to the U.S. more secure. In addition to the U.S. reserves, which are mostly in California, Utah, Wyoming, Kansas and Arkansas, it is estimated that there are some 700 billion barrels of heavy oil in Canada and an even larger amount in the Orinoco tar belt of Venezuela.

Hyperactivity and Drugs

At times almost all children are unduly fidgety, noisy, disruptive or generally difficult to handle, but clearly some are more so than others; hyperkinesis, or hyperactivity, is recognized as a distinct syndrome. For some years it has been considered a specific "disease," and it has been treated by the administration of amphetamines and other stimulant drugs. The effect of the drugs is often dramatic, with the children quieting down, paying attention to their schoolwork and in some cases doing better in school. Although many psychiatrists have felt that the stimulants are not getting at the causes of a hyperactive child's disturbance, drug therapy has seemed at least to interrupt the characteristic record of failure in school and to make it possible to treat a child by other methods (see "Hyperactive Children," by Mark A. Stewart; SCIENTIFIC AMERICAN, April, 1970).

The practice and concept of drug therapy for hyperkinesis has been strongly criticized by Lester Grinspoon of the Harvard Medical School and Susan B. Singer, then of the Massachusetts Mental Health Center, in an article in the *Harvard Educational Review.* They base their article on an intensive review of the literature on hyperactivity and its treatment. They conclude that there is no justification for "the supposition that [the behavioral syndrome] is the result of a specific disorder of the central nervous system," and they question whether behavior that may stem from many different causes can be treated rationally by one type of drug. It has been assumed, they point out, that the apparent calming effect of stimulants on hyperactive children is a paradoxical effect, indicating that such children are physiologically different from other people; that assumption in turn reinforces the idea that the drug therapy is appropriate for them. In fact, the effect may not be paradoxical at all. Many studies have shown that increased attention and better performance are normal results of treatment with stimulants. The children may be reacting just like an adult who finds he can work better if he takes Benzedrine or Ritalin or some other stimulant.

Grinspoon and Singer estimate that some 200,000 children are now given stimulants routinely, and they consider it "impossible to believe" that all of them suffer from organic brain damage or chemical deficiencies. They suggest that the syndrome should be treated as a social problem, not a physical disease.

The Big Cannonade

The samples of lunar rock and soil brought back by the Apollo astronauts presented investigators with two contradictory kinds of evidence. On the one hand, radioactive dating techniques based on the ratios of isotopes of rubidium and strontium seemed to show that there were no lunar rocks that had crystallized more than four billion years ago. On the other hand, the same dating techniques simultaneously provided ample evidence that the moon, or at least its crust, formed some 4.5 billion years ago. The implication seemed to be that a series of bombardments, culminating in some kind of cataclysm about 3.9 billion years ago, obliterated most traces of earlier events in the lunar rocks. Recent analyses of the abundance ratios of isotopes of uranium, thorium and lead in rock samples from the lunar highlands independently confirm this implication.

According to Fouad Tera, D. A. Papanastassiou and G. J. Wasserburg, writing in *Earth and Planetary Science Letters,* the uranium-lead and thorium-lead data indicate that the lunar crust formed about 4.4 billion years ago, which falls short of the age of 4.6 billion years that is generally assigned to the planets. If the lunar crust also began forming 4.6 billion years ago, then perhaps it evolved in a multistage process that lasted .1 to .2 billion years.

The cataclysm of 3.9 billion years ago is associated with the events that created the vast lunar "sea" Mare Imbrium. It also probably gave rise to Mare Crisium, Mare Orientale and several other major basins, all within the relatively short span of 200 million years or less. The events could have been the outcome of either of two separate processes or both. The first possibility is that they resulted from the production of second-generation magmatic and metamorphic rocks from materials formed 4.4 billion years ago by sources of heat within the moon. The second possibility is that the events resulted from a massive series of impacts on the lunar surface and the subsequent shock-melting of the crust. For Tera, Papanastassiou and Wasserburg the second possibility is the "most appealing explanation." They conclude that the lunar-highland soils are not a good average sample of the lunar crust; they may have been enriched in certain elements because they were volatilized during the cataclysmic events.

The analyses raise three major questions. First, what was the source of the massive debris that bombarded the moon at such a relatively late date and then abruptly disappeared? Second, is the current orientation of the moon with respect to the earth the result of the major collision or collisions of 3.9 billion years ago? Third, was the last major bombardment localized in the earth-moon system or did it affect other bodies in the solar system as well?

Cables of the Cell

When a cell divides, its two sets of chromosomes are separated by the system known as the mitotic apparatus. This system consists of bodies called centrosomes at each pole of the dividing cell and a network of microtubules: fibers that connect the poles with the chromosomes. Although there are several competing theories of how the mitotic apparatus pulls the chromosomes apart, it appears that the microtubules probably provide the motive force.

A method for studying the action of the microtubules in isolation from the cell has been devised by Lionel I. Rebhun, Joel Rosenbaum, Paul Lefebvre and George Smith. Writing in *Nature*, they report that the mitotic apparatus can be made to perform at least some of its functions in vitro. Moreover, when the apparatus is deliberately disorganized, it can be restored by incubation in a suitable chemical medium.

Rebhun, Rosenbaum and their co-workers isolated the mitotic apparatus from the eggs of the surf clam (*Spisula solidissima*). The isolated apparatus retained several characteristics of mitotic apparatus in vivo, including two properties that were exploited in a subsequent series of experiments: birefringence and sensitivity to low temperatures. Birefringence is the refraction of a light ray into two beams; it results from the structure and dimensions of the microtubules. The sensitivity of the microtubules to low temperatures is demonstrated by the loss or reduction of birefringence when the temperature is lowered. At about zero degrees Celsius birefringence disappears because the microtubules disintegrate.

Microtubules are made up mainly of protein; they are polymers whose basic unit is a protein called tubulin. Rebhun, Rosenbaum and their co-workers reduced or abolished the birefringence of microtubules by exposing them to low temperatures, then incubated them at room temperature with tubulin. The birefringence returned, indicating that the microtubules had reassembled.

Some of the reassembled microtubules were observed to grow in length, one of the processes involved in mitosis. In some cases the reassembled apparatus assumed a configuration much like that observed in the last phase of cell division. This suggests, Rebhun, Rosenbaum and their co-workers conclude, that it may be possible to assemble from component parts a functional mitotic apparatus, that is, one that will separate chromosomes outside the cell.

Borrowed Countermeasure

Many insects regurgitate or defecate when they are threatened, a phenomenon that is familiar to almost anyone who has held a grasshopper in his hand. It is believed that such discharges help to protect the insect against predation. When the larva of the sawfly (*Neodiprion sertifer*) is disturbed, it quickly turns and dabs a droplet of a viscous resin from its mouth on the offending object. This oral effluent is obnoxious to many predators of the larva; ants and spiders, for example, promptly flee when the larva places a dab of resin on them.

The sawfly larva feeds on the needles of conifers such as the Scotch pine. It normally consumes not only the shaft of the needle but also the lower portion, which contains resin from the branch. The resin repels other insects, so that it is normally a part of the tree's own defenses against insect attack.

The composition and origin of the oral effluent of the sawfly larva have been intensively studied by Thomas Eisner, Judith S. Johnessee, James Carrel, Lawrence B. Hendry and Jerrold Meinwald of Cornell University, who report their findings in *Science*. Chemical analysis of the effluent revealed that it contained two volatile terpenes (alpha pinene and beta pinene) and seven viscous resin acids (which may serve to prevent the evaporative loss of the pinenes). All these components were found in resin from the branches of the Scotch pine, but only three were present in the resin from the needles of the tree. Further investigation showed that very young larvae, which nibble only the shafts of the needles, have an oral effluent that closely resembles the resin.

Dissection of larvae revealed that the resin is stored in two pouches of the foregut. Analysis of the midgut and of fecal pellets revealed no trace of pinenes or resin acids. The resin apparently never serves as a nutrient: the resin pouches of starved larvae remain full.

Even after the sawfly larva has spun its cocoon it responds to a scraping of the cocoon by turning until its mouth is at the site of the disturbance. If the cocoon is pierced, the larva regurgitates on the intruding object. When the skin of the larva is shed by the pupa, the foregut and its resin pouches are also left behind. Pouches recovered from cocoons as much as three years old were full of the resin, which was still clear and unhardened. Even the volatile pinenes were detectable.

Eisner and his colleagues conclude that the sawfly has achieved the capability of "crashing through" the chemical defenses of the pine without detoxifying its weaponry. The larva, they write, "simply appropriates intact the resin of the plant, thereby obviating the need for metabolic production of an alternative defensive agent of its own. One wonders to what extent the defensive enteric discharges of animals generally derive their effectiveness from unaltered dietary components. In fact, the whole question of transmission along food chains of non-nutritive substances... is worthy of increased consideration."

Unified Theories of Elementary-Particle Interaction

Physicists now invoke four distinct kinds of interaction, or force, to describe physical phenomena. According to a new theory, two, and perhaps three, of the forces are seen to have an underlying identity

by Steven Weinberg

One of man's enduring hopes has been to find a few simple general laws that would explain why nature, with all its seeming complexity and variety, is the way it is. At the present moment the closest we can come to a unified view of nature is a description in terms of elementary particles and their mutual interactions. All ordinary matter is composed of just those elementary particles that happen to possess both mass and (relative) stability: the electron, the proton and the neutron. To these must be added the particles of zero mass: the photon, or quantum of electromagnetic radiation, the neutrino, which plays an essential role in certain kinds of radioactivity, and the graviton, or quantum of gravitational radiation. (The graviton interacts too weakly with matter for it to have been observed yet, but there is no serious reason to doubt its existence.) A few additional short-lived particles can be found in cosmic rays, and with particle accelerators we can create a vast number of even shorter-lived species [*see top illustration on page 52*].

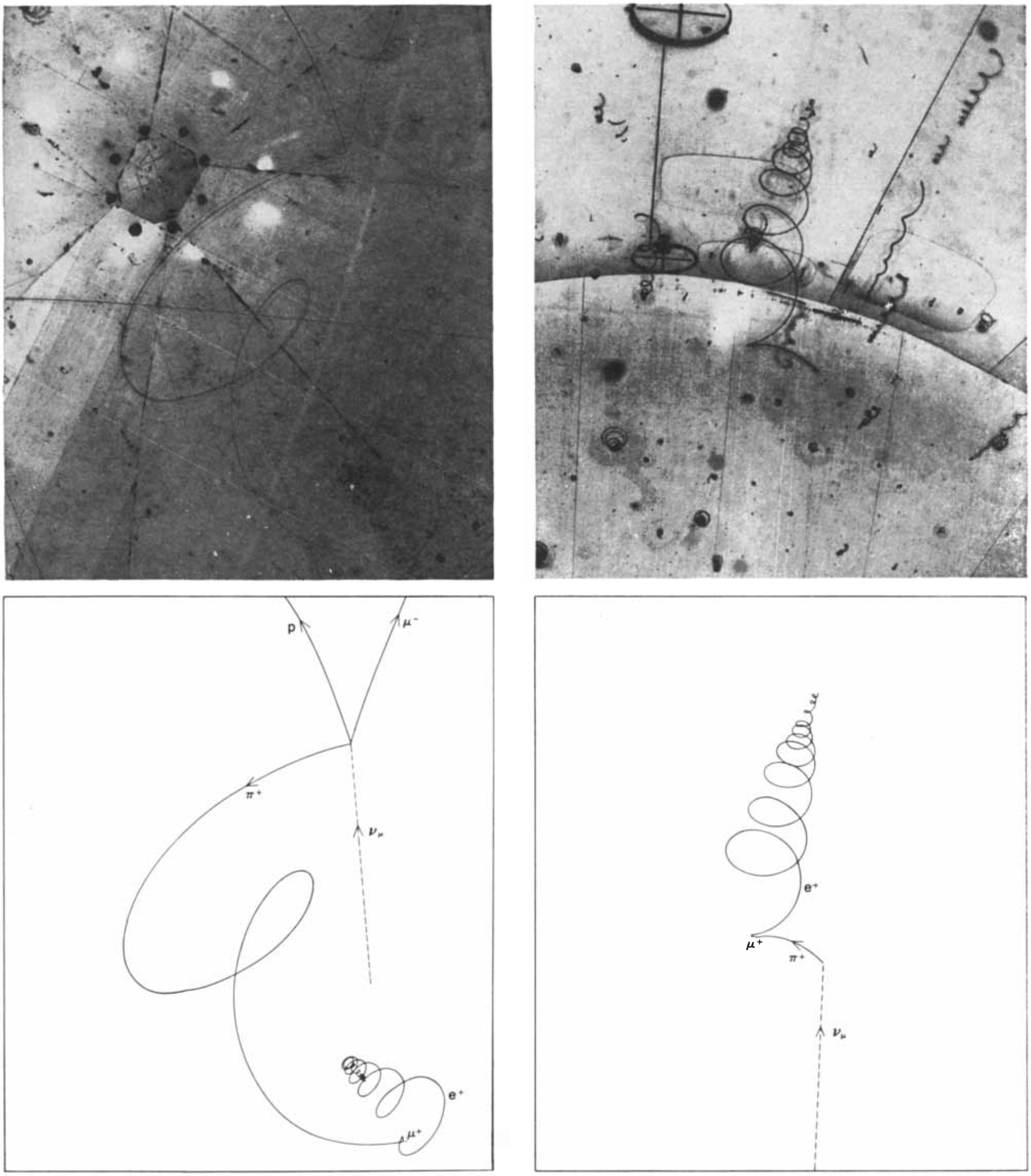
Although the various particles differ widely in mass, charge, lifetime and in other ways, they all share two attributes that qualify them as being "elementary." First, as far as we know, any two particles of the same species are, except for their position and state of motion, absolutely identical, whether they occupy the same atom or lie at opposite ends of the universe. Second, there is not now any successful theory that explains the elementary particles in terms of more elementary constituents, in the sense that the atomic nucleus is understood to be composed of protons and neutrons and the atom is understood to be composed of a nucleus and electrons. It is true that the elementary particles behave in some respects as if they were composed of still more elementary constituents, named quarks, but in spite of strenuous efforts it has been impossible to break particles into quarks.

For all the bewildering variety of the elementary particles their interactions with one another appear to be confined to four broad categories [*see bottom illustration on page 52*]. The most familiar are gravitation and electromagnetism, which, because of their long range, are experienced in the everyday world. Gravity holds our feet on the ground and the planets in their orbits. Electromagnetic interactions of electrons and atomic nuclei are responsible for all the familiar chemical and physical properties of ordinary solids, liquids and gases. Next, both in range and familiarity, are the "strong" interactions, which hold protons and neutrons together in the atomic nucleus. The strong forces are limited in range to about 10^{-13} centimeter and so are quite insignificant in ordinary life, or even on the scale (10^{-8} centimeter) of the atom. Least familiar are the "weak" interactions. They are of such short range (less than 10^{-15} centimeter) and are so weak that they do not seem to play a role in holding anything together. Rather, they are manifested only in certain kinds of collisions or decay processes that, for whatever reason, cannot be mediated by the strong, electromagnetic or gravitational interactions. The weak interactions are not, however, irrelevant to human affairs. They provide the first step in the chain of thermonuclear reactions in the sun, a step in which two protons fuse to form a deuterium nucleus, a positron and a neutrino.

From this brief outline one can see that a certain measure of unification has been achieved in making sense of the world. We are still faced, however, with the enormous problem of accounting for the baffling variety of elementary-particle types and interactions. Our prospects for further progress would be truly discouraging were it not for the guidance we receive from two great products of 20th-century physics: the development of quantum field theory and the recognition of the fundamental role of symmetry principles.

The Necessity of Fields

Quantum field theory was born in the late 1920's through the union of special relativity and quantum mechanics. It is easy to see how relativity leads naturally to the field concept. If I suddenly give one particle a push, this cannot produce any instantaneous change in the forces (gravitational, electromagnetic, strong or weak) acting on a neighboring particle because according to relativity no signal can travel faster than the finite speed of light. In order to maintain the conservation of energy and momentum at every instant, we say that the pushed particle produces a field, which carries energy and momentum through surrounding space and eventually hands some of it over to the neighboring particle. When quantum mechanics is applied to the field, we find that the energy and momentum must come in discrete chunks, or quanta, which we identify with the elementary particles. Thus relativity and quantum mechanics lead us naturally to a mathematical formalism, quantum field theory, in which elementary-particle interactions are explained by the exchange of elementary particles themselves.



EVIDENCE FOR NEUTRAL CURRENTS, the existence of which would support theories showing a connection between electromagnetic interactions and weak interactions, was recently obtained in an experiment conducted at the Argonne National Laboratory with a neutrino beam from the zero-gradient synchrotron and with a 12-foot bubble chamber filled with liquid hydrogen. The bubble-chamber photograph at left and the map below it show an example of a familiar kind of charged-current process ($\nu_\mu + p \rightarrow \mu^- + p + \pi^+$) in which a unit of electric charge is exchanged between leptons (ν_μ, μ^-) and other particles. The photograph at the right and the map below it show an example of a neutral-current process ($\nu_\mu + p \rightarrow \nu_\mu + n + \pi^+$) distinguished by the absence of outgoing negative muon (μ^-) or proton (p) tracks. In such photographs tracks are left only by charged particles, so that the incoming neutrino (ν_μ) and the outgoing neutrino and neutron (n) in the neutral-current process are invisible. Moreover, the bubble chamber is subjected to an intense magnetic field, which causes charged particles to follow curved tracks, clockwise for negative charge and counterclockwise for positive charge. In both of these photographs the positive pion (π^+) is seen to decay into a positive muon (μ^+), which then decays into a positive electron (e^+), visible as a tightly wound spiral. This experiment is more recent than similar ones that have been conducted at European Organization for Nuclear Research (CERN) and at National Accelerator Laboratory. It provides the first evidence for the specific neutral-current reactions $\nu_\mu + p \rightarrow \nu_\mu + n + \pi^+$ and $\nu_\mu + p \rightarrow \nu_\mu + p + \pi^0$.

		PARTICLE	SYMBOL	CHARGE	MASS (10^6 ELECTRON VOLTS)	LIFETIME (SECONDS)
		PHOTON	γ	0	0	∞
LEPTONS		NEUTRINO	$\nu_e \bar{\nu}_e$	0	0	∞
LEPTONS		NEUTRINO	$\nu_\mu \bar{\nu}_\mu$	0	0	∞
LEPTONS		ELECTRON	$e^{\pm}$	$\pm e$	0.511	∞
LEPTONS		MUON	$\mu^{\pm}$	$\pm e$	105.66	2.199×10^{-6}
HADRONS	MESONS	PION	$\pi^{\pm}$	$\pm e$	139.57	2.602×10^{-8}
HADRONS	MESONS	PION	π^{0}	0	134.97	0.84×10^{-16}
HADRONS	MESONS	KAON	$K^{\pm}$	$\pm e$	493.71	1.237×10^{-8}
HADRONS	MESONS	KAON	K^{0}	0	497.71	0.882×10^{-10}
HADRONS	MESONS	ETA	η	0	548.8	2.50×10^{-17}
HADRONS	BARYONS	PROTON	$p \bar{p}$	$\pm e$	938.259	∞
HADRONS	BARYONS	NEUTRON	$n \bar{n}$	0	939.553	918
HADRONS	BARYONS	LAMBDA HYPERON	$\Lambda \bar{\Lambda}$	0	1,115.59	2.521×10^{-10}
HADRONS	BARYONS	SIGMA HYPERON	$\Sigma^+ \bar{\Sigma}^+$	$\pm e$	1,189.42	8.00×10^{-11}
HADRONS	BARYONS	SIGMA HYPERON	$\Sigma^0 \bar{\Sigma}^0$	0	1,192.48	$< 10^{-14}$
HADRONS	BARYONS	SIGMA HYPERON	$\Sigma^- \bar{\Sigma}^-$	$\pm e$	1,197.34	1.484×10^{-10}
HADRONS	BARYONS	CASCADE HYPERON	$\Xi^0 \bar{\Xi}^0$	0	1,314.7	2.98×10^{-10}
HADRONS	BARYONS	CASCADE HYPERON	$\Xi^- \bar{\Xi}^-$	$\pm e$	1,321.3	1.672×10^{-10}
HADRONS	BARYONS	OMEGA HYPERON	$\Omega^- \bar{\Omega}^-$	$\pm e$	1,672	1.3×10^{-10}

PARTIAL LIST OF OBSERVED ELEMENTARY PARTICLES identifies all those with lifetimes greater than 10^{-20} second. Apart from the photon, all the observed particles fall into two broad families: leptons and hadrons. The leptons are either massless or low-mass particles that do not take part in the "strong" interactions; the hadrons are heavier and do take part. Hadrons are further divided into mesons and baryons according to rotational angular momentum and other properties. A symbol with a bar above it denotes an antiparticle. Neutrinos and antineutrinos are of two types: electron-type, ν_e, and muon-type, ν_μ. In three cases (photon, neutral pion and eta meson) the particle is its own antiparticle. Charges are given in units of charge, e, on the electron, equal to 1.602×10^{-19} coulomb. Masses are in energy units; one million electron volts (MeV) equals 1.783×10^{-27} gram.

	GRAVITATIONAL	ELECTRO-MAGNETIC	STRONG	WEAK
RANGE	∞	∞	$10^{-13}-10^{-14}$ CM.	$<<10^{-14}$ CM.
EXAMPLES	ASTRONOMICAL FORCES	ATOMIC FORCES	NUCLEAR FORCES	NUCLEAR BETA DECAY
STRENGTH (NATURAL UNITS)	$G_{NEWTON} = 5.9 \times 10^{-39}$	$e^2 = \frac{1}{137}$	$g^2 \approx 1$	$G_{FERMI} = 1.02 \times 10^{-5}$
PARTICLES ACTED UPON	EVERYTHING	CHARGED PARTICLES	HADRONS	HADRONS LEPTONS
PARTICLES EXCHANGED	GRAVITONS	PHOTONS	HADRONS	?

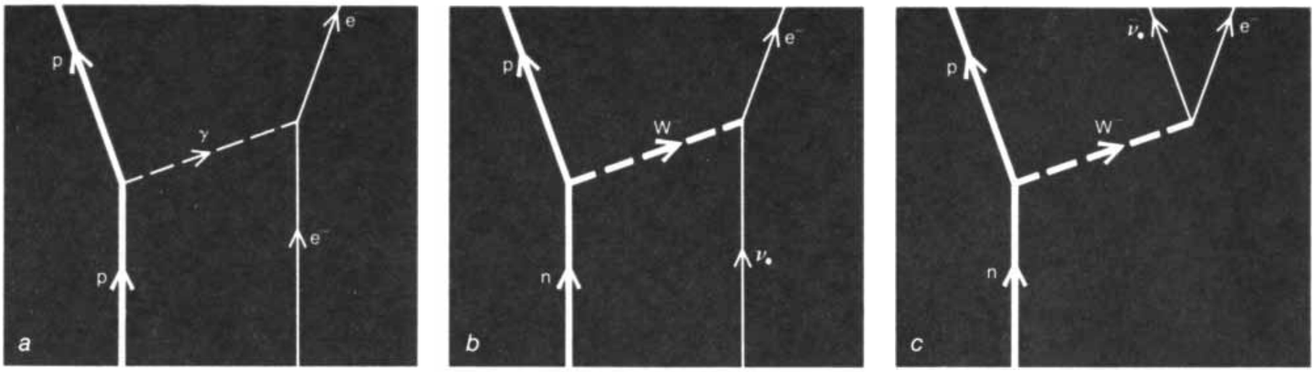
FOUR TYPES OF INTERACTION among particles are believed to account for all physical phenomena. "Range" is the distance beyond which the interaction effectively ceases to operate. In two cases the range is believed to be infinite. "Strength" is a dimensionless number that characterizes the strength of the force under conditions typical of current observations. Thus the gravitational force is some 39 orders of magnitude weaker than the strong force.

It is a simple consequence of the uncertainty principle in quantum mechanics (which states that the uncertainties in our knowledge of the momentum and the position of a particle are inversely proportional to each other) that the range of the force should be inversely proportional to the mass of the exchanged particle. (For an exchanged mass equal to that of the proton the range is about 2×10^{-14} centimeter.) Thus electromagnetism and gravitation, which seem to be of infinite range, are due to the exchange of particles of zero mass: the familiar photon and the hypothetical graviton. The strong interactions are generally believed to arise from the exchange of a large variety of strongly interacting particles, including protons, neutrons, mesons and hyperons of various kinds. Since the weak interactions have a much shorter range than the strong interactions, they must be produced by the exchange of much heavier particles, presumably particles too heavy to have yet been created with existing accelerators.

The Intermediate Vector Boson

For many years it has been speculated that there may be a deep relation between the weak interactions and the electromagnetic interactions, with the difference in their apparent strengths being due simply to the large mass of the particle exchanged in the weak interactions. This hypothesis is supported by the observation that the angular momentum exchanged in weak processes such as nuclear beta decay [*see illustration on opposite page*] has the same value as the angular momentum of a single photon (equal to the Planck constant: 1.0546×10^{-27} erg-second). In fact, the hypothetical particle, or quantum, exchanged in weak interactions has long had a name: the intermediate vector boson. (The term "vector" is used because any particle with this angular momentum is usually described by a field that is a four-dimensional vector, like the vector potential used to describe the photon in James Clerk Maxwell's theory of electromagnetism. The term boson refers to the entire class of particles whose angular momentum is an integer multiple of Planck's constant.)

If we assume that the intrinsic interaction strength of the putative intermediate vector boson that is exchanged in the weak interactions is the same as it is for a photon, then the weak force will have the same strength as the electromagnetic force at short distances; it only appears weaker because of its much



ELECTROMAGNETIC AND WEAK PROCESSES exhibit striking similarities when depicted in the form of Feynman diagrams. Such diagrams symbolize the interactions that underlie subnuclear phenomena, for example the collision between two particles, which physicists refer to as a scattering event. Thus diagram *a* indicates that the electromagnetic scattering of an electron by a proton is due to the exchange of a photon, which transfers energy and momentum from one particle to another. Since time proceeds upward in these diagrams, the photon in *a* is traveling from the proton to the electron, but the diagram is intended also to cover the equally important case in which the photon is traveling in the opposite direction on a trajectory from lower right to upper left. It is precisely this feature of lumping together different processes in a single graph that constitutes the great conceptual value of the "language" of the Feynman diagrams. The scattering of a neutrino by a neutron (*b*) represents a weak interaction in which a heavy particle as yet undetected, the intermediate vector boson, $W^{\pm}$, is believed to play a role analogous to that of the photon in electromagnetic scattering. Here the W particle is assigned a negative charge because it is assumed to be traveling from left to right. It could equally well be regarded as carrying a positive charge and traveling from right to left. The intermediate vector boson is also thought to mediate the radioactive decay of a neutron into a proton, an electron and an antineutrino (*c*). Note that diagram *b* can be obtained from diagram *a* simply by changing some of the particles into others of different charge. Diagram *c* can be obtained from diagram *b* by replacing the incoming neutrino by an outgoing antineutrino. Note also that the total charge is conserved at each vertex in the diagrams.

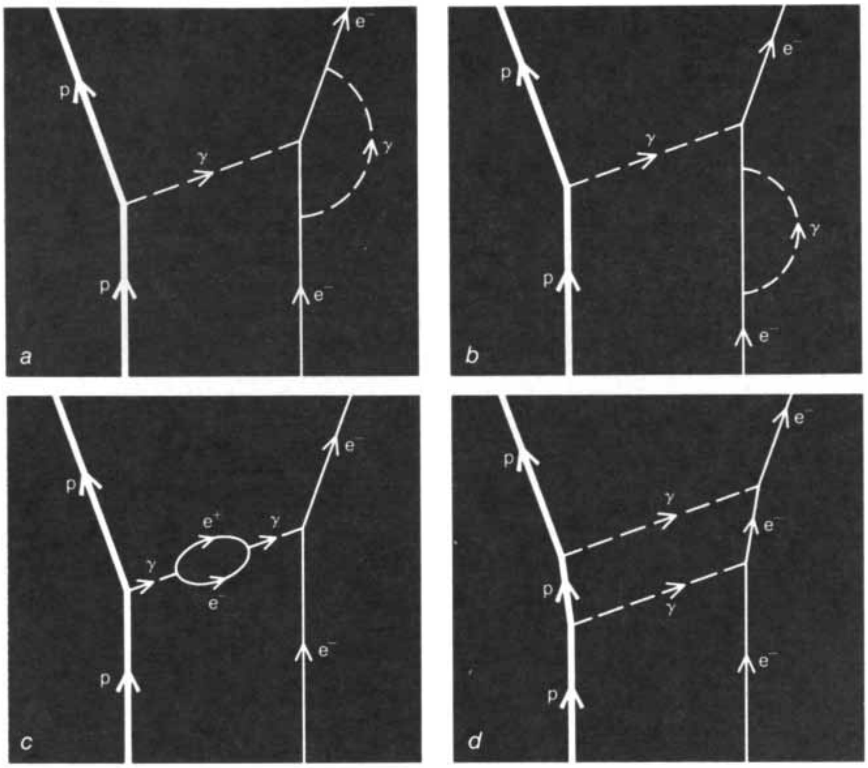
shorter range. The effect of the weak force is reduced in any given process by a factor given by the square of the ratio of the typical masses involved in the process to the mass of the intermediate vector boson. For processes characterized by masses comparable to the mass of the proton, the weak force is roughly 1,000 times weaker than the electromagnetic force. Hence, taking the square root, we conclude that the mass of the intermediate vector boson is roughly 30 proton masses. By applying the conservation of charge to processes such as nuclear beta decay we also see that the intermediate vector boson must carry either a positive or a negative electric charge equal in magnitude to the charge of the proton or of the electron.

A quantum field theory tells us how to calculate the rate for any process in terms of a sum of individual processes, each symbolized by a Feynman diagram such as the one shown above and elsewhere in this article. This useful method for visualizing subnuclear events was introduced some 25 years ago by Richard P. Feynman, and led to a solution of the one great problem that had plagued quantum field theory almost from its birth: the problem of infinities. It was found in the 1930's that the contributions produced by processes more complicated than single-particle exchanges usually turn out to be infinitely large. In fact, the electrostatic repulsion within a single electron produces an infinite self-energy, which manifests itself whenever a photon is emitted and reabsorbed by the same electron [*see illustration on next page*]. These infinities arise only in Feynman diagrams with loops, and they can be traced to the infinite number of ways that energy and momentum can flow through the loop from one particle to another. As is usually the case when paradoxes arise in science, the problem of infinities is both a curse and a blessing: a curse because it keeps us from getting on with calculations we would like to perform, and a blessing because when the solution is found, it may work only for a limited class of theories, among which one hopes to find the true theory.

That is just what seems to have happened with the problem of infinities. In the late 1940's a group of young theoreticians working independently (Feynman, then at Cornell University, Julian Schwinger at Harvard University, Freeman J. Dyson at the Institute for Advanced Study and Sin-itiro Tomonaga in Japan) found that in a certain limited class of field theories the infinities occur only as "renormalizations," or corrections, of the fundamental parameters of the theory (such as masses and charges) and can therefore be eliminated if one identifies the renormalized parameters with the measured values listed in tables of the fundamental constants. For example, the measured mass of the electron is the sum of its "bare" mass and the mass associated with its electromagnetic self-energy. In order for the measured mass to be finite, the bare mass must have a negative infinity that cancels the positive infinity in the self-energy. One simple version of the field theory of electromagnetic interactions not only was found to be renormalizable in the sense that all infinities could be eliminated by a renormalization of the electron's mass and charge but also led to electrodynamic calculations whose agreement with experiment is without precedent in physical science. Thus the theory predicts that the value of the magnetic moment of the electron (in natural units) is 1.0011596553, whereas the observed value is 1.0011596577. The uncertainty in both figures is in the ninth place: ±.0000000030.

Tougher Infinity Problems

In spite of this stunning success, attempts to construct renormalizable field theories of the other elementary-particle interactions long proved to be unsuccessful. For the strong interactions there was no lack of possible renormalizable theories; rather, the trouble was (and indeed still is) that the strength of the interaction invalidates any simple approximation scheme that might be used to draw consequences from a given field theory that could be checked by experiment. (Roughly speaking, the probability of exchanging a set of strongly interacting particles in a high-energy collision is independent of the number of particles exchanged, so that very com-



HIGHER-ORDER CONTRIBUTIONS TO SCATTERING RATES were formerly impossible to calculate when a photon was emitted and reabsorbed by the same electron (*a, b*) or when a photon gave rise to an electron-positron pair that subsequently recombined into a photon (*c*). When physicists see "loops" of this kind in Feynman diagrams, they are prepared to encounter infinities in trying to calculate reaction rates. Thus in diagram *b* the electrostatic repulsion within a single electron manifests itself as an infinite self-energy. It was found in the late 1940's that such infinities can be handled by redefining the mass and charge of the electron, a process called renormalization. Infinity problem does not arise, however, in scattering events such as that in *d*, where the loop has four corners or more.

plicated exchanges have to be taken into account in even the lowest approximation.)

For the gravitational interactions we have a well-known field theory, Einstein's general theory of relativity, which accounts very well for phenomena on the scale of the solar system but seems not to be renormalizable and presumably therefore needs modification for phenomena at very short distances. The problem here is the opposite of that for the strong interactions: gravitational effects are so weak that one can get no help from experimental measurements, at their current level of precision, in finding the correct theory.

The weak interactions present an intermediate case: they are strong enough so that good experimental data are available (although nowhere near as copious as for the strong interactions) and yet weak enough so that approximate calculations are practicable. Even though the weak interactions are believed to be similar to the electromagnetic interactions, however, the theory, as it existed until a few years ago, does not appear to be renormalizable. To be more specific, the exchange of pairs of intermediate vector bosons in processes such as neutron-neutrino scattering [*see illustration on page 57*] leads to infinities that cannot be absorbed in a renormalization of the parameters of the theory. Hence although the quantum field theory of intermediate vector bosons gave a perfectly good approximate picture of the observed weak interactions, it broke down as soon as it was pushed beyond the lowest approximation.

What is the difference between photons and intermediate vector bosons that makes the infinities so much worse for the latter? A detailed analysis enables us to trace the difference back to the fact that the photon has zero mass whereas the intermediate vector boson has ponderable mass. Like all other zero-mass particles, the photon can exist as a superposition of at most two pure states, characterized by left or right circular polarization, in which the axis of rotation is respectively in the same direction as the direction of motion or in the opposite direction. On the other hand, the intermediate vector boson, like any other massy particle whose angular momentum equals Planck's constant, can exist in any one of three states, characterized by an axis of rotation that points in the direction of motion, points in the opposite direction or points in a direction perpendicular to the direction of motion. It is the exchange of intermediate vector bosons whose axes of rotation are perpendicular to the direction of motion that produces the nonrenormalizable infinities.

Types of Symmetry

Before we can see how this problem can be resolved we must first consider the role that symmetry principles have come to play in theoretical physics. Considerations of symmetry have always been important in science, but they acquired special significance with the advent of quantum mechanics. That is because the energy or mass levels of any quantum-mechanical system subject to a symmetry principle are generally required to form certain well-defined and easily recognized families. (In mathematical language one says that the collection of all the mathematical operations on fields that leave the form of the field equations unchanged constitute a "group"; the levels with a given energy or mass are said to form a representation of that group.)

For example, the quantum-mechanical equations that describe the hydrogen atom obey the symmetry principle that all spatial directions are equivalent. As a result the energy levels of hydrogen form families with an odd number of members (1, 3, 5 and so on), the levels within each family being distinguished by the orientation of the axis of rotation [*see top illustration on opposite page*]. When we look at a table showing the masses of elementary particles, we find a similar grouping into families: the proton and the neutron have nearly equal mass; the three sigma hyperons (Σ^+, Σ^0, Σ^-) likewise have nearly the same mass, and so on. For this reason it is believed that the field equations of elementary-particle physics obey isotopic spin symmetry, a symmetry principle analogous to rotational symmetry except that the "rotation" alters the value of the particle's electric charge rather than its spatial orientation. In the early 1960's it was further realized that the various particle pairs, triplets and so on are themselves grouped into larger superfamilies of eight, 10 or even more members, reflecting an approximate symmetry larger than isotopic spin symmetry [*see bottom illustration at right*].

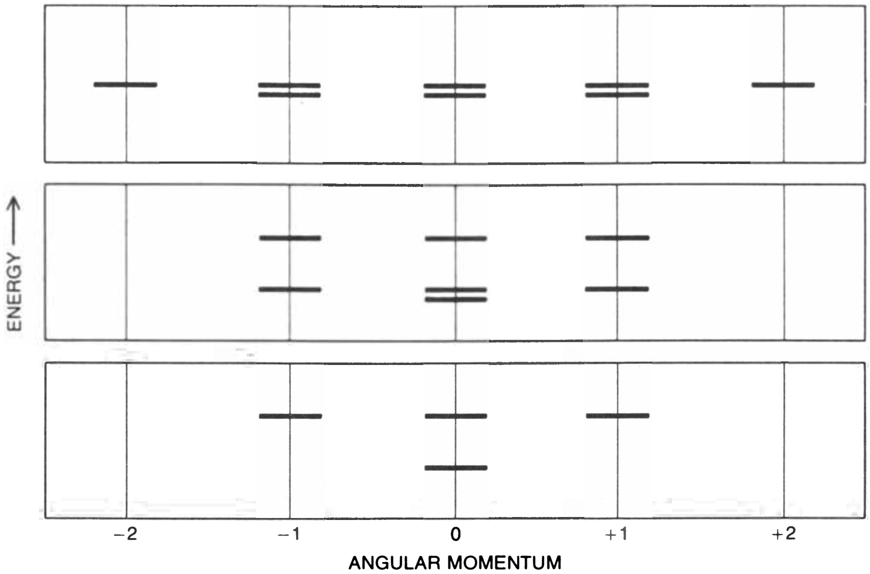
All these symmetry principles require that the field equations do not change when we simultaneously perform well-defined "rotations" on some labeled characteristic of a family or superfamily of particles everywhere in space. One can imagine a much more powerful requirement: that the equations should not change when we perform such rotations of labeled characteristics independently at each point in space and time. The first and less stringent kind of symmetry operation is comparable to giving each apple in a basket the same rotation from one orientation in space to another. The second and more general kind of symmetry operation is comparable to rotating each apple in the basket separately to different new orientations. Invariance under the second kind of symmetry operation is known as a gauge symmetry.

It has been known for many years that the Maxwell field equations of electromagnetism obey a gauge symmetry, based on the group of rotations in two rather than in three directions. Indeed, the logic can be turned backward: assuming this gauge-symmetry principle, one can deduce all the properties of electromagnetism, including Maxwell's equations and the fact that the mass of the photon is zero. It is difficult to conceive of any better example of the power of symmetry principles in physics.

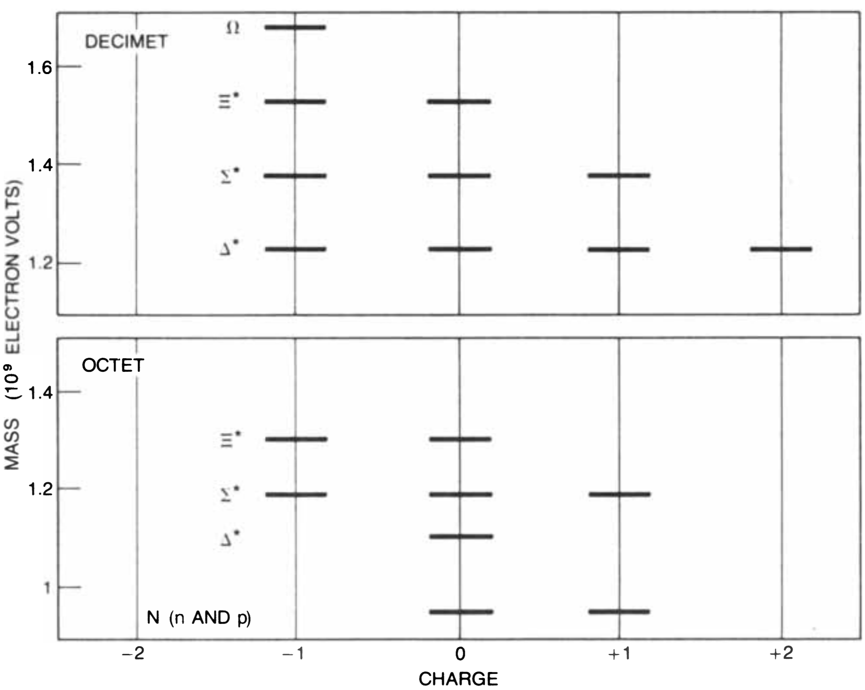
Appearance and Reality

Our hopes of perceiving an underlying identity in the weak and the electromagnetic interactions lead us naturally to suppose there may be some larger gauge symmetry that forces the photon and the intermediate vector boson into a single family. (Indeed, the mathematical theory of generalized gauge symmetries has been understood since the work in 1954 of C. N. Yang and Robert L. Mills, who were then working at Brookhaven National Laboratory.) For this to be possible, however, the intermediate vector boson, like the photon, would have to have zero mass, and we have already seen that its mass is actually much greater than that of any known particle. How can there be any family connection between two such different particles?

The answer to this conundrum lies in considerations of appearance and reality. Inasmuch as symmetry principles govern the form of the field equations, they are generally regarded as providing information about the laws of nature on the deepest possible level. Is it conceivable for a symmetry principle to be valid on this level and yet not be manifest in



SYMMETRY PRINCIPLES require the energy levels of an atom, such as the lower levels of hydrogen shown here, to cluster in well-defined families. Each quantum state of the hydrogen atom is indicated by a short bar. The value of the energy is indicated schematically by the bar's vertical position. (The breaks between panels indicate energy gaps.) The value of the angular momentum around any fixed direction (in units of Planck's constant) is indicated by the horizontal position of the bar. The exact equality of energies within the various triplets, quintuplets and so on is a consequence of the rotational symmetry of the equations that describe the atom, whereas the approximate equality of energies within the various larger families is dictated by more detailed features of the dynamics, such as the weakness of the magnetic coupling between proton and electron, the small value of the electron's charge and the decrease in electrostatic attraction with distance squared.



FAMILIES OF ELEMENTARY PARTICLES are believed to be a consequence of a symmetry principle known as isotopic spin symmetry, analogous to the rotational symmetry that produces the families of quantum states within the hydrogen atom. The grouping of these families of elementary particles into superfamilies (octets, decimets and so on) was proposed independently in the early 1960's by Murray Gell-Mann and Yuval Ne'eman. The bars' vertical position indicates the mass of the particles. Their horizontal position indicates the electric charge in units of the proton's charge. Masses differ slightly from particle to particle even within families, but these differences are too small to be shown here. Particles with an asterisk are very-short-lived states, not included in the table at top of page 52.

the masses and other observed properties of physical particles? The familiar phenomenon of ferromagnetism provides an example of how this can happen.

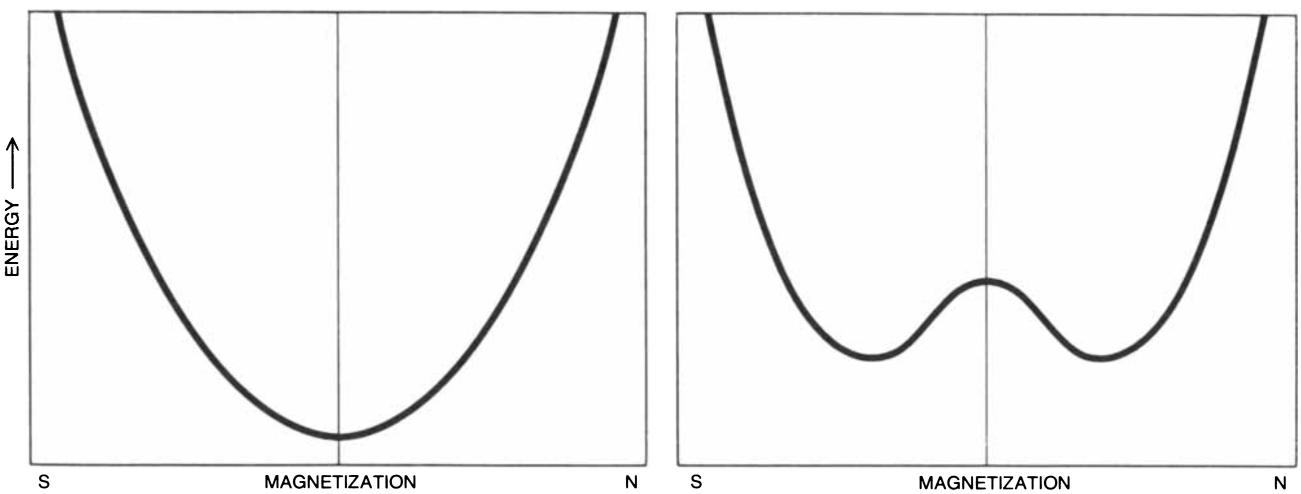
The equations governing the electrons and iron nuclei in a bar of iron obey rotational symmetry, so that the free energy of the bar is the same whether one end is made the north pole by magnetization or the south. At high temperatures the curve of energy versus magnetization has a simple U shape that has the same rotational symmetry as the underlying equations [*see illustration below*]. The equilibrium state, the state of lowest energy at the bottom of the U, is also a state of zero magnetization, which shares this symmetry. On the other hand, when the temperature is lowered, the lowest point on the U-shaped curve humps upward so that the curve resembles a W with rounded corners. The curve still has the same rotational symmetry as the underlying equations, but now the equilibrium state has a definite nonzero magnetization, which can be either north or south but which in either case no longer exhibits the rotational symmetry of the equations. We say in such cases that the symmetry is spontaneously broken. A tiny physicist living inside the magnet might not even know that the equations of the system have an underlying rotational symmetry, although we, with our superior perspective, find this easy to recognize. Reasoning by analogy, we see that a symmetry principle might thus be exactly true in a fundamental sense and yet not be visible at all in a table of elementary-particle masses.

The first proposed example of a broken symmetry of this kind in elementary-particle physics was a non gauge symmetry known as chiral symmetry. (The term chiral is from the Greek for hand, and it is used here because the symmetry consists of independent three-dimensional rotations on fields of left-handed or right-handed polarization. This symmetry contains within it the unbroken three-dimensional rotation group of isotopic spin symmetry.) Chiral symmetry has led to great successes in predicting the properties of low-energy pi mesons, but a discussion of such matters would take us too far afield.

New Theories and Predictions

In 1967 I suggested that the weak and electromagnetic interactions are governed by a broken gauge-symmetry group. (A similar suggestion was made independently some months later by Abdus Salam of the International Center for Theoretical Physics in Trieste.) The proposed group contains within it the unbroken gauge-symmetry group of electromagnetism and therefore requires the photon to have zero mass, but the other members of the photon's family are associated with broken symmetries and therefore pick up a large mass from the symmetry-breaking. In the simplest version of this theory the relatives of the photon would consist of a charged intermediate vector boson (long referred to as the W particle) with a mass greater than 39.8 proton masses plus an additional neutral intermediate vector boson (which I called the Z particle) with a mass greater than 79.6 proton masses. (A theory of this kind is more closely analogous to superconductivity than to ferromagnetism: in a superconductor electromagnetic gauge symmetry is broken and the photon itself acquires mass, as is shown by the fact that a magnetic field can penetrate only a short distance into a superconductor. In particle physics the appearance of vector boson masses in this way is called the Higgs mechanism because it first became known as a mathematical possibility through a 1964 paper by Peter Higgs of the University of Edinburgh.)

At the time I proposed my theory there was no experimental evidence for or against it and no immediate prospect of getting any. There was, however, an internal test of the theory that could be made without help from experiment. We have seen that the infinities in the quantum field theory of pure electromagnetism can be renormalized away, whereas this cannot be done with the existing theory of weak interactions in which intermediate vector bosons have mass. Thus one can ask: Does a field theory become renormalizable when the intermediate vector bosons belong to the same family as the photon and acquire mass only through the spontaneous breakdown of a gauge symmetry? I had suggested in 1967 that this might be the case, but the renormalizability of the theory was not demonstrated until four years later, when it was first shown by Gerhard 't Hooft, then a graduate student at the University of Utrecht. (The proof has since been made more rigorous through the work of many theorists, par-



EXAMPLE OF "BROKEN" SYMMETRY can be found in the two different curves that result when one plots the free energy v. magnetization for a bar magnet at high temperature (*left*) or at low temperature (*right*). The magnet naturally seeks a state of minimum free energy. At high temperature this is a state of zero magnetization, a state that exhibits perfect symmetry between north and south. At low temperature the equilibrium state shifts to one of nonzero magnetization, which can be either north or south, even though the free-energy curve is still perfectly symmetrical between north and south. In this case physicists say that the symmetry is spontaneously broken. The author invokes a similar breaking of symmetry to unify the electromagnetic and weak interactions.

ticularly B. W. Lee, J. Zinn-Justin, M. Veltman and 't Hooft himself.) It turns out that the various multiparticle exchanges involving photons, charged intermediate vector bosons, neutral intermediate vector bosons and other particles add up so as to cancel all nonrenormalizable infinities [*see top illustration on next two pages*].

Once the renormalizability of the theory was established it became clear that the long-sought goal of a unified field theory of weak and electromagnetic interactions might finally be at hand. It then became crucially important to test the theory against experiment. Until such time as intermediate vector bosons can be produced directly the best way to test the theory is to look for effects attributable to the newly predicted neutral intermediate vector boson —the new Z particle—that must appear in the same family as the photon and the charged intermediate vector bosons. The neutral boson does not contribute to processes such as beta decay, in which a charge must be exchanged between the nucleus and the emitted particles. It does, however, contribute along with the charged bosons in processes such as the scattering of "ordinary" neutrinos by electrons [*see "a" and "b" in bottom illustration on next two pages*] and materially changes their rate. Finally, there are processes such as the elastic scattering of "muon-type" neutrinos by electrons and the elastic scattering of any type of neutrino by protons or neutrons [*see "c" and "d" in bottom illustration on next two pages*] that could be produced only by exchanges of neutral intermediate vector bosons.

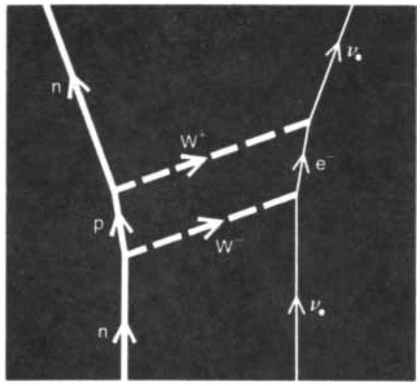
For some years these neutral current processes, as they are called, remained at the edge of detectability, and many physicists doubted their existence. Within the past year, however, evidence for neutral-current processes has at last begun to appear. A pan-European collaboration involving some 55 investigators from seven different institutions, working at the European Organization for Nuclear Research (CERN) in Geneva, has found two events in which muon-type antineutrinos are scattered by electrons and several hundred events in which they are scattered by protons or neutrons. Such scattering events can apparently be explained only by the exchange of a neutral intermediate vector boson, or Z particle, and are therefore direct evidence for a new kind of weak interaction. Moreover, the inferred collision rates agree well with rates predicted by the new theory. An American consortium working at the National Accelerator Laboratory in Batavia, Ill., and another group working at the Argonne National Laboratory have apparently also found neutral-current events [*see illustration on page 51*]. Further experiments aimed at the detection of neutral-current processes and the measurement of their rates are in train at various laboratories in Western Europe, the U.S. and the U.S.S.R.

The existence of neutral-current processes is not yet definitely established, and in any case the general idea of a renormalizable unified field theory based on a spontaneously broken gauge symmetry does not depend absolutely on the existence of neutral-current processes. For instance, in one model suggested by Howard Georgi and Sheldon L. Glashow of Harvard the photon and the charged intermediate vector boson form a family by themselves, although this simplification is achieved at the cost of introducing new particles of other kinds. (There are now a host of other ingenious models suggested by more theorists than can be named here.) There is no doubt, however, that the apparent detection of neutral-current processes has brought welcome encouragement to field theorists.

Some Further Implications

At the same time that experimentalists have been working to test the consequences of unified weak and electromagnetic field theories, theoreticians have been discovering that the new theories freshly illuminate a number of outstanding problems. One, for instance, has to do with the dynamics of the giant stellar explosions known as supernovas. It is believed that supernovas occur at a certain point in the life of a very massive star when the core of the star becomes unstable and begins to implode, or collapse. It has long been a puzzle how the implosion can be reversed and become an explosion, and how the star can shed enough of its outer layers to reach stability as an ultradense neutron star, only 10 or 20 kilometers in diameter. (There is observational evidence that at least two pulsars, believed to be rapidly rotating neutron stars, are embedded in the remnants of past supernovas.)

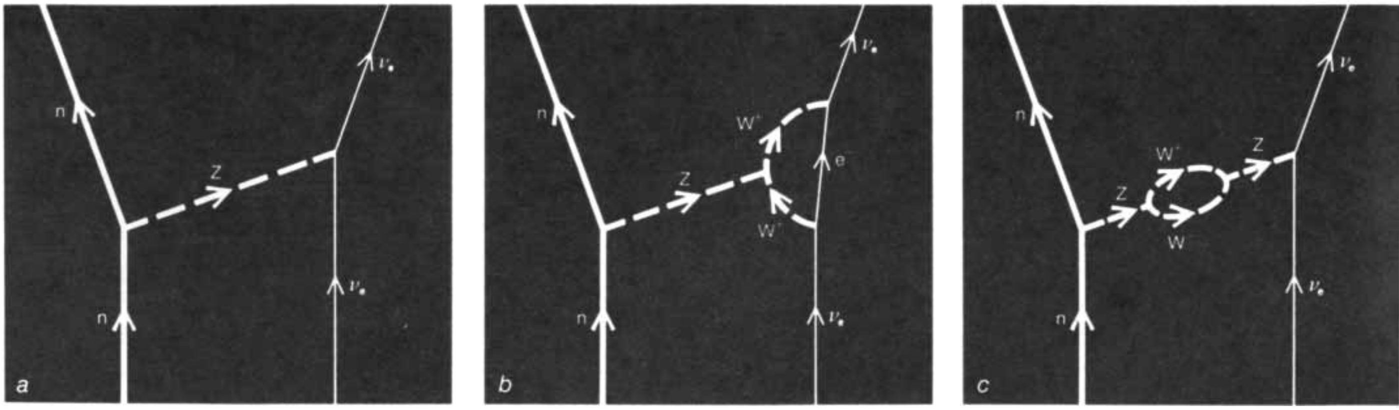
In 1966 Stirling A. Colgate and R. H. White of the New Mexico Institute of Mining and Technology suggested that the outer layers of an exploding star might be blown off by the pressure of neutrinos produced in the hot stellar core, but detailed calculations by James Wilson of the Lawrence Livermore Laboratory of the University of California,



INFINITY PROBLEM ARISES in calculating the rates of neutron-neutrino scattering events involving the sequential (and hypothetical) exchange of two intermediate vector bosons, W^- and W^+. The first exchange converts the neutron into a proton and the neutrino into an electron. The second exchange restores the original cast of characters. Note that the Feynman diagram of this neutron-neutrino event has the same form as diagram *d* in illustration on page 54, in which infinity problem does not arise. The author's recent work, however, indicates how infinities can be removed in processes involving the intermediate vector boson.

using the then available theories of weak interactions, did not support the conjecture. It has recently been pointed out by Daniel Freedman of the State University of New York at Stony Brook that neutral currents can produce "coherent" neutrino interactions, in which a neutrino interacts with an entire nucleus rather than with its individual neutrons and protons. This leads to a much stronger interaction between neutrinos and the relatively heavy nuclei, chiefly the nuclei of iron, in the outer layers of the stellar core. According to Wilson's latest calculations, the increased neutrino pressure is apparently sufficient to produce a supernova.

Another old problem that may be solved through the development of unified gauge theories of weak and electromagnetic interactions concerns the origin of the slight departures from perfect isotopic spin symmetry. The masses of particles within a given family are not precisely equal, generally differing by less than 1 percent to several percent. (The masses of the best-known family pair, the neutron and the proton, differ by only .13 percent.) The differences in mass are about what one would expect if isotopic spin symmetry were respected by the strong interactions but violated by the electromagnetic ones. Calculations along these lines, however, never seem to work. For instance, the electromagnetic self-energy of the proton not



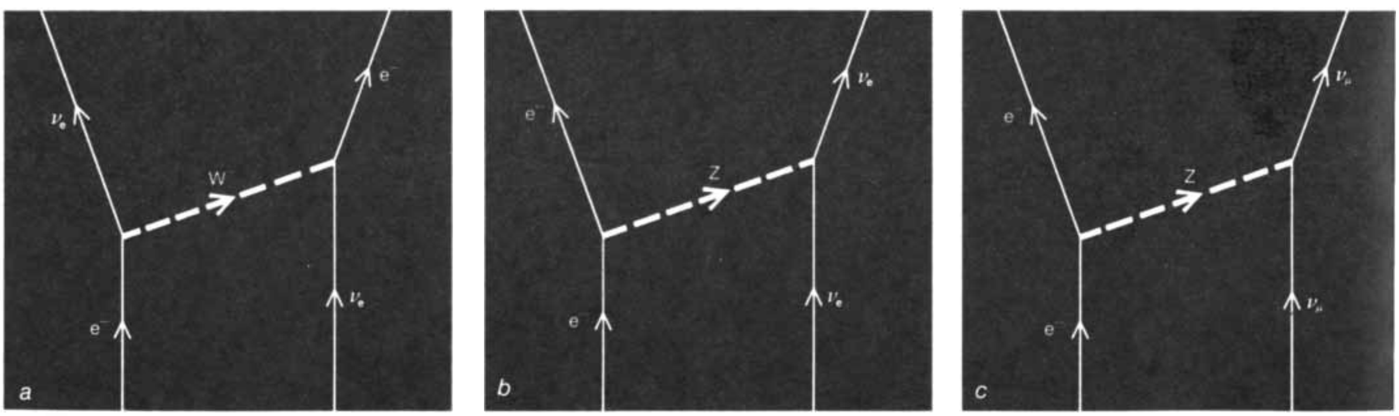
POSSIBLE SOLUTION OF INFINITY PROBLEM in calculating rates of neutron-neutrino scattering may be achieved by postulating the existence of the *Z* particle, a neutral intermediate vector boson. Such a particle is predicted by unified theory of weak and electromagnetic interactions proposed by the author. The *Z* particle should lead to a variety of neutrino-scattering events of the kind diagrammed here. When such processes are added to those involving the charged intermediate vector boson, $W^{\pm}$, it is found that the

only turns out to be positive, contrary to the observation that the neutron is slightly heavier than the proton, but also has an infinite value. This infinity is of the type discussed above, but it cannot be eliminated by renormalization of the bare mass of the proton, if we insist that the bare masses of the proton and the neutron are equal.

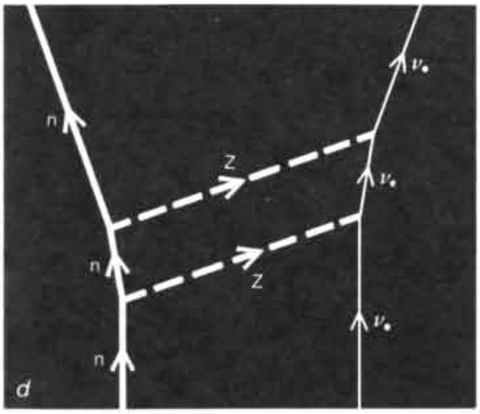
If, as now seems possible, the weak interactions really have an intrinsic strength comparable to that of the electromagnetic interactions, they can provide additional corrections to isotopic spin symmetry that can cancel the infinities due to electromagnetism and leave a finite correction of the right magnitude and sign. Before such calculations can be effectively carried out, however, it is necessary to settle on a detailed model not only of the weak interaction of electrons and neutrinos, as in the 1967 theory, but also of the weak interactions of the strongly interacting particles. This task is still in progress.

Since the weak and electromagnetic interactions seem to be described by a unified gauge-symmetric field theory, it is natural to ask whether the strong interaction can be brought into this picture. There are in fact good reasons for seeking a description of strong interactions in terms of gauge field theories. Possibly the most important is that for a certain class of such theories it is possible to prove that the strong and electromagnetic interactions must necessarily exhibit symmetries between right and left and between matter and antimatter, as is in fact observed to be the case, even though these symmetries are not respected by the weak interactions. As we have seen, the difficulty in testing such field theories is not the lack of experimental data but rather the lack of a method of calculation that can cope with the strength of the strong interactions. Within the past year, however, there has been a theoretical breakthrough that may at last make possible a solution of this problem. David Politzer, a graduate student at Harvard, and independently David Gross and Frank Wilczek of Princeton University, have discovered that in certain gauge field theories the effective strength of the strong interactions at a given energy decreases as the energy rises. In such "asymptotically free" theories it is possible to carry out approximate calculations with the same methods one uses for the weak and electromagnetic interactions, provided that one works at an energy sufficiently high (no one really knows how high) for the strong interactions to be sufficiently weak. Some of the calculations carried out in this way seem to agree quite well with experiment and others do not.

Although it is too early to tell how this will all work out, the development of asymptotically free gauge theories has already led Gross, Wilczek, Politzer, Georgi, Glashow, Helen Quinn and me to an intriguing series of conjectures. If the effective interaction strength be-

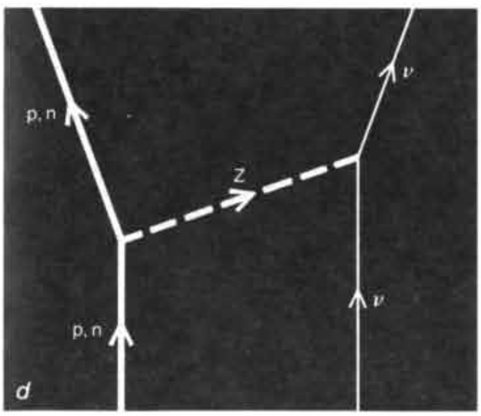


TESTS FOR EXISTENCE OF Z PARTICLE can be made by studying interactions of neutrinos with electrons or with protons and neutrons. The scattering of an electron-type neutrino, ν_e, by an electron can involve the exchange of either a charged intermediate vector boson $W^{\pm}$ (*a*) or a neutral intermediate vector boson *Z* (*b*). Hence the process can be used to test for the *Z* particle only if the rate of the process is carefully measured and compared with theory. In contrast the scattering of a muon-type neutrino, ν_μ, by an electron (*c*) or of any kind of neutrino by a proton or neutron (*d*) can occur only by exchange of a *Z* particle and therefore provides



infinities appearing in the sum of the contributions to the total rate, symbolized by all such diagrams, can be absorbed into a renormalization of parameters of new theory.

comes small at high energies and short distances, then it must become large at low energies and large distances. Perhaps this explains why the ordinary elementary particles cannot be broken up into quarks: as a quark is pulled away from the rest of the particle the forces may increase without limit. Perhaps the intrinsic interaction strength of the strong interactions is really of the same order of magnitude as that of the weak and electromagnetic interactions and only appears stronger because our present experiments happen to be carried out at relatively low energies and large distances. Perhaps the strong interactions are really caused by the exchange of particles that belong to the same family as the photon and the intermediate vector bosons that are responsible for the electromagnetic and weak interactions. If these speculations are borne out by further theoretical and experimental work, we shall have moved a long way toward a unified view of nature.



direct evidence for a new kind of weak interaction. These interactions are neutral-current processes such as the one shown in the bubble-chamber picture from Argonne National Laboratory at top right on page 51.

THE GLORY

This halo of prismatic colors is most often seen around the shadow of an airplane on a cloud. Its cause is not the same as that of the common rainbow, and involves phenomena at the frontier of physics

by Howard C. Bryant and Nelson Jarmie

"If it be shortly after sunup of a morning when the fog has obliterated the highway below, I am then rewarded with a spectacle rare to witness. Looking up the coast toward Nepenthe... the sun rising behind me throws an enlarged shadow of me into the iridescent fog below. I lift my arms as in prayer, achieving a wingspan no god ever possessed, and there in the drifting fog a nimbus floats about my head, a radiant nimbus such as the Buddha himself might proudly wear. In the Himalayas, where the same phenomenon occurs, it is said that a devout follower of the Buddha will throw himself from a peak—'into the arms of Buddha.'"

So does Henry Miller, in *Big Sur and the Oranges of Hieronymus Bosch,* describe his observation of the meteorological phenomenon known as the glory. He conveys his feeling of apotheosis on viewing his shadow on a fog bank, "glorified" with colored rings similar to those of the rainbow. The spectacle is indeed a rare one for ground-based observers such as the solitary hiker in Miller's narrative, because it requires an unusual configuration of the sun, the observer and a cloud composed of droplets of uniform size. It is seen regularly, however, by air travelers, particularly those who know where to look. In fact, it is sometimes called the pilot's bow. Other names for the glory are the anticorona and the brocken bow.

Like the common rainbow, the glory is caused by the scattering of sunlight by droplets of water. Like the primary rainbow, the brightest in a series of rainbows, it consists of concentric rings of color, with red the outermost and violet the innermost, encircling a bright central region in the direction opposite to that of the sun. Unlike the primary rainbow, whose red ring is invariably at an angle of 42 degrees from the direction of the shadow cast by the observer, the glory has rings whose angular diameter varies inversely with the diameter of the droplets that give rise to them. The primary rings are often accompanied by as many as four similar sets of rings of larger angular diameter. Typically the innermost red ring has a diameter of two or three degrees.

To see a glory close up you must view the cloud of uniform water droplets in such a way that your shadow is projected on the cloud. You will be rewarded by a vision of the shadow of your head surrounded by a series of colored haloes. Moreover, a feeling of uniqueness may be oddly enhanced: if someone else is with you, his shadow will not appear to be so endowed. One may even speculate that the artistic practice of rendering the heads of holy or powerful personages with luminous and sometimes colored haloes or nimbuses could have arisen from the observation of such haloes on fog banks by solitary mystics on well-illuminated heights. The use of the halo is not restricted to Christian iconography: glorylike structures can be seen surrounding the heads of Roman emperors and Greek gods, and of icons from China, Burma and India, suggesting that such representations may have a universal natural origin.

The first scientific record of the glory was a drawing made by Antonio de Ulloa during a French expedition to Peru in 1735. Both he and Pierre Bouguer also wrote descriptions, translations of which can be found in R. A. R. Tricker's *Introduction to Meteorological Optics.* Balloonists in the 19th century were able to see the glory encircling the shadow of the basket of the balloon, as Gaston Tissandier relates in his *Observations météorologiques en ballon.* These sightings and others are described in detail in the classic work *Meteorologische Optik,* by Josef M. Pernter and Felix M. Exner, published in Vienna in 1910.

C. T. R. Wilson built the first cloud chamber in 1895 for the purpose of recreating the glory in the laboratory. He put aside his original objective when he found that energetic charged particles leave visible tracks of water droplets in the moist air. In his Nobel prize lecture of 1927 he reported: "In September, 1894, I spent a few weeks in the observatory which then existed on the summit of Ben Nevis, the highest of the Scottish hills. The wonderful optical phenomena shown when the sun shone on the clouds surrounding the hilltop, and particularly the colored rings surrounding the sun (coronas) or surrounding the shadow cast by the hilltop or observer on mist or cloud (glories), greatly excited my interest and made me wish to imitate them in the laboratory."

What is the explanation of this lovely apparition? We have indicated that the glory is caused by the scattering of light from water droplets back toward the source of the light. In the process the scattered light is enhanced. To be sure, light is enhanced when it is scattered backward from many things other than fog or clouds, including plowed fields, foliage, the eyes of many animals and dewy grass. Let us first discuss some of these nonglory effects to illustrate the variety of the mechanisms at work.

The "cornfield effect," or backscattering from a plowed field or foliage, can actually be observed on any rough surface that casts small shadows. When we look at the surface from the direction of the illumination, we do not see the shadows and the surface looks unusually bright. The reason is that the brightness is contrary to our expectation: from other directions parts of the surface are darkened by the shadows, and the eye

GLORY FROM THE AIR surrounding the shadow of an airplane projected on a cloud below is seen in this photograph made by Fritz Goro. Since the glory is often visible under these circumstances, it is sometimes called the pilot's bow. Around the shadow there is a glow covering a circular area half-obscured by the shadow. Surrounding the central area are two sets of concentric colored rings with blue on the inside and red on the outside. To unaided eye the colors are sometimes much brighter than they appear here.

GLORY FROM THE GROUND was photographed by John C. Brandt of the National Aeronautics and Space Administration at the Zodiacal Light Observatory on Haleakala Crater on the Hawaiian island of Maui. The picture was taken before sunset as the crater was filled with cloud or fog. Five glory rings are visible; their angular radii, measured from the red rings, are 1.2 degrees, 3.0 degrees, 4.9 degrees, 6.7 degrees and 8.3 degrees. Normally the shadow of the person viewing the glory is visible, but here it is not.

integrates the image so that the surface appears uniformly darker.

Backscattering from animal eyes arises from the fact that the illuminating light ray is reversed with high precision. Light entering the animal's eye from the direction of the observer is brought to a focus on the animal's retina; then some portion of it is scattered back toward the lens of the animal's eye and refracted back in the direction from which it came. The result is that the animal's eye appears to be illuminated from within.

The effect is enhanced in cats, dogs, rabbits and other animals because they have a reflecting layer behind the retina. Man lacks this layer, but on occasion his eyes backscatter very well. David L. MacAdam, editor of *The Journal of the Optical Society of America*, comments: "Any photographer who has taken many close-up color pictures with a flash lamp built into his camera, close to the lens, will recall his dismay when some of his pictures were ruined by brilliant red spots coincident with the pupils of the eyes of some of his subjects. A considerable portion of fair-haired, light-eyed persons have such strong reflection from the fundus of the eye as to produce this Heiligenschein [the German for halo]. It can also be seen clearly by another person over whose shoulder a bare incandescent tungsten bulb is shining directly into the eyes of the subject, in an otherwise poorly lighted room."

It is effects of the Heiligenschein type that are seen against a background of dewy grass. When one looks at the grass from the direction of illumination, the shadow of one's head appears to be surrounded by a bright area. As with animal eyes, the water droplets, which are more or less spherical, serve as miniature converging lenses that collect the light and focus it on the blades of grass on which they rest. Much of the focused light is scattered in all directions by the leaf, but some of it reenters the droplet and is refracted backward in the direction from which it came.

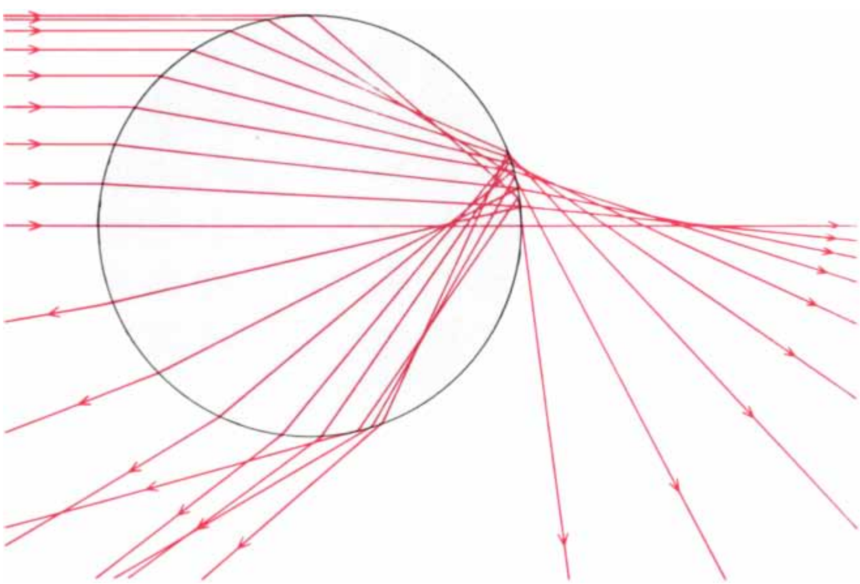
A completely different mechanism is needed to account for the glory. In 1947 the Dutch astronomer H. C. van de Hulst put forward the explanation that the light of the glory is sent back from the edges of the spherical water droplets in the cloud. For the moment let us set aside the question of how the light is returned from the droplet's edge and concentrate on understanding how the glory would be produced in that way by a random distribution of uniform droplets.

Since each droplet is returning the light from its edge, each is effectively a ring-shaped light source sending light back toward the sun. One way to simulate examining the optics of a field of backward-scattering water droplets is to replace the droplets with an opaque screen that has ring-shaped apertures in it and illuminate the screen from behind with a beam whose rays are parallel [*see bottom illustration on page 64*]. When the screen is viewed from a distance, the resulting diffraction pattern, or distribution of the intensity of the scattered light, will be very much like the diffraction pattern of the glory.

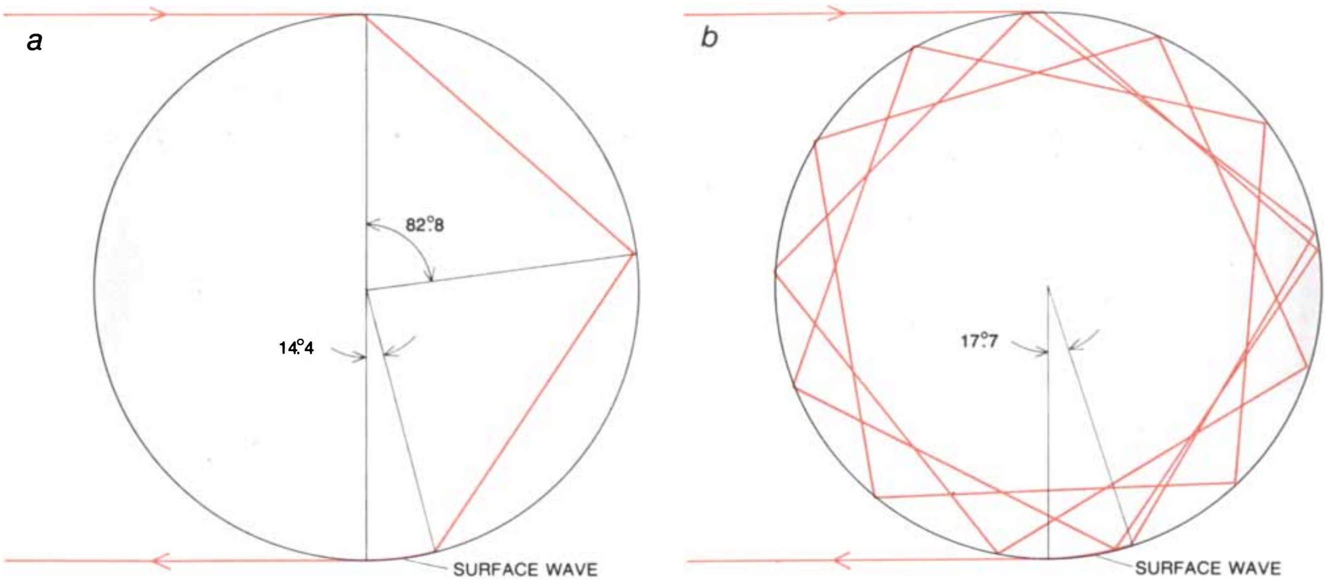
Why should this be so? The diffraction pattern from a single ring can be understood by regarding each point on the ring as a separate source of coherent light waves. This means that all the wavelets coming from each point on the ring will be in phase with one another. The light arriving at any particular point on another screen at some distance from the aperture will consist of contributions from all the points around the ring. Only on the part of the screen that is directly in front of the ring will the wavelets be exactly in phase, since it is only there that the distance each wavelet has traveled from the ring's edge is the same as the distance every other wavelet has traveled. At that point on the screen there will be a circular spot of maximum brightness.

In a region at a small angle away from the spot of maximum brightness the path each wavelet takes is either longer or shorter than the path of its neighbor, and the wavelets begin to interfere destructively with one another. The intensity of the light falls to a minimum, and there is a ring of minimum brightness around the central bright spot. At a greater angle away from the bright spot the light intensity begins to increase again to a second-order maximum: wavelets from opposite sides of the ring have a difference of one wavelength in the distance they travel to the screen and are back in phase again. This second-order maximum (and successive maximums at even greater angles, for which the differences in path length differ by two, three, four or any other whole number of wavelengths) is not as intense as the principal maximum in the center because it is only at that one central spot that all the wavelets are in phase.

So far we have described what happens with only one ring. To illustrate what happens in a glory we prepared a random array of many ring sources. First we drew 241 circles two millimeters in diameter with black ink on a piece of white paper about 12 centimeters on a side. Then we photographed the piece of paper to get a 35-millimeter negative; the negative showed transparent rings on a black background reduced in size by a factor of 16. We set up a parallel-ray source of coherent light by passing the red beam of a helium-neon laser through a shutter into a microscope lens with a pinhole at its focal point. The beam was then directed through a converging lens whose focal point was also at the pin-



LIGHT THAT ACCOUNTS FOR THE COMMON RAINBOW is not sent straight back toward the observer. The primary rainbow is caused by parallel light rays that are incident from the left being refracted off to lower left by a spherical water droplet. The rays to the right of the droplet are those that go straight through without contributing to the rainbow.

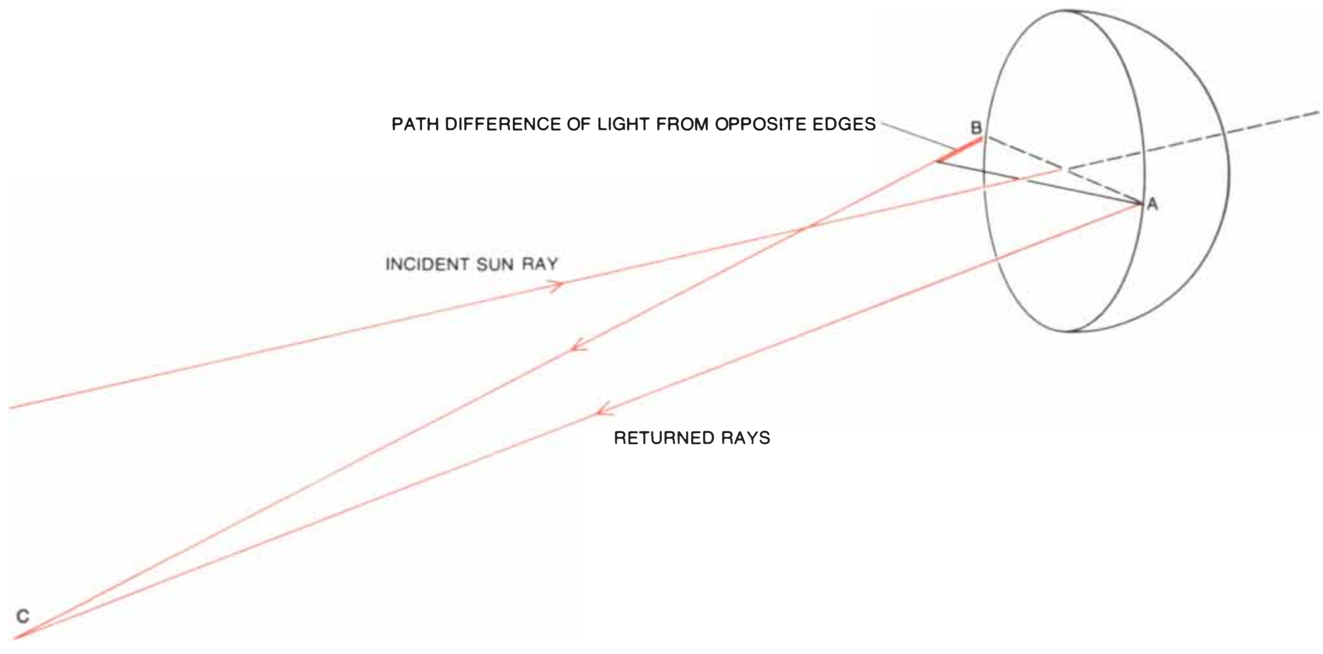


LIGHT THAT ACCOUNTS FOR THE GLORY follows paths that are different from those that are responsible for the common rainbow. The paths consist of light that is reflected repeatedly at an angle of 82.8 degrees within the droplet, together with small segments of surface waves in which the light clings to the surface of the droplet and is conveyed the rest of the way around the droplet to the backward direction. When a number of different paths give rise to light waves that are in phase, there is a resonance, or enhancement, in the backscattered light. The glory is believed to be due principally to the paths in which the light travels halfway around the droplet (*a*) and those in which it travels three and a half times around the droplet (*b*) before being sent straight back.

hole. This train of optical devices gave us a parallel-ray beam of coherent light two centimeters in diameter, into which we inserted the photographic negative of the array of rings.

We now had a field of coherent ring sources, each about 125 micrometers across, which together simulated the effect of a uniform field of spherical droplets that are backscattering light. In order to record the diffraction pattern as it would appear at a distance from the array we put a lens with a focal length of one meter in front of the array and placed a sheet of photographic film in its focal plane. The resulting photograph of concentric bright and dark rings approximately represents the appearance of a glory seen through a red filter [*see illustration on page 65*].

We can now explain the presence of

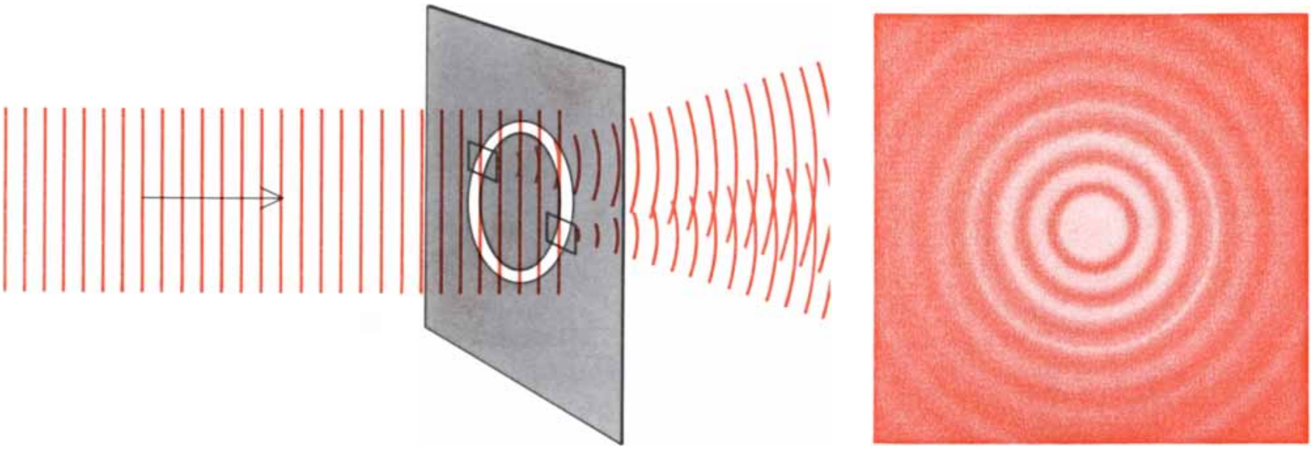


EDGES OF WATER DROPLETS of uniform size scatter sunlight back toward the observer in the formation of a glory. Here the path of rays returning from a single droplet is traced; the droplet is cut in half to show the geometry of the rays. If the observer is straight back from the droplet, the waves on every ray from the edge of the droplet will be in phase and will reinforce one another. As the angle of the observer to the path of the light falling on a droplet increases, the waves from opposite edges begin to go out of phase and interfere. As the angle increases further, they come back into phase and reinforce again. A secondary maximum in brightness is reached near the point where the difference between paths *AC* and *BC* is one wavelength. The rings of the glory are colored because the many different wavelengths in sunlight go through their cycles of maximum and minimum brightness at different angles.

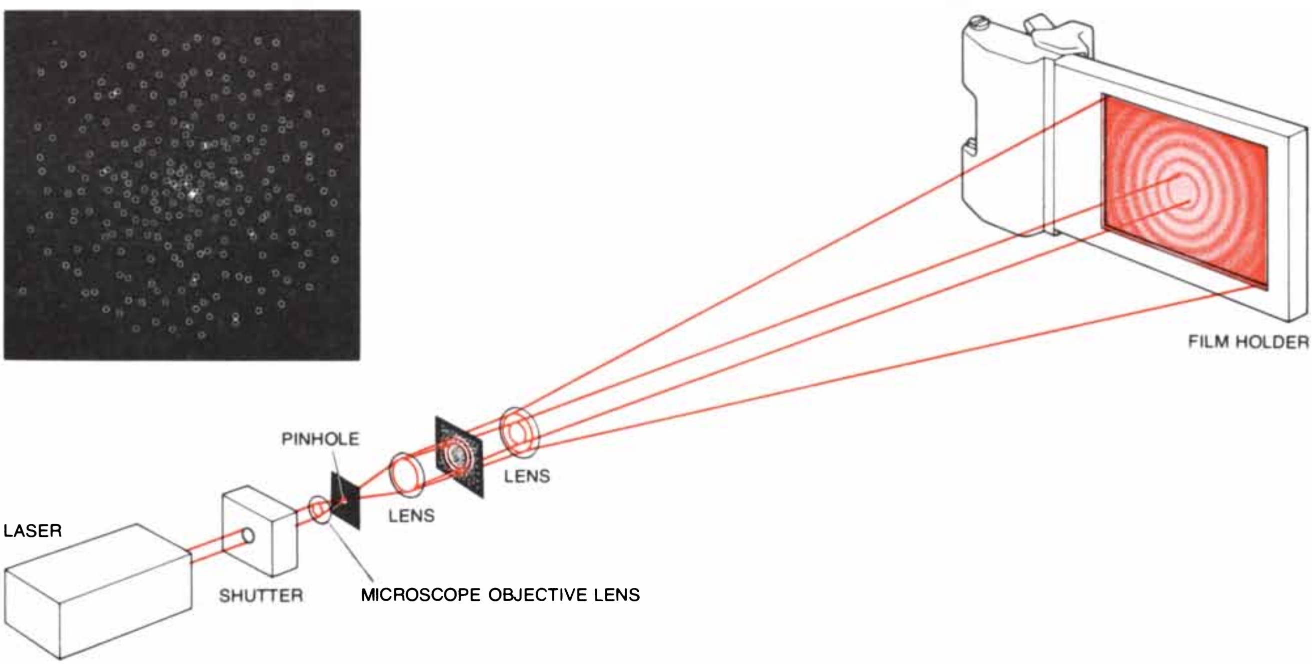
colored rings in the glory. When each wavelength in the solar spectrum is backscattered from an array of droplets of uniform diameter, it will give rise to a pattern similar to the pattern of concentric rings in the photograph. The longer the wavelength, the larger the diameter of the diffraction rings. Although all wavelengths will contribute to the bright central maximum, at other angles only light of certain wavelengths will be at a maximum. Therefore the light in the glory away from the central maximum will be colored. In every sequence of spectral colors red will be at the largest angle from the center, since it has the longest wavelength. Thus the outermost ring of the glory is always red.

Of course, our artificial ring sources are only an approximation of the optical properties of an array of backscattering water droplets. One difference is that the laser light giving rise to the diffraction pattern in the artificial array is not backscattered. We have left out the effects of the polarization of the light; we have also neglected the beam directly reflected from the center of the droplet. The correct treatment of polarization alters the distribution of the light's intensity somewhat, and the inclusion of the central point source produces an enhancement or diminution of every other ring, depending on its phase.

We now return to the question of how droplets scatter light backward from their edges. The fact that they do so can be demonstrated by constructing a large transparent sphere and directing sunlight at it. Permanently mounted in a lecture hall at the University of New Mexico is a heliostat, an optical mechanism made up of a series of mirrors that tracks the sun across the sky during the course of a day and sends a sunbeam 35



SINGLE RING APERTURE (*left*) **illuminated by plane-parallel light** (*color*) **gives rise to a diffraction pattern that is a series of concentric circles. The ring aperture can be regarded as one droplet in the cloud that causes the glory. Each point on the ring** (*indicated by the two rectangles*) **acts as a separate source of wavelets of coherent light. Wavelets interfere either constructively or destructively at different angles away from direction of light coming straight back. This interference produces the light and dark rings** (*right*).



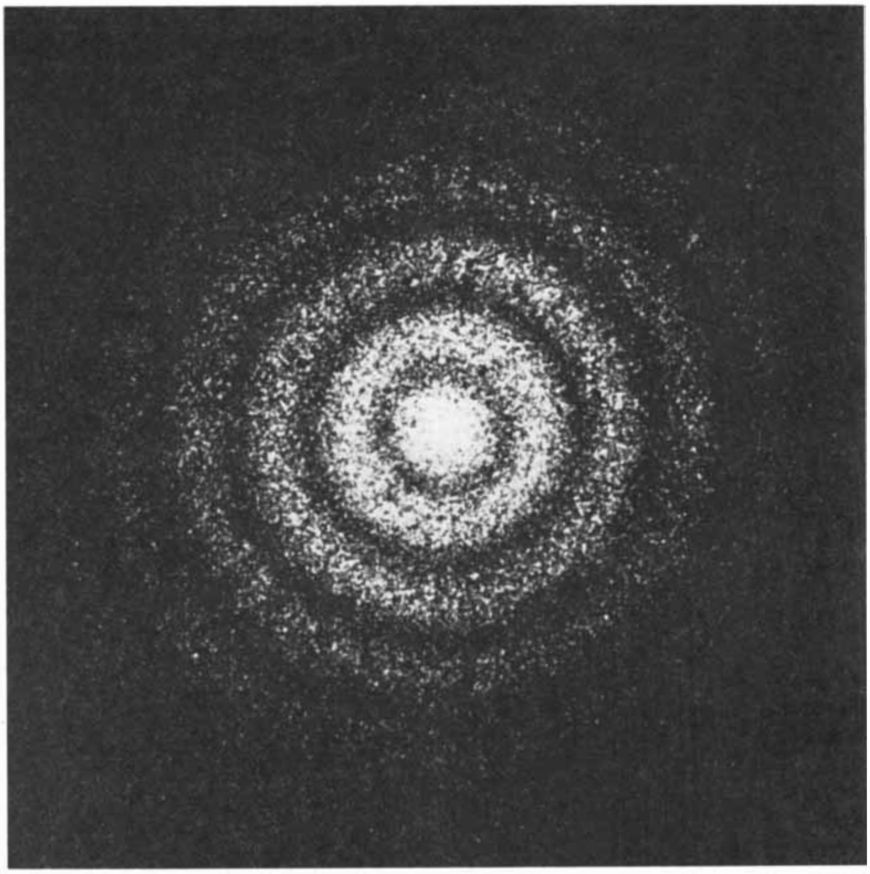
MANY RING APERTURES together (*insert at top left*) **act as a cloud of uniform water droplets and generate a diffraction pattern that resembles the glory. First a series of 241 randomly placed circles were drawn on white paper. Then a much reduced photographic negative was made from the drawing. A parallel-ray beam of red light from a helium-neon laser was directed through the array of rings; a lens formed an image of the diffraction pattern that was recorded on Polaroid film** (*see illustration on opposite page*).

centimeters in diameter across the front of the hall. Into this beam we can insert a Lucite sphere with a diameter of 30.48 centimeters (12 inches).

In addition to producing a marvelous rainbow that covers one of the white walls of the lecture hall, the system demonstrates the "backward" optics of transparent spheres. We can view the sphere from the direction of the sunbeam by inserting a small mirror into the beam. It can be seen that the backward-directed light indeed comes principally from the edge of the sphere, with a small additional contribution from specular reflection at the center of the sphere's face.

When the sphere is viewed from a point that is even at a slight angle to directly backward, its appearance changes considerably. Its edge is still well illuminated, and the equatorial region of the edge is particularly bright. The specular reflection from the face is off center.

If a plant leaf or a piece of paper is placed five centimeters behind the sphere, one can see the Heiligenschein. The sphere becomes very bright. The presence of the leaf greatly enhances the amount of light sent backward in the direction from which it came, to the extent that the glow reflected back to the mirror of the heliostat can be seen throughout the lecture hall.



DIFFRACTION PATTERN that approximately represents the appearance of a glory seen through a red filter was photographed with the apparatus that is shown in the illustration at the bottom of the opposite page. Graininess in the picture is caused by interference among individual apertures; if the number of apertures were increased, graininess would decrease.

This procedure with sunlight and the Lucite sphere is instructive, but more relevant studies have been made in the laboratory with laser light and real droplets of water suspended in midair. Theodore S. Fahlen, who was then working at the university, found that droplets could be suspended in two different ways. In the first method a resonator made from a cylindrical piezoelectric crystal and a circular watch glass produced an acoustical standing wave. Droplets as large as a millimeter in diameter could be suspended in midair at the region of maximum acoustical pressure for up to several minutes without appreciable vibration.

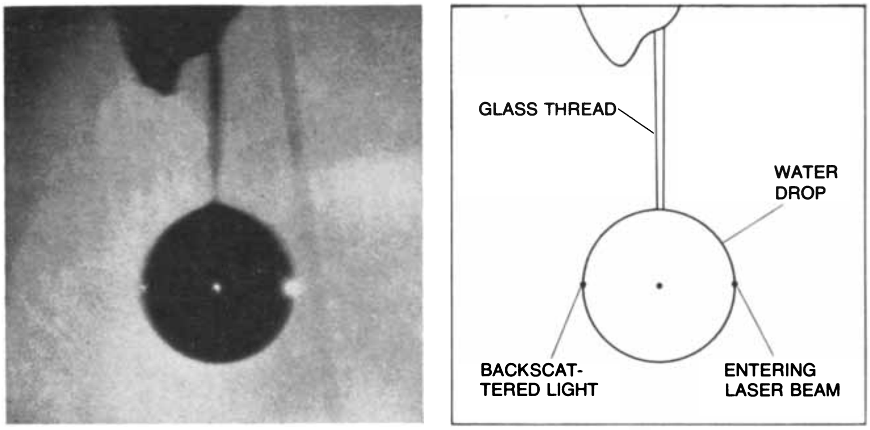
The second method simply entailed suspending the droplet by its own surface tension from a glass fiber whose end had been enlarged to form a tiny bead; the fiber could hold droplets two or three millimeters in diameter. This method proved to be more useful than the other one. If the optical effects are to be observed, the surface of the droplet must be exceedingly quiet and smooth, much quieter and smoother than the droplets that were levitated in the acoustical resonator. M. J. Saunders of Bell Laboratories has used the fiber method to make and study water droplets as small as nine microns in diameter. (His fibers were bits of spider web.) Saunders has verified that even with such small droplets the backscattered light comes from the edge of the sphere.

Why should the light be scattered mostly from the edge? In trying to answer that question we approach some current frontiers of optics and particle physics, even though the precise solution of how an electromagnetic wave (in this case light) is scattered from a transparent sphere has been known since the beginning of the century.

In 1908 the German physicist Gustav Mie showed that the intensity of an electromagnetic wave scattered from a sphere can be calculated as precisely as one wants for any angle, including angles in the backward direction. He showed that the intensity of the light scattered at any angle can be represented as a sum of a series of algebraic terms, each composed of involved mathematical expressions. These terms are not of the type that can be easily computed on the back of an envelope. In addition the number of terms that must be evaluated in order to arrive at the intensity at a given angle for a specific wavelength is somewhat larger than the circumference of the sphere in nanometers divided by the wavelength in nanometers. For a droplet one millimeter in diameter, for example, and green light of a wavelength of 500 nanometers, some 6,300 terms have to be evaluated and added together. And to get the entire intensity pattern for just one wavelength of light requires repeating the process for a number of angles.

Thus in addition to a desire to learn the details of Mie's predictions for a given case one must have access to the services of a high-speed computer. It is only within the past decade that Mie calculations for large spheres have been conducted to any extent. To make matters worse, the intensity of the backscattered light is extremely variable. As A. J. Cox has demonstrated at the University of New Mexico, a very small change in the wavelength of the light can result in a change of intensity as large as a factor of 100.

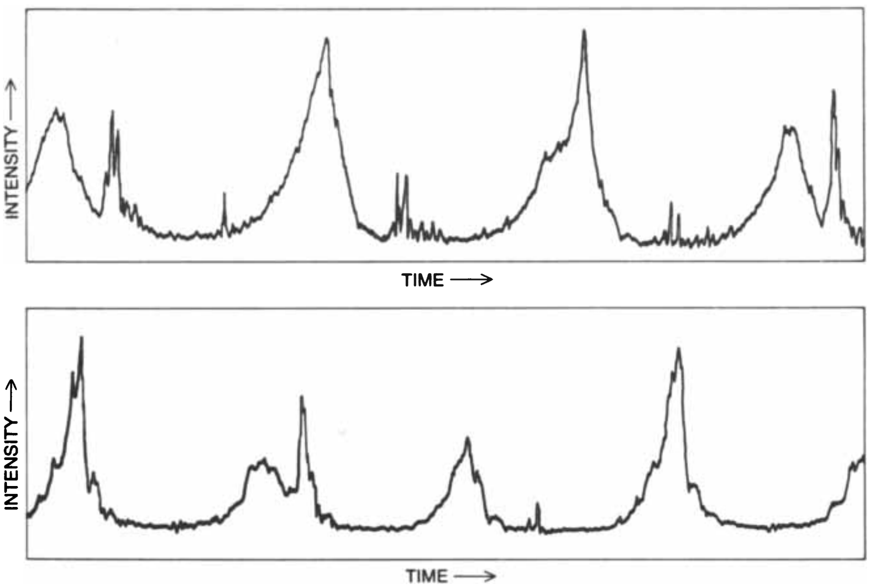
There are less exact but perhaps more instructive ways to treat the scattering



WATER DROPLET SUSPENDED from a glass thread returns a beam of laser light straight back toward the observer. The beam entered at the right edge of the droplet and was sent back at the left edge. Experiment demonstrates that in giving rise to the glory, light waves must be conveyed around the surface of the droplets. Droplet is two millimeters in diameter; thread is 40 micrometers thick. Central bright spot on droplet is due to backlighting.

of light. For instance, primary and secondary rainbows can be explained in terms of geometrical optics: the primary bow is produced by light that is reflected once inside a droplet of water and the secondary bow is produced by light that is reflected twice. The glory cannot be explained on this basis, because in an ordinary rainbow the rays striking the edge of the droplet do not come straight back. It is nonetheless possible to understand in terms of ray optics the small contribution made to the glory along the axis of the incident light beam.

Rays are reflected not only from the outer surface of the drop but also from the inner surface. Since only some 2 percent of the light falling perpendicularly on an interface between air and water is reflected, the amount of light traversing paths that involve more than one such reflection will be negligible compared with the two direct reflections. The relative phase of the two rays comprising the axial contribution varies with the diameter of the droplet. The two beams interfere with each other to give rise to an intensity that varies on a sine curve as the diameter of the droplet varies. This property has been used, in fact, to make precise determinations of the rate at which droplets of various sizes evaporate.



OSCILLOGRAPH TRACINGS of the intensity of a laser beam emerging from the margin of an evaporating droplet over a period of 20 minutes show a repetitive pattern of three humps on which are imposed spikes. The tracing at the top was made when the droplet had a diameter of 1,153 micrometers; the tracing at the bottom was made when it had shrunk to 741 micrometers. Humps are due principally to paths taken by light traveling halfway around droplet and 3½ times around it. Spikes are resonances from hundreds of paths.

The explanation of why the light is returned from the edge of the droplet is surprising even to many people familiar with physical optics. There must be a process that takes light impinging on one edge of the droplet and sends it back in the opposite direction from the opposite edge. The mechanism that produces the rainbow, in which rays are refracted into the sphere, are reflected internally once and are refracted back out again at a preferred angle of 42 degrees from the backward direction, at first appears to be inadequate to explain the glory. Van de Hulst proposed that an internally reflected ray is important but that it is combined with another mechanism we have not yet considered: the surface wave.

When a light beam strikes the air-water interface from the water side, there is a critical angle at which the beam refracted into the air is parallel to the interface. At that angle the beam has a long "tail" on the "downstream" side. This tail is the surface wave, or more properly the lateral wave, and it carries some of the light farther around the circumference of the droplet than one would expect from geometrical optics alone. Thus the surface wave provides the necessary ingredient that makes van de Hulst's hypothesis viable.

As the diameter of the droplets that give rise to the glory is increased, the intensity of the glory shows periodic resonances, or fluctuations. These resonances indicate that the effect must be due to two or more beams whose difference in phase depends on the size of the droplet. When the two or more beams interfere constructively, the intensity of the glory will be high; when the beams interfere destructively, the intensity will be low.

The existence of such multiple-beam resonances has been confirmed experimentally. Fahlen suspended a very quiet evaporating droplet and focused light from a helium-neon laser on it through a small hole in a mirror set at an angle of 45 degrees. The returning light was reflected by the same mirror through a system of variable-aperture lenses onto the photocathode of a photomultiplier, which converts the light into an electric current. When the laser beam was directed along a tangent to one side of the droplet, the returning beam was received coming back along a tangent to the opposite side of the droplet. The amplitude of the current produced by the photomultiplier was recorded with a light-beam oscillograph, an instrument that writes on a rapidly moving strip of photosensitive paper with a narrow ultraviolet beam. The deflection of the beam along the paper was made proportional to the amount of laser light received by the photomultiplier [*see bottom illustration on opposite page*].

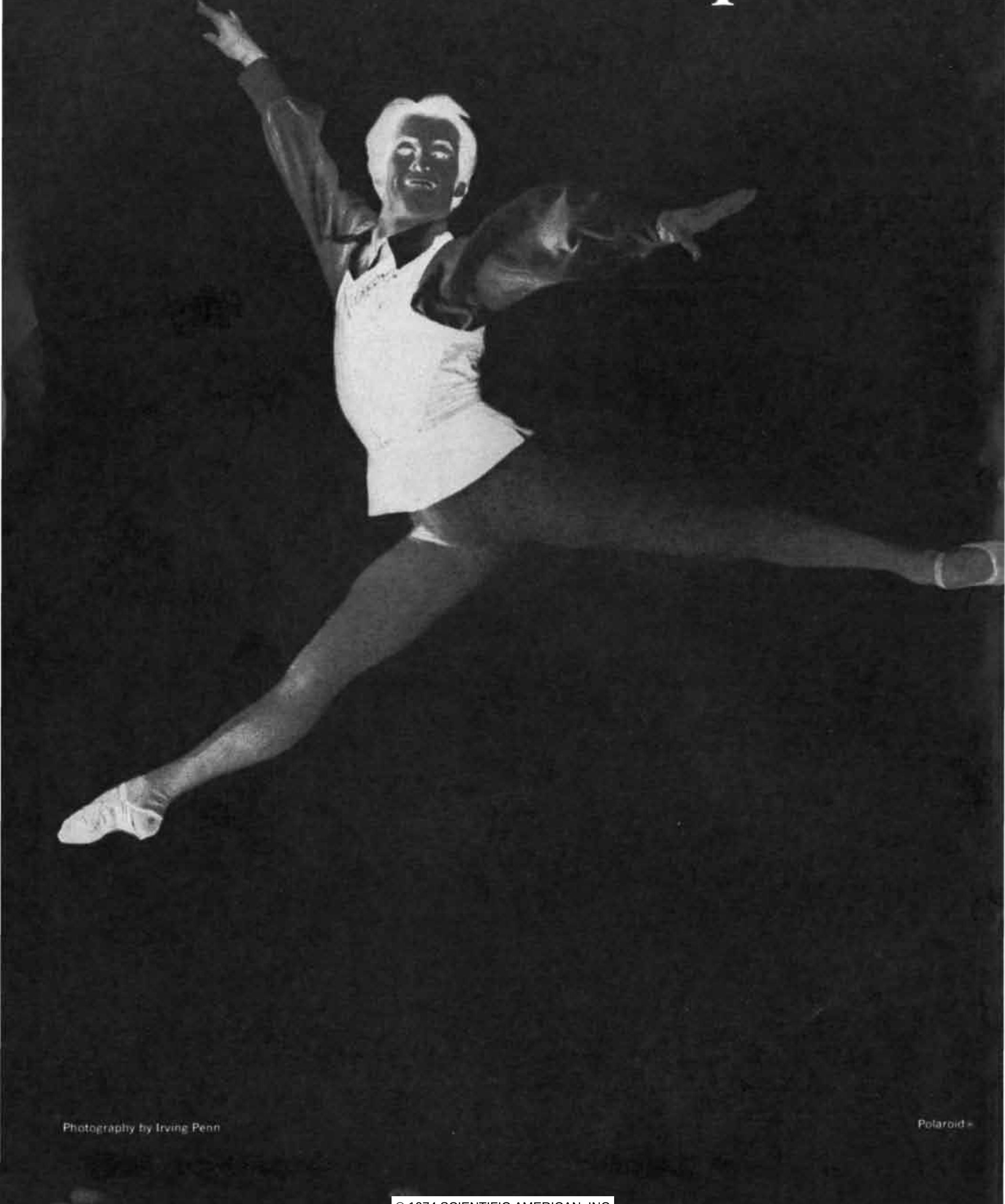
The oscillograph tracings clearly showed the highly structured periodic fluctuations in the intensity of the returning ray as the droplet evaporated. The signal was composed of a periodic set of humps with a series of sharp spikes on top of them; the time that elapses between repetitions of a certain spike structure is about three times longer than the time between adjacent humps.

Although it is not apparent from short sections of the traces, the oscillograph record covering the 20-minute lifetime of the droplet shows striking long-term changes in the basic three-hump structure. The spikes slowly change their magnitude and their position on the humps. A new spike rises to prominence at about every 72nd hump. If we number the humps in cycles of three, we find that the most prominent spike of all occurs on the same hump in the cycle every 73rd cycle. From this Fahlen and one of us (Bryant) concluded that the spike period is about 73/74 of the three-hump cycle.

What is the significance of these humps and spikes? The elapsed time between humps can be directly related to the change in the diameter of the droplet by monitoring this intensity of the axial ray. When the intensity goes from a maximum to a minimum and back to a maximum again, wave optics tells us that the diameter of the water droplet has changed by about three-eighths of the wavelength of the light. This fact enables us to determine that the hump period and the spike period respectively correspond to changes in the droplet diameter of .09 and .26 wavelength. The fluctuations in intensity therefore reflect very small changes indeed. For example, for the red laser beam, which has a wavelength of 630 nanometers, the hump period corresponds to a change in the droplet diameter of only .056 micrometer. For a droplet one millimeter in diameter that is a change of only 56 parts per million.

Needless to say, great pains are required to obtain a droplet quiet enough to yield these results, which indicates how complicated a complete explanation of the glory must be. Although the droplets studied in the laboratory have diameters that are 50 times greater than those of the droplets in the clouds where the glory occurs in nature, Mie calculations undertaken by Cox and Fahlen show

Polaroid announces a pack film



for instant prints and negatives.

A new film for cameras and instruments now using Polaroid 3¼"x 4¼" film.

Introducing Type 105.

A brand new Polaroid Land film for serious amateurs and professionals that delivers exceptionally high-quality prints and negatives. And delivers them economically and conveniently.

Just 30 seconds after you shoot with Type 105 you get a black and white print of exceptional tonal range, sharpness and luminosity.

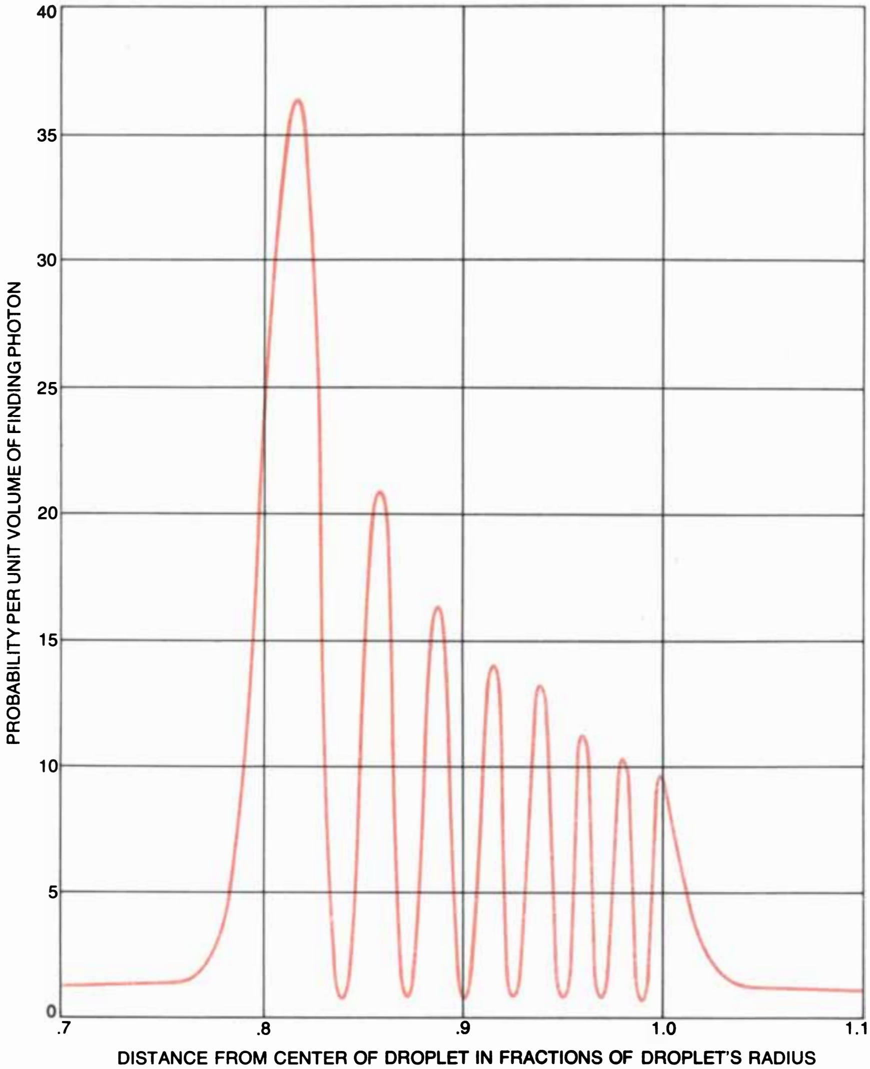
And a quality negative to match. With a resolution of over 150 lines/mm so enlargements retain original sharp detail.

What's more, this new positive/negative film comes in our convenient 3¼" x 4¼" eight-shot pack format. So it fits any instrument, camera, and Polaroid Land camera that already uses our Type 107 or Type 108 pack film.

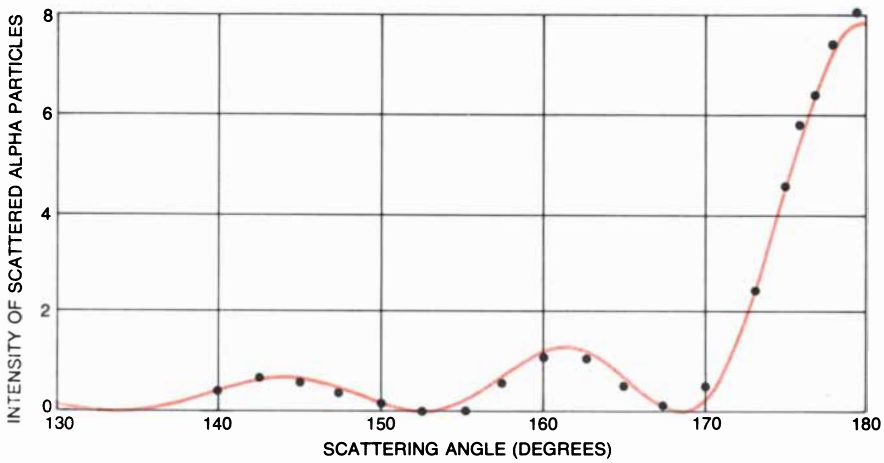
(For 4 x 5 cameras and instruments we're introducing a new adapter for all our 3¼" x 4¼" pack film called the Model 405 Land Pack Film Holder.)

Ask at your photo store for our new Type 105 Land film. A superb black and white film for instant prints and reprints.

Polaroid's instant positive/negative pack film.



PROBABILITY OF FINDING A PHOTON at a given distance from the center of a droplet shows a set of maximums and minimums that one would expect from rays traveling around the droplet at a resonance similar to that producing one of the narrower spikes seen in the oscillograph tracings in the bottom illustration on page 66. Example corresponds to a droplet with a diameter of 40.44 micrometers illuminated by light from a helium-neon laser.



SUBATOMIC PARTICLES PRODUCE A GLORY, as can be seen in this distribution of alpha particles (helium nuclei) at an energy of 29 million electron volts backscattered by nuclei of calcium 40. The solid line is the prediction based on a theory of the glory developed by the authors; the points are actual measurements made by A. Budzanowski and colleagues at the Institute of Nuclear Physics at Cracow in Poland. The maximums at the angles of 180 degrees, 162 degrees and 144 degrees are analogous respectively to the central bright region and the side rings of the optical glory. Intensity is given in arbitrary units.

that the behavior of light in the smaller droplets is much the same as it is in the larger ones.

Fahlen and one of us (Bryant) have been able to devise a simple mathematical model that can predict intensity fluctuations that are in qualitative agreement with experiment and with the exact predictions from the Mie theory. In principle the model includes an infinite series of paths through and around the surface of the sphere. H. M. Nussenzveig of the University of Rochester has developed a rigorous treatment of the scattering of waves by a sphere. Like the Mie theory his analysis consists of an infinite series of terms, but unlike the Mie theory each term can be directly interpreted as representing a light ray being internally reflected one, two, three or more times as it passes around and through the droplet. The humps observed in the oscillograph traces are principally the result of interference between a strong rainbowlike ray (a ray that is internally reflected only once) and a ray that is internally reflected 14 times, traveling around the droplet three and a half times. The spikes, being very narrow, must be the result of constructive interference among a large number of rays of similar intensities.

In master's theses at the University of New Mexico, Robert Thede and later Jaime Wong analyzed a certain class of spikes according to a description of light waves that neglects polarization effects. Their results can be interpreted to obtain the probability of finding a particle of light (a photon) at a given point in space. Thede and Wong found that there are several different types of spike that can be classified according to the probability of finding a photon at a given distance from the center of the water droplet. One such probability plot is shown in the top illustration at the left. We call it a Type 8 spike because it has eight maximums. Spikes of Type 6, Type 7 and Type 9 were found to be prominent for droplets with a diameter of 30 microns. The peaks and troughs in the probability curve can be interpreted in terms of waves reflecting around the inside of the droplet and interfering with one another constructively (the peaks) and destructively (the troughs). These waves correspond to light rays bouncing around inside the droplet at close to the critical angle.

In 1971 Lawrence Sromovsky of the University of Wisconsin, following on the work of Nussenzveig, showed in detail that the spikes are efficiently described by a mathematical concept close to the hearts of some elementary-particle theorists. The concept is that of the Regge pole, introduced by the Italian physicist Tullio Regge. The Regge pole is an abstract mathematical way of representing an entire class of physical situations. For example, all our Type 8 spikes can be described as originating from one Regge pole, so that instead of needing an infinite number of mathematical terms to describe the behavior of Type 8 resonances, one can make do with the Regge pole alone. Thus we find that an important component of the explanation of the glory is quite similar to one concept in the theory of elementary particles.

Since the wave-particle duality exhibited by light is shared by all elementary particles, beams of particles scattered in accelerator experiments may also show optical effects such as the rainbow and the glory. In fact, the concept of a surface wave in a water droplet can be applied to the behavior of high-energy particles bombarding an atomic nucleus. There is no direct correspondence, however, between the scattering of light by a transparent sphere and the scattering of particles by a nucleus. In the nuclear optical model the nucleus is rather like a muddy droplet with a poorly defined surface, so that surface waves would be strongly absorbed and not as clearly defined as they are in a droplet of clear water. Nevertheless, a simple extension of the glory mechanism serves very well in describing certain instances of the backscattering of high-energy particles by nuclei.

Like elementary particles and nuclei, entire atoms and molecules exhibit wave behavior when they are accelerated and scattered. In such experiments nothing comparable to the backward-scattered glory is observed, but wave effects resembling the rainbow have been. Moreover, there is a distinct interference pattern in the forward direction; it is often called the forward glory. Changes in such patterns produced by molecular beams have been utilized by chemists to investigate the energetic threshold of chemical reactions.

We hope that, having read this article, the reader who has never seen a glory will now be on the lookout for one. The easiest way to see a glory in this age of air travel is to watch for the small shadow of one's airplane on a layer of cloud below. If the conditions are right, the shadow will be surrounded by the bull's-eye pattern of the glory—a striking demonstration of the wave nature of light and a colorful reminder of the underlying unity of the physical world.

LEGAL NOTICE TO PURCHASERS OF SWIFT BRAND MICROSCOPES

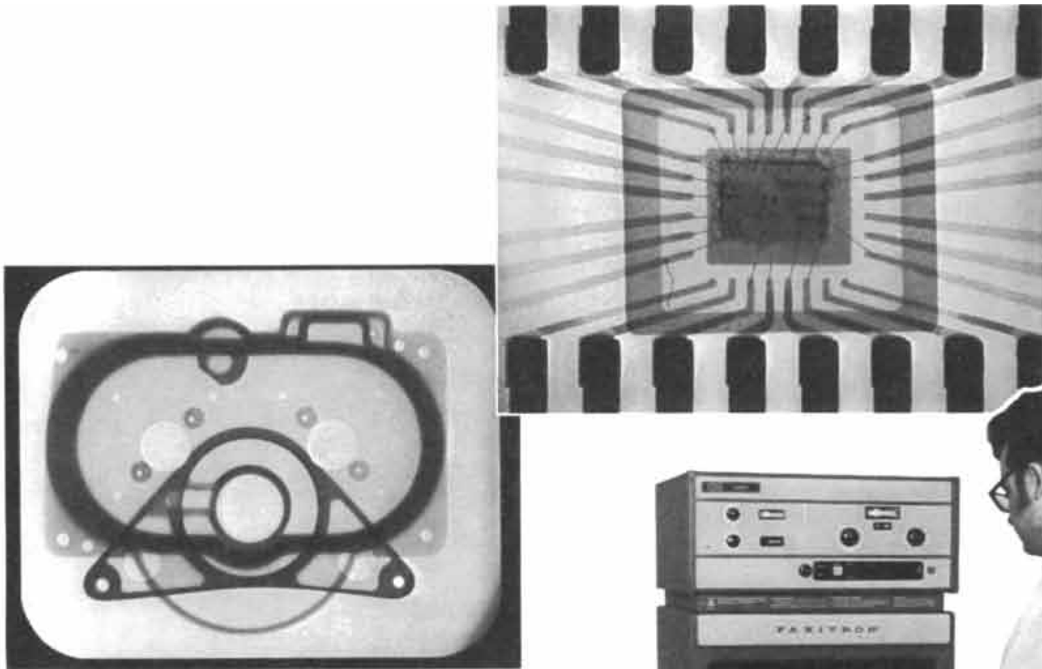
SETTLEMENT OF CLASS ACTION

There is now pending before the United States District Court in San Francisco a class action, Fremont Unified School District vs. Swift Instruments, Inc., seeking to recover damages for alleged violations of Section 1 of the Sherman Act. Plaintiff class consists of all retail purchasers of Swift Brand Microscopes, microscope parts and accessories in the United States, at any time during the period from February 1, 1969 through June 6, 1970, from Swift Instruments, Inc. and/or from any dealer (s) of Swift Microscopes.

Defendant has proposed to settle this litigation, without any admission of liability on terms which will result in total payments to the class in an amount estimated to exceed 9% of total purchases. *You may be a member of the class described above:* if so, your rights will be affected by this settlement. Each class member who desires to receive payment must submit a Statement of Claim no later than August 26, 1974. A hearing will be held before the Court at 2:00 P.M. September 16, 1974, to review and approve the settlement and if approved, to enter judgment dismissing said action.

The Court will, upon request to the undersigned Clerk, exclude any member of the class who so requests in writing on or before August 15, 1974. Any judgment entered pursuant to the settlement shall include all class members who do not request exclusion. Any class member who does not elect to be excluded from the class may but need not enter an appearance through his own counsel. Further details, including forms required for Statement of Claim are available upon written request to counsel for plaintiff class: Darrell Salomon, Esq., Law Offices of Joseph L. Alioto, 111 Sutter Street, San Francisco, California 94104.

s/s F. R. Pettigrew
Clerk
U.S. District Court
Northern District of California
450 Golden Gate Avenue
San Francisco, California 94102



The HP 43805 cabinet X-ray system gives you correctly exposed films automatically—like these radiographs of a defective printed circuit board and an aluminum casting.

X-ray inspection unravels structural mysteries routinely, right at your workbench.

Not only in life science and crime laboratories but in industrial workshops and labs of every kind, X-ray inspection can solve problems that defy solution by any other method. Often radiography is the only way to examine what's inside a sample without destroying it. To check the registration of a printed circuit board, for example. Or defects in a metal casting, the condition of an encapsulated component, the presence of microfossils in sedimentary rock, the morphology of an organ.

Many investigators dismiss radiography as a practical tool because it has traditionally required bulky, expensive cabinets and a specially shielded room. Although this was once an accurate assessment, radiography now is practical and convenient for any lab or workbench with the HP 43805 cabinet X-ray system.

The 43805 is as easy to operate as an office copier. Place your sample and film in the 43805 cabinet and close the door: the machine tells you what voltage to set for best picture contrast, automatically selects the correct exposure time, makes the exposure, and shuts itself off. You get a good radiograph on the first try, without wasting film and without extensive training.

The 43805 makes sharp prints of a variety of samples because it has a wide enough voltage range (10 to 130 kV) to assure excellent contrast with materials as thin as a postage stamp or as dense as a pistol. Use Polaroid® films for 15-second, self-developed prints; Kodak Industrex 600® paper film; standard film in sizes up to 14 by 17 inches; or ultra-high-resolution wet film for enlargement of micro-detail. For quick inspections, you can dispense with film entirely and use the optional fluoroscope and manipulator.

Most important, you can use the 43805 right in your lab or work area. Just plug it into a wall outlet. Shielded to exterior radiation levels below those specified by NBS Handbook 93 for use in populated areas, the 43805 also has double safety interlocks that prevent operation with the door open.

With prices starting at $3350,* the 43805 brings the benefits of instant X-ray inspection into laboratories that could not previously afford it, and enhances its feasibility for classroom use. The optional fluoroscope starts at $300,* and the manipulator at $150.*

Think what you can do with 4.8 million bytes of memory in a desk-top computing calculator.

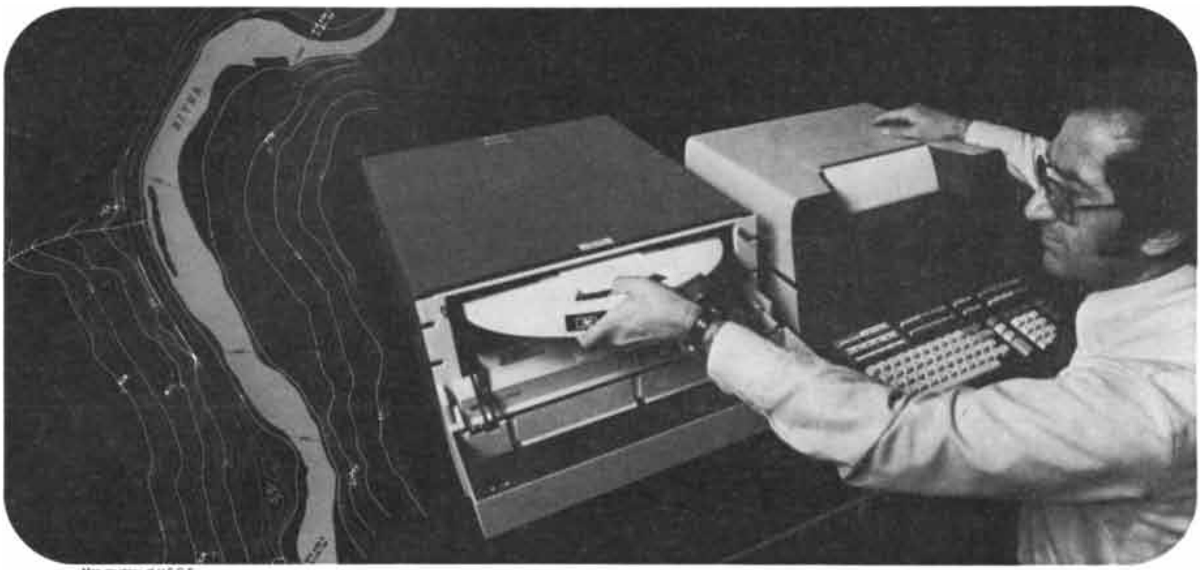
The new HP 9880B Mass Memory has been developed specifically for the HP 9830A Programmable Calculator, dramatically extending the calculator's storage and instruction capabilities.

For instance, its 4.8 million bytes of storage can accommodate large-data-base programs in statistical analysis, structural design, and inventory control —programs that previously required a sizeable computer system.

With the 9830 calculator and Mass Memory, you have enough storage capacity, for example, to store twenty or more years of rainfall and stream flow data for an entire river drainage system. And, through random access, the ability to retrieve any part of this data rapidly. Thus expected runoff and flood levels under varying conditions can be predicted, dam sites selected, and optimum dam sizes determined.

With a disc drive of one fixed and one removable platter, the Mass Memory can store 2.4 million 8-bit bytes on each platter—as much as can be stored on 38 magnetic tape cassettes. If you want even greater capacity, a second disc drive with two more platters can be added to the system. Because access is random, it is extremely fast. You can retrieve any 1,000-word program and load it into the 9830 in just one second. A comprehensive BASIC language command set provided with the Mass Memory lets you access files by name using simple commands such as "read" or "print."

A full line of peripherals is available for the Programmable Calculator system: card readers, line printers, plotters, digitizers, and tape punches. The lease price per month for the 9830A Calculator and 9880B Mass Memory starts at $890.*



Map courtesy of U.S.G.S.

For more information on these products write to us. Hewlett-Packard, 1502 Page Mill Road, Palo Alto, California 94304.

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.
Palo Alto, California 94304

*Domestic USA prices only. 00431

Mail to Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304
Please send me information on:
( ) HP 43805 Cabinet X-ray System
( ) HP 9830A Calculator
( ) HP 9880B Mass Memory

Name ______ Title ______
Company ______
Address ______
City ______ State ______ Zip ______

A FAMILY OF PROTEIN-CUTTING PROTEINS

The serine proteases act as catalysts in numerous processes, from the digestion of food to the clotting of blood. They have a common mode of action and apparently evolved from a common ancestor

by Robert M. Stroud

Almost all the chemical processes in a living organism are controlled by the catalytic proteins called enzymes. They are the three-dimensional jigs and tools of biochemistry, with which molecules are assembled, transformed and destroyed. Each enzyme acts by combining with the molecules involved in a particular reaction (the substrates of the reaction) and speeding up the making or breaking of a specific covalent chemical bond. Most of the bonds in biological molecules are very stable and rarely break by chance; consequently without enzymes few biological molecules would react with others at physiological temperatures. Enzymes catalytically promote reactions between molecules; like all other catalysts, they are never consumed in the course of the reaction.

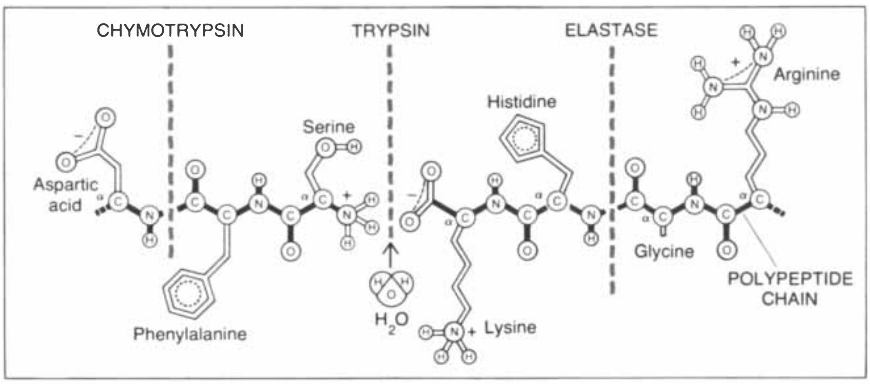
The proteolytic, or protein-cutting, enzymes constitute a large group [see "Protein-digesting Enzymes," by Hans Neurath; SCIENTIFIC AMERICAN, December, 1964]. They are proteins whose function is to alter or decompose other proteins by splitting them into fragments. Within this group the serine proteases form an important family whose members are essential to a variety of biological activities. Serine proteases participate in digestion, in the formation and dissolution of blood clots, in the immune reaction to foreign cells and organisms, in the fertilization of the ovum by the spermatozoon. Even though their physiological functions are diverse, however, they all seem to employ the same catalytic mechanism to promote the same chemical process: the cleavage of a particular kind of chemical bond common to all proteins.

The three-dimensional structure of some of the serine proteases—in particular those of the digestive enzymes—has been determined in atomic detail. The analysis of structure not only reveals what the molecule "looks like" but also suggests how it might work. Any proposed model of the catalytic mechanism must be consistent with the structure of the enzyme and the substrates.

Knowledge of structure has provided another insight into the nature of the serine proteases. It reveals that most of these enzymes are related, that they evolved from a common ancestor. It appears that a single mechanism for protein cutting, which may have emerged early in the history of life, was conserved during subsequent evolution. Moreover, the same mechanism was applied to a variety of tasks, some of which must be far removed from the task of the ancestral enzyme.

The bond cut by the serine proteases is the one that joins amino acids together to form proteins. Each amino acid consists of an amino group (NH_2) and a carboxyl group (COOH) attached to a single carbon atom, the alpha carbon. To the alpha carbon are also attached a hydrogen atom and one of some 20 side chains, by which the amino acid is identified. In the synthesis of proteins the carboxyl group of one amino acid is connected to the amino group of the next, by the extraction of a water molecule, to form the linkage —CO—NH—. This bond is called a peptide bond; a protein is thus a polypeptide. Each protein is characterized by a unique sequence of amino acids, specified in the genetic material of the cell. That sequence in turn determines the architecture and function of the molecule.

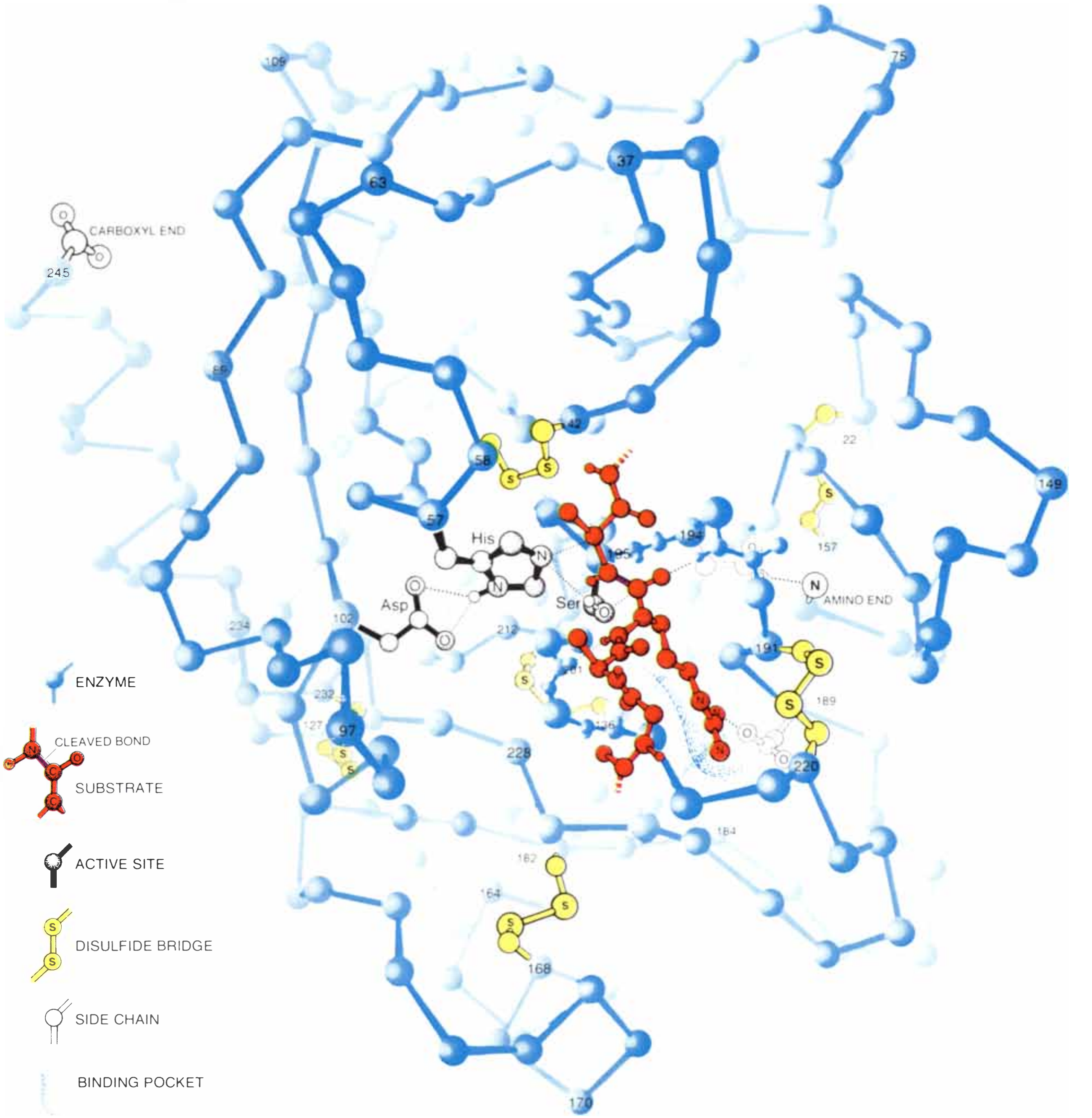
The process of protein decomposition mediated by the serine proteases is the reverse of the process followed in protein synthesis. A water molecule (a hydrogen atom and a hydroxyl group) is added for



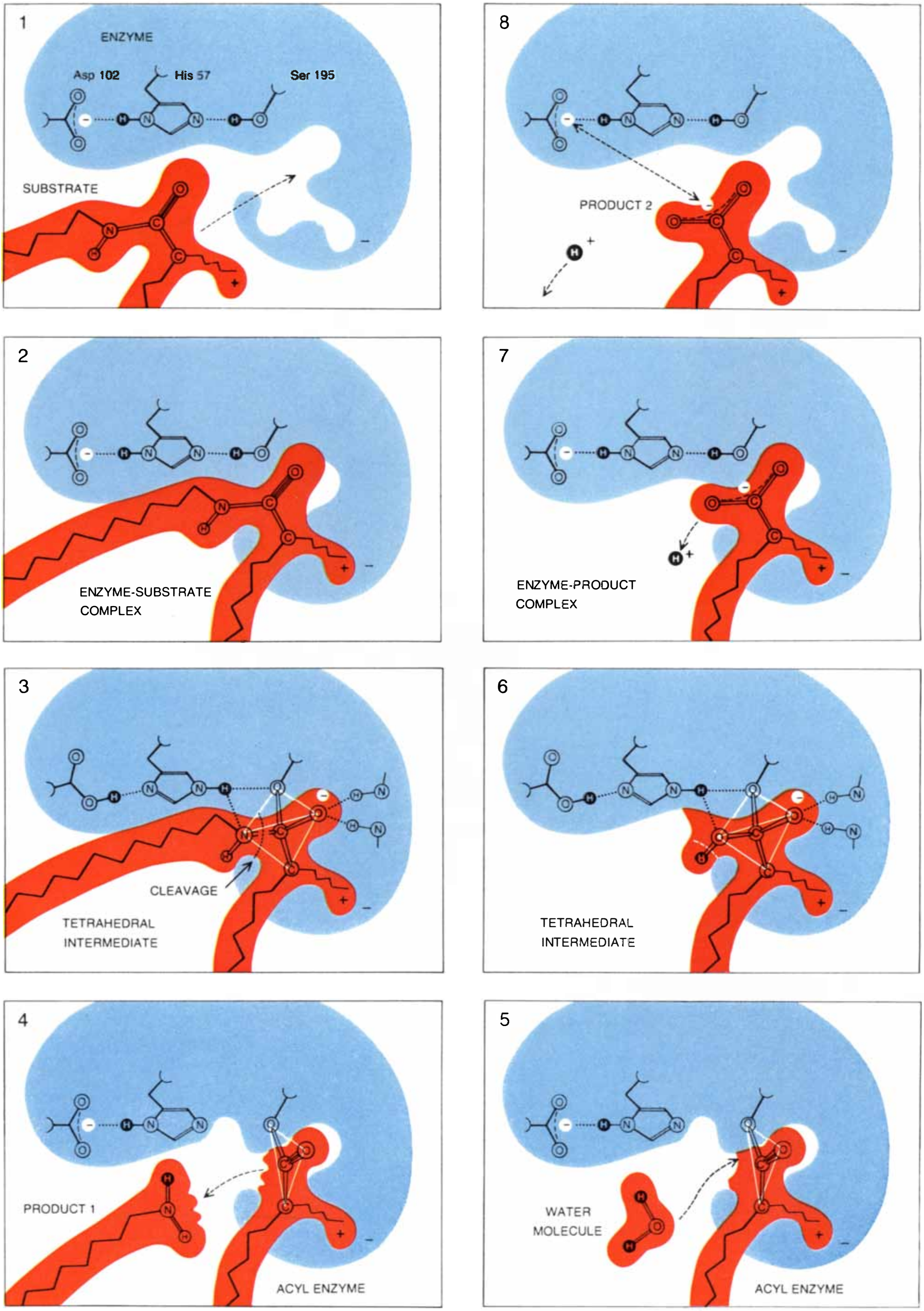
POLYPEPTIDE CHAIN is cleaved at a different position by each of the three serine proteases involved in digestion. In this diagram the peptide bonds, which join amino acid units, are those to the right of the nitrogen atoms (*N*). Not all the peptide bonds, however, are susceptible to attack by the serine proteases. Chymotrypsin breaks only those that (*reading from right to left*) follow large, hydrophobic side chains, such as phenylalanine. Trypsin cuts peptide bonds that follow either of two positively charged side chains, lysine (*shown at the cleavage site*) and arginine (*shown at another position on this hypothetical peptide*). Elastase attacks bonds that lie next to very small side chains, such as the side chain of glycine. The cleavage of each bond is achieved by the addition of a water molecule.

each peptide bond broken, restoring the amino and carboxyl groups at the site of the cleavage to their free amino acid form. Each of the fragments produced by the cutting of the protein is thus a complete peptide, with an amino terminus and a carboxyl terminus. Because the process involves the addition of a water molecule it is termed hydrolysis [*see illustration on opposite page*].

Under physiological conditions the hydrolysis of peptide bonds will proceed in the absence of enzymes, but only at an exceedingly low rate. (If the rate were not low, proteins would spontaneously disintegrate.) The effect of an enzyme is to greatly speed the reaction; moreover, it favors a particular reaction over all others involving the same substrates. In this way enzymes steer the sequence of reactions along specified pathways



TRYPSIN AND A SUBSTRATE fit together like pieces of a three-dimensional jigsaw puzzle. The substrate is held in place by its correspondence to the shape of the enzyme and stabilized in position by several hydrogen bonds (*dotted lines*). Of particular importance in determining the orientation of the substrate is the "binding pocket," which receives a substrate side chain of lysine or arginine. The positively charged groups at the end of these side chains interact electrostatically with the negatively charged carboxyl group of unit 189 in the enzyme; it is because of the form and electrical environment of the binding pocket that trypsin cleaves only those peptide bonds that are adjacent to these side chains. Except at the active site, this schematic representation of the trypsin molecule shows only the skeleton of alpha carbon atoms. The active-site side chains (serine 195, histidine 57 and aspartic acid 102) act together to cleave the susceptible bond of the substrate. The enzyme consists of 223 amino acid units, numbered by analogy to the sequence of chymotrypsin. It is folded to form a three-dimensional structure stabilized by six disulfide bonds. The structure was determined by the author, Lois Kay and Richard E. Dickerson; the substrate-binding model was developed by the author and Monty Krieger.



1
ENZYME
Asp 102
His 57
Ser 195
SUBSTRATE
2
ENZYME-SUBSTRATE COMPLEX
3
CLEAVAGE
TETRAHEDRAL INTERMEDIATE
4
PRODUCT 1
ACYL ENZYME
5
WATER MOLECULE
ACYL ENZYME
6
TETRAHEDRAL INTERMEDIATE
7
ENZYME-PRODUCT COMPLEX
8
PRODUCT 2

toward products employed in subsequent biochemical processes.

Energy must be supplied to join amino acids together, whereas energy is liberated in the hydrolysis of peptide bonds. The overall change in energy indicates that the hydrolyzed peptide is more stable, and also suggests the second important function of enzymes. In the absence of enzymes the hydrolysis of a peptide bond in a neutral solution is very slow because there are high energy barriers between the starting materials and the products. At physiological temperatures the reactants only rarely attain the high internal energy required for the hydrolysis to proceed. Enzymes allow the reaction to follow a different pathway from the starting materials to the products, and thereby reduce the energy barriers. New intermediate states are established in the course of the reaction; the enzyme stabilizes the states of highest energy and thereby lowers the internal energy barriers: the high-energy transitions between one intermediate and the next [*see illustration on next page*].

The ability to make or acquire amino acids and to link them together in the correct sequence to form proteins is one of the fundamental talents of the living cell. Some single-celled organisms, such as bacteria, can synthesize all the required amino acids from simple nutrients. Higher animals, however, have lost the ability to make some of these amino acids and must obtain them from their food supply. Man, for example, requires about a gram a day of eight amino acids. They are obtained by decomposing dietary proteins. Proteolytic enzymes are essential to this process; without them it would take 50 years to digest a meal.

The digestive enzymes are among the most thoroughly studied of all enzymes, principally because they are easily isolated and purified. Three of them—trypsin, chymotrypsin and elastase—are serine proteases. They originate in the pancreas as proenzymes, or inactive precursors, which are secreted into the duodenum. There they are activated and catalyze the breakdown of proteins into fragmentary peptides; in concert with other enzymes the serine proteases ultimately reduce the peptides to individual amino acids, which are absorbed by the intestine and transported to the sites of protein synthesis.

The digestive enzymes act cooperatively, attacking proteins at different positions. For example, the enzymes of one class, the carboxypeptidases, release free amino acids one by one from the carboxyl end of the polypeptide chain. Trypsin, chymotrypsin and elastase, on the other hand, generally act on bonds in the middle of the chain. Of the three trypsin is the most selective; it cleaves only those peptide bonds adjacent to the amino acid units lysine or arginine, which are relatively large, carry a positive charge and are hydrophilic (have an affinity for water). Chymotrypsin is somewhat less specific; it hydrolyzes the bonds adjacent to any of several amino acids that are large but hydrophobic (repel water). Elastase acts on bonds adjacent to glycine, alanine or serine, which are the smallest of the amino acids found in nature. Although the three enzymes cannot cut all peptide bonds, they form a team that can reduce any food protein to small, soluble fragments.

The three-dimensional structure of all three of these enzymes has now been determined. In this procedure the sequence of amino acid units is studied first, utilizing methods developed in the 1950's by Frederick Sanger of the University of Cambridge. The protein is repeatedly broken into many fragments of different length; then the fragments are chemically tagged and analyzed. The manner in which this linear chain of amino acids folds up into a more or less globular three-dimensional structure is determined from the diffraction patterns produced by passing X rays through a crystallized sample of the protein; the technique was developed mainly by Sir Lawrence Bragg, J. D. Bernal and Max Perutz [see "X-Ray Crystallography," by Sir Lawrence Bragg; SCIENTIFIC AMERICAN, July, 1968].

The first of the serine proteases for which a three-dimensional structure was known was chymotrypsin; the structure was determined by David Blow and his colleagues at Cambridge and later, in a different form, by David Davies of the National Institute of Arthritis and Metabolic Diseases [*see illustration on page 80*]. The structure of elastase was solved by Herman Watson and David Shotton at Cambridge (who were later at the University of Bristol) and by Brian Hartley at Cambridge [*see illustration on page 81*].

In 1968 I joined Richard E. Dickerson's group at the California Institute of Technology and with Lois Kay began to study the structure of trypsin. Of the three enzymes trypsin is the most difficult to work with because it is the most active and within minutes it is damaged by autolysis: the destruction of an enzyme molecule by another molecule of the same enzyme. The activity of the enzyme varies with *p*H, and the problem of autolysis can be eliminated by studying the enzyme in an acid environment. We wished to study a catalytically potent molecule, however, and it was therefore necessary to carry out our investigations where the enzyme is most active, at *p*H 7 to 8. We achieved this by employing a method in which the enzyme is activated only as it is crystallizing. By 1970 we had built a model of the trypsin molecule by fitting the results of our X-ray-crystallography studies to the amino acid sequence that had been determined by Hans Neurath of the University of Washington and Frantisek Sorm of Charles University in Czechoslovakia. It was rather like solving a three-dimensional jigsaw puzzle.

One of the objects of these investigations was to determine how the substrate is bound to the enzyme. A number of clues were already available. More than a decade ago E. F. Jansen, A. K. Balls and M. D. F. Nutting of Purdue University discovered that blocking a particular serine unit in chymotrypsin inactivates the enzyme. It is because of this crucial serine that the family of enzymes is called serine proteases. Subsequent investigation showed that a his-

CATALYTIC MECHANISM of trypsin and other serine proteases accelerates the hydrolysis of peptide bonds by providing intermediate states for the reaction and by smoothing the transition from one intermediate to the next. Enzyme and substrate must first come together (*1*) to form a precisely oriented complex (*2*). The oxygen atom of serine 195 then bonds covalently to the substrate carbon, forming a tetrahedral intermediate (*3*); the proton, or hydrogen ion, from serine 195 is transferred to the substrate nitrogen. The proton transfer breaks the peptide bond, and the first product is liberated (*4*). The remaining complex is called an acyl enzyme; it breaks down to regenerate free enzyme in steps that are symmetrical with those of the first half of the process. A water molecule enters the reaction (*5*), its hydroxyl group forming with the substrate another tetrahedral intermediate (*6*). The hydrogen ion from the water molecule is transferred to serine 195, breaking the covalent bond between enzyme and substrate (*7*). The second product is now freed (*8*); its departure is hastened by repulsion between negatively charged carboxyl groups of product and aspartic acid 102. Histidine 57 and aspartic acid 102 participate in the proton transfers.

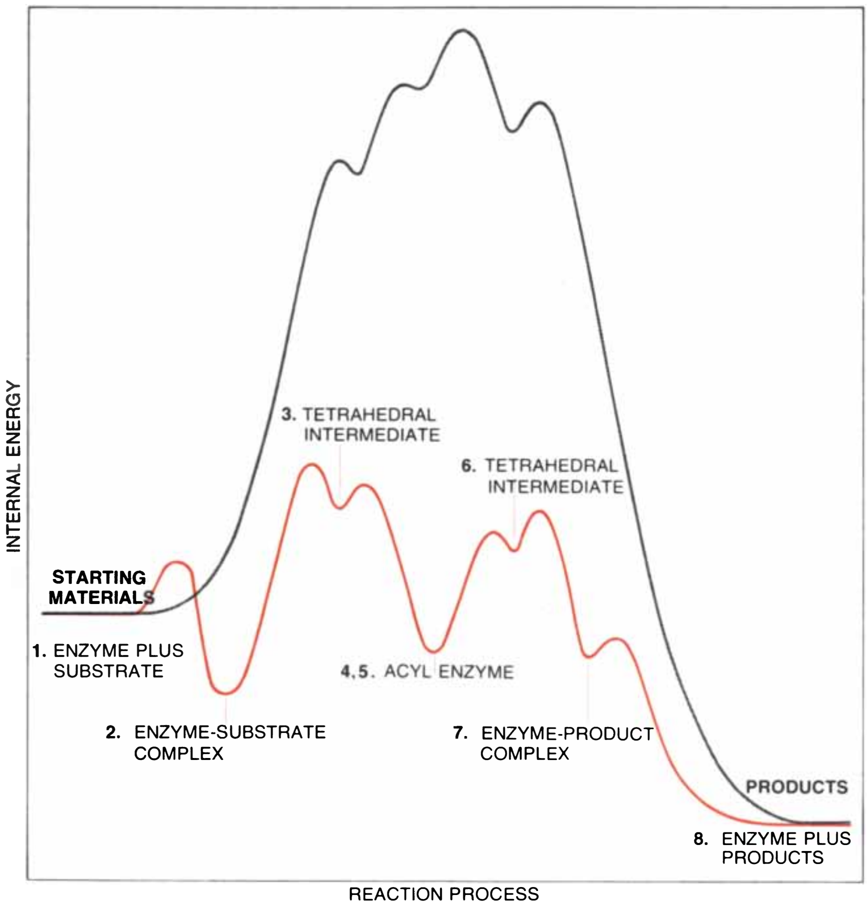
tidine unit elsewhere in the molecule is equally crucial to its catalytic activity; furthermore, the same two amino acids were shown to be essential in trypsin and elastase as well as in chymotrypsin. It is now known that the immediate neighbors of these amino acids are also invariable in all the enzymes. In a standard numbering system that relates the amino acid sequences of the other enzymes to the sequence of chymotrypsin, the essential amino acid units are serine 195 and histidine 57. The enzymes are inactivated if either of these units is chemically altered in any way.

These discoveries were made before the three-dimensional form of any of the enzymes was known. When the structure of chymotrypsin had been determined, it was found that an aspartic acid unit, numbered 102, is also involved in the enzyme's catalytic activity. These three units, although widely separated in the linear sequence of amino acids, are brought close together by the folding of the polypeptide chain. They are at the active site: the region where the enzyme and its substrate are bound to each other and where the catalyzed reaction takes place.

Although chemical methods alone had suggested the probable location of the active site, only the structure of the molecule could explain the specificity of the enzymes. It was apparent that all of them bring the same chemical groups to bear on the substrate; it was also evident that some mechanism capable of discriminating between the side chains of the amino acids of the substrate must be incorporated in the enzyme molecule at or near the active site. Such a mechanism emerged when the structure of chymotrypsin was solved.

Near the active site of chymotrypsin is a "pocket" in the surface of the molecule whose interior environment is hydrophobic. Chymotrypsin is specific to peptide bonds adjacent to large hydrophobic side chains, and Blow and his colleagues proposed that this pocket could act as a receptor for the side chain and help to orient the substrate against the enzyme. The hypothesis was supported by their discovery that *N*-formyl-l-tryptophan, a small molecule very much like one of the amino acids for which chymotrypsin is specific, binds to the hydrophobic pocket.

The same recognition mechanism is found in the other digestive serine proteases. In elastase the pocket is filled by amino acid side chains not present in chymotrypsin, and it is blocked to all but the smallest side chains on the substrate. In trypsin the pocket is not blocked but has at its back an aspartic acid unit, which carries a negative charge. Thus positively charged lysine and arginine side chains are held in the pocket by electrostatic forces.

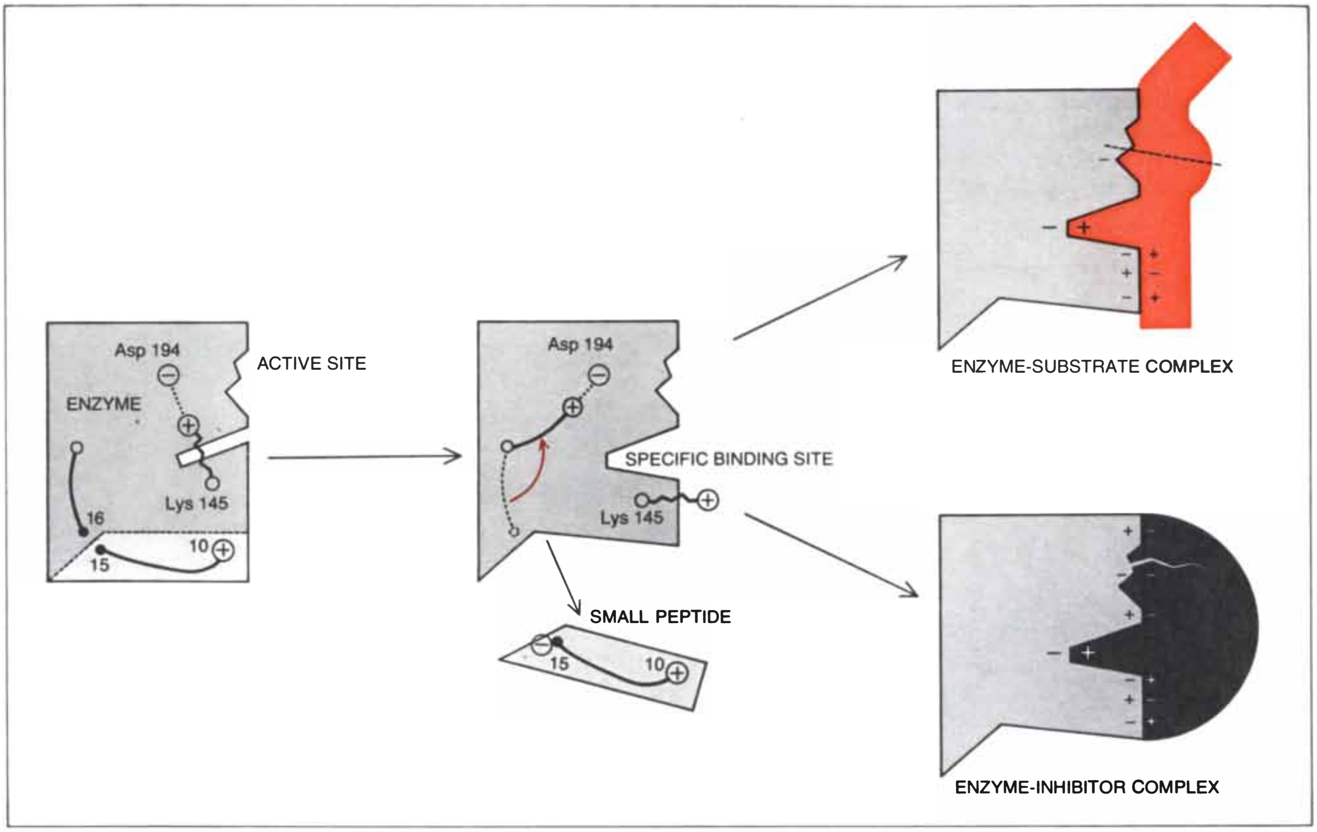


ENERGY ANALYSIS of the hydrolysis of a peptide bond shows that the overall reaction proceeds with net release of energy, since the internal energy of the products is less than that of the reactants. Under physiological conditions, however, peptide bonds are rarely broken in the absence of enzymes because a large amount of energy must be added in order to initiate the reaction (*black curve*). Enzymes such as the serine proteases reduce this energy barrier by providing an alternate pathway for the reaction through intermediate products of lower internal energy (*colored curve*). The labeled intermediates of the enzymatic hydrolysis correspond to the intermediates depicted in the illustration on page 76.

Although the pocket for specific side chains is obviously important in binding, it is not the only binding site. Enzyme and substrate must be secured at several points, so that the susceptible bond is oriented at the active site in precisely the right configuration. The ideal way to investigate the spatial relations of the enzyme and the substrate would be to crystallize a specimen of the enzyme with the substrate bound and examine it by X-ray crystallography. This is not possible. Once the enzyme-plus-substrate complex is formed, the catalyzed reaction takes place almost immediately (within a hundredth of a second), and the system becomes an enzyme-plus-product complex. This too breaks down quickly to yield free enzyme and products, but even if it were possible to examine the enzyme-plus-product complex, one would have arrived too late for the crucial event.

The best explanation of how trypsin binds to a substrate comes from studies of the structure of one of the naturally occurring trypsin inhibitors, a small protein of molecular weight 6,000. The inhibitor binds extremely tightly to trypsin, in almost the same manner as a substrate, but it is not cleaved and dispatched as a normal substrate would be. Instead it blocks the active site and so inactivates the enzyme.

Protease inhibitors such as this one form a part of the control or defense mechanisms that are crucial to the well-being of the organism, and in the pancreas trypsin inhibitors are particularly important. Trypsin activates the proenzymes of chymotrypsin and elastase as



ACTIVATION AND INHIBITION are means of regulating the serine proteases. The enzymes are produced as precursors, or proenzymes, with an active site but without a functional binding pocket. In the case of trypsin, which is diagrammed here, the enzyme is activated by the cutting of a single bond (between amino acid units 15 and 16 in the standard numbering system). The cutting liberates a small peptide. The new amino terminal at unit 16 then interacts ionically with the aspartic acid unit 194, altering the orientation of the lysine 145 unit and inducing a structural change that generates the specific binding site. Once activated, the enzyme can bind to and cleave a substrate (*top right*), or it may encounter an inhibitor molecule (*bottom right*). Inhibitors resemble substrates, but they bind to the enzyme without being cleaved. The result is that they inactivate the enzyme molecule. Generally the initial stages of hydrolysis do proceed, but the reaction is halted in an intermediate state, and both of the products remain bound to the enzyme.

well as the proenzyme of trypsin itself; thus if one molecule of trypsin were prematurely activated in the pancreas, it would be sufficient to start an autocatalytic chain reaction. The enzymes would destroy any protein within reach, including themselves. Such catastrophes are averted by the presence in the pancreas of inhibitor molecules, which must have evolved simultaneously with the enzymes they inactivate. The system of inhibitors specifically blocks active trypsin, but not its precursor proenzyme.

Robert Huber's group at the Max Planck Institute for Biochemistry in Munich determined the three-dimensional structure of a bovine pancreatic inhibitor in 1971. Chemical studies had already shown what part of the inhibitor molecule interacts with trypsin, so that we were able to test models of the two molecules against each other. They fit together beautifully.

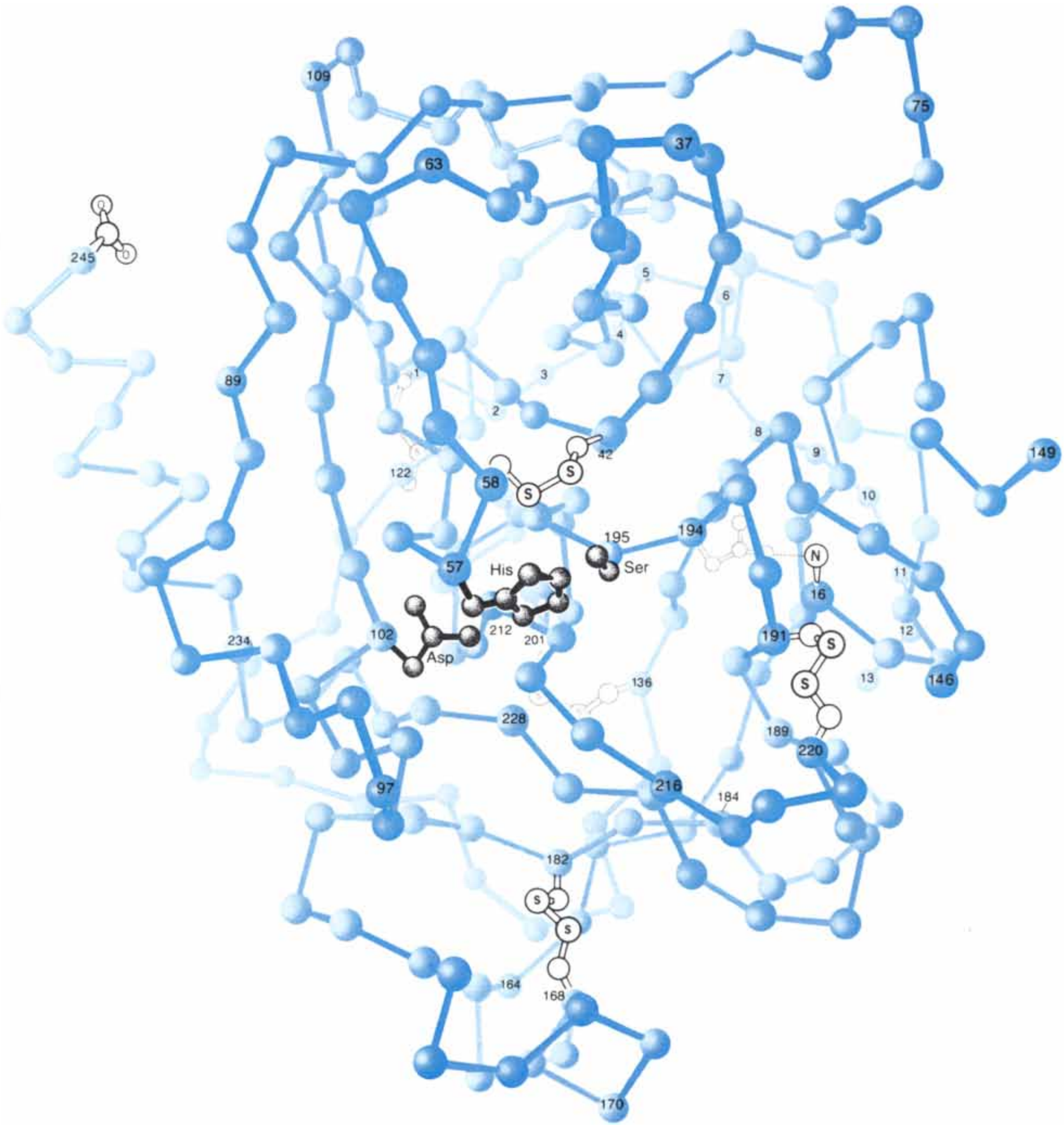
Trypsin normally operates on denatured or partly unraveled proteins that can be oriented on the enzyme surface. The inhibitor has evolved a configuration in the binding region that closely resembles that of a bound substrate, which is part of the reason it binds to the enzyme so tightly.

From this model and from X-ray studies in which a small molecule resembling the side chain of arginine was bound to trypsin, Monty Krieger and I developed a model for substrate binding to trypsin [*see illustration on page 75*]. Several hydrogen bonds and a hand-in-glove geometrical association, which includes the insertion of the side chain into the binding pocket, orient the susceptible peptide bond close to the active site. Huber and Blow independently developed a similar model for the binding of the pancreatic inhibitor to chymotrypsin. Since then both Huber and Blow have also solved the structure of a trypsin-plus-inhibitor complex; these complexes suggest the same mode of enzyme-substrate binding. This model probably provides the best view we shall ever have of how an extended substrate binds to trypsin. Together with the vast knowledge of enzyme chemistry available, the model is close enough to reality to suggest ways in which chemical groups on the enzyme catalyze the hydrolysis of the peptide bond [*see illustration on page 76*].

The enzyme serves first as a template on which the substrate is bound to form the precisely oriented enzyme-plus-substrate complex. The –CO– and –NH– groups that are joined to form the peptide bond lie in a single plane at the active site, with the arginine or lysine side chain in the enzyme binding pocket.

The hydrolysis of the bond begins when the hydroxyl group of serine 195 pivots to approach and attack the carbonyl carbon of the substrate (the carbon of the –CO– group). At the same time the hydrogen ion, or proton, of the hydroxyl group is transferred to the nearby histidine 57, and the serine hydroxyl oxygen forms a covalent bond with the carbonyl carbon. As a result the double bond connecting the carbon and oxygen atoms in the substrate carbonyl group is transformed into a single bond; the oxygen assumes a negative charge and the carbon bonds change from a planar to a tetrahedral configuration. The tetrahedral form is a short-lived intermediate, but it is stabilized both by the covalent bond to the serine oxygen and by hydrogen bonds formed between the negatively charged substrate oxygen and another region of the enzyme. Both Richard Henderson at Cambridge and Joseph Kraut (together with his co-workers) at the University of California at San Diego have emphasized the importance of these hydrogen bonds, which help to speed up the reaction by reducing the highest energy barriers between intermediate states.

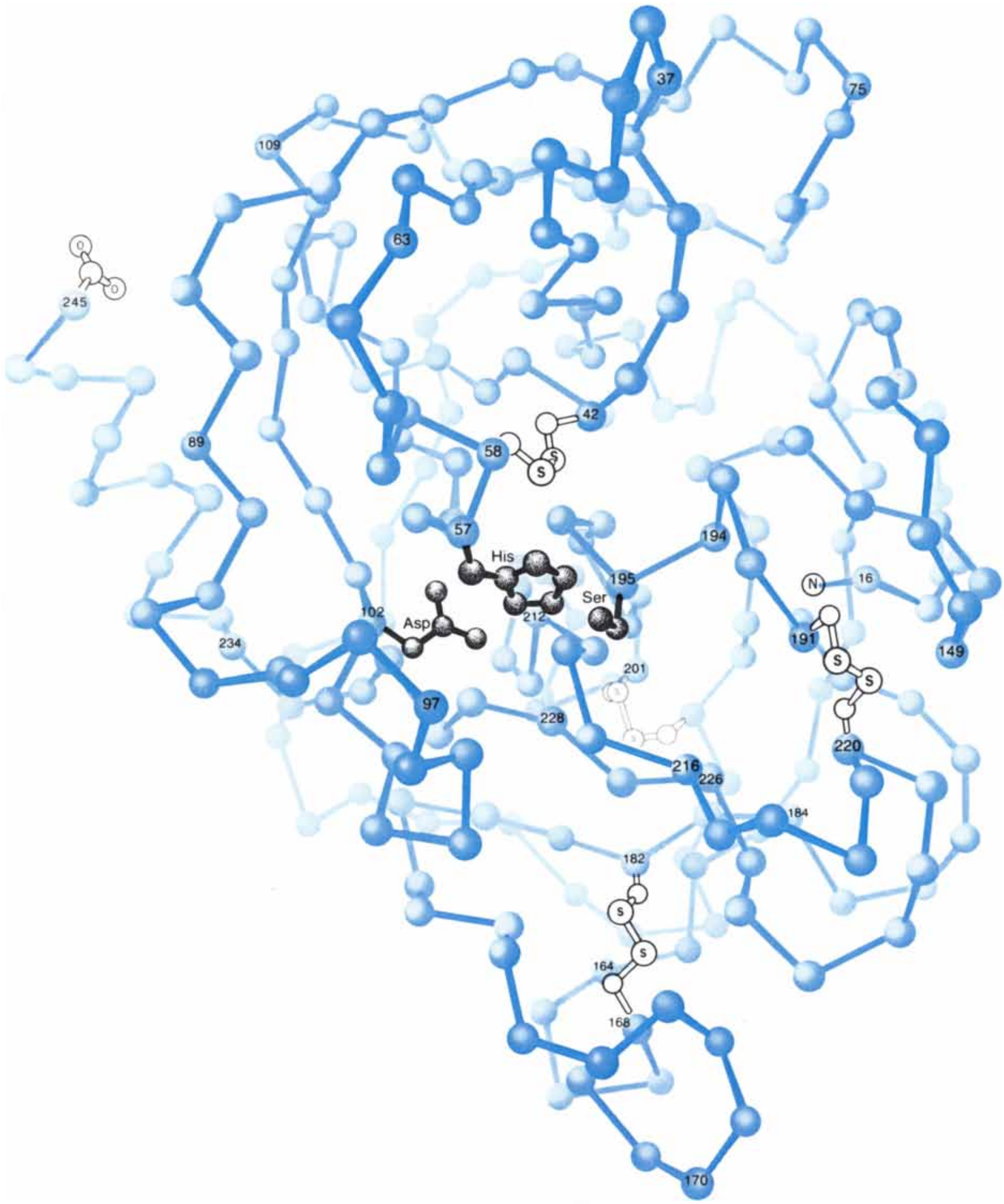
The proton delivered to histidine 57 remains there only briefly before it is



CHYMOTRYPSIN folds to form a molecule with architecture very similar to that of trypsin, providing strong evidence of kinship. The three amino acid units at the active site, serine 195, histidine 57 and aspartic acid 102 (*black*), are identically positioned in the two enzymes. The binding pocket of chymotrypsin, below and to the right of serine 195, accepts large, hydrophobic side chains. In this representation only the alpha carbon atoms of the polypeptide chain are shown. Those numbered 14 and 15 are excised when the enzyme is activated. The chymotrypsin structure was determined by David Blow and his colleagues at the University of Cambridge.

transferred to the –NH– group in the substrate on the other side of the peptide bond that is being broken. The histidine greatly facilitates this process, yet even so the proton transfer is a relatively slow step in the hydrolysis.

John H. Richards, Michael Hunkapillar and Stephen Smallcombe of the California Institute of Technology recently showed histidine 57 in trypsin is neutral, or uncharged, and therefore it cannot maintain the positive charge brought to it by the proton even for a short time. Consequently another proton "recoils" from the other side of the histidine and is accepted by aspartic acid 102, the third of the three catalytically important amino acid units at the active site. Within the environment of the enzyme under physiological conditions aspartic acid 102 ordinarily has a negative charge; this



ELASTASE, another relative of trypsin, has the same active site as that enzyme and as chymotrypsin, and is similar in overall structure. The binding pocket, however, is filled by side chains (*not shown*) attached to amino acid units 216 and 226; as a result elastase cleaves only bonds adjacent to very small side chains. The similarity of trypsin, chymotrypsin and elastase in function and in three-dimensional structure is achieved in spite of extensive differences in amino acid sequence; only 24 percent of the units are identical in all three enzymes. Elastase structure was worked out by Herman Watson and his colleagues at the University of Bristol.

charge is neutralized by the proton transferred from histidine 57 just as a negative charge is simultaneously developed on the carbonyl oxygen of the tetrahedral intermediate. At no stage in the process is it necessary to "pull" electrons from one place, making it positive, to generate negative charge elsewhere. As in a dynamo, that would require extra energy.

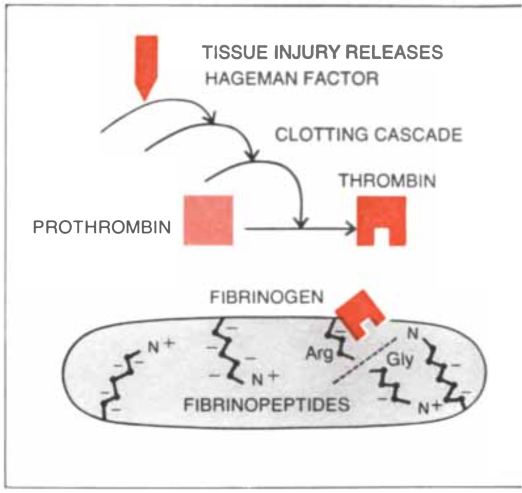
When the nitrogen atom in the substrate receives the proton from histidine 57, the peptide bond between the nitrogen and the carbonyl carbon becomes unstable and breaks. The first product of hydrolysis—the portion of the protein on the amino side of the cleaved bond—is free to diffuse away, taking with it a proton from the enzyme. At the same time the part of the substrate that remains bound to the enzyme is rearranged to form an acyl intermediate. (Acyl refers to an organic acid in which the hydroxyl portion of the carboxyl group is lacking.) The carbonyl carbon reverts to the planar configuration and the hydrogen bonds to the carbonyl oxygen are probably broken, but the covalent bond to the serine oxygen remains intact.

The breakdown of the acyl intermediate to regenerate native enzyme takes place in steps that retrace in reverse order those of the preceding process. In this process, however, a water molecule, drawn from the surrounding aqueous environment, enters the reaction. The water molecule's hydroxyl group binds to the substrate carbon, once again forming a tetrahedral intermediate, which is, as before, stabilized by hydrogen bonds to a negatively charged oxygen atom. At the same time histidine 57 accepts the hydrogen ion from the water molecule and donates the proton from the other side of its side chain to aspartic acid 102. Finally, the second product of the reaction is formed when the covalent bond between the substrate carbon and the serine oxygen is broken. The cleavage of this bond results in the formation of a carboxyl group on the remaining fragment of the substrate and releases the fragment from its attachment to the enzyme. At the same time histidine 57 transfers the proton it has just accepted to serine 195, restoring the serine hydroxyl group, and aspartic acid 102 once again gains its negative charge. When the second product reenters the aqueous medium, its carboxyl group is ionized, leaving the peptide with a negative charge; C. H. Johnson and Jeremy Knowles of the University of Oxford have suggested that the repulsion between this charge and that of aspartic acid 102 helps to push the second product away from the enzyme. The trypsin molecule has now resumed its original state and is free to hydrolyze another substrate molecule.

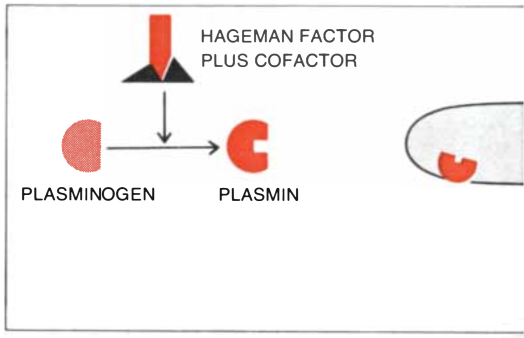
One prerequisite to catalysis implied by this model is that aspartic acid 102 be without its carboxyl proton and therefore negatively charged. If the *p*H is lowered to below neutrality (*p*H 7), a process that adds protons to the solution, the enzyme is temporarily inactivated. It would have been reasonable to assume that when the *p*H is lowered, a proton would attach first to histidine 57, giving it a positive charge; Richards and his colleagues have shown, however, that histidine 57 remains electrically neutral down to *p*H 4, and they have inferred that when the *p*H is lowered below 7, it is aspartic acid 102 that takes up the proton, neutralizing its charge. By a different method Roger E. Koeppe II and I have since shown directly that one of the two aspartic acid units in the interior of the trypsin molecule (almost certainly aspartic acid 102) is the recipient of the proton. That is not the normal behavior of aspartic acid and histidine in solution; it is a result of the environment provided by the enzyme structure. Richards and his colleagues have emphasized that it is only because of these unusual electronic states that histidine 57 and aspartic acid 102 are able to efficiently transfer protons back and forth between the enzyme and the substrate; this efficient shuttle represents one of the important differences between the enzyme-mediated reaction and the nonenzymatic hydrolysis of proteins. A relay of protons between serine 195, histidine 57 and aspartic acid 102 was first proposed by Blow, Jens Birktoft and Hartley. Later refinements in the model of this mechanism were made possible by more exact knowledge of the relative proton-binding abilities of the three groups in their molecular environment.

As far as is known all the serine proteases employ the same three amino acid units to hydrolyze peptide bonds; the diversity of the enzymes results entirely from the way they accommodate their specific substrates. It is clear from this model of the catalytic mechanism that if the enzyme is to operate, the substrate must fit the active site exactly.

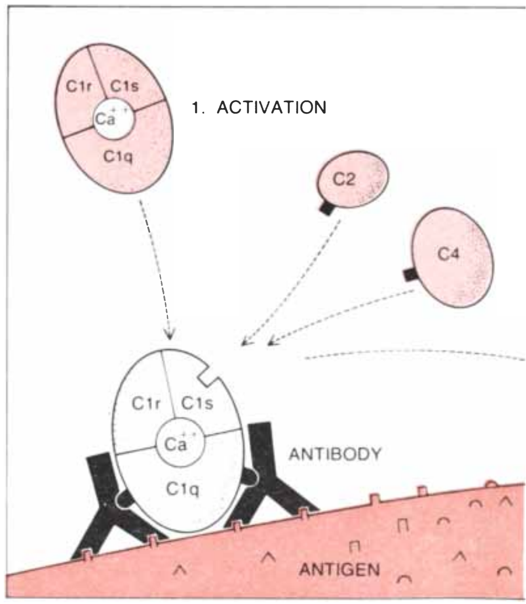
With these insights into the structural compatibility of enzyme and substrate one can see how a single catalytic mechanism can be of such great utility to living organisms. It is employed in many physiological processes where much greater control must be maintained than is required in digestion. One key factor that contributes to the versatility of the



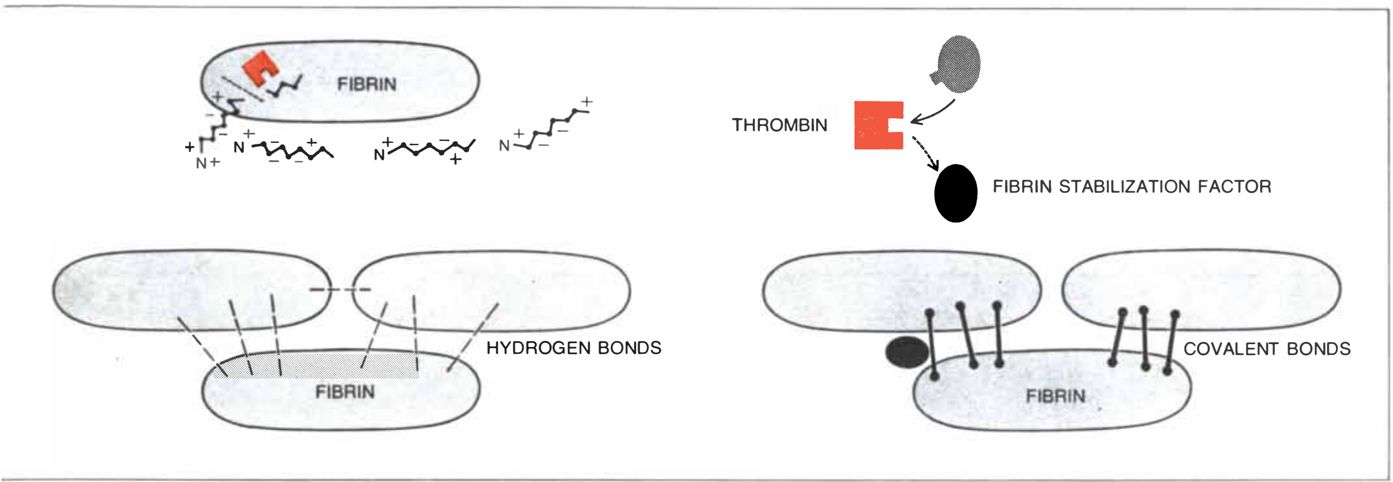
BLOOD CLOTTING is controlled by a consort of enzymes, some of which are serine proteases. The sequence of clotting begins when Hageman factor, released at the site



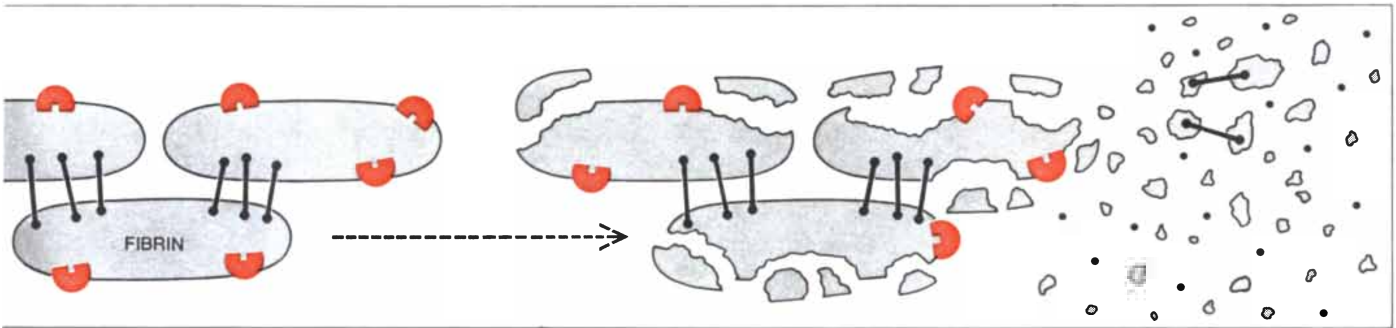
DISSOLUTION of blood clots also requires the action of a serine protease. Plasmin, activated from plasminogen by Hageman fac-



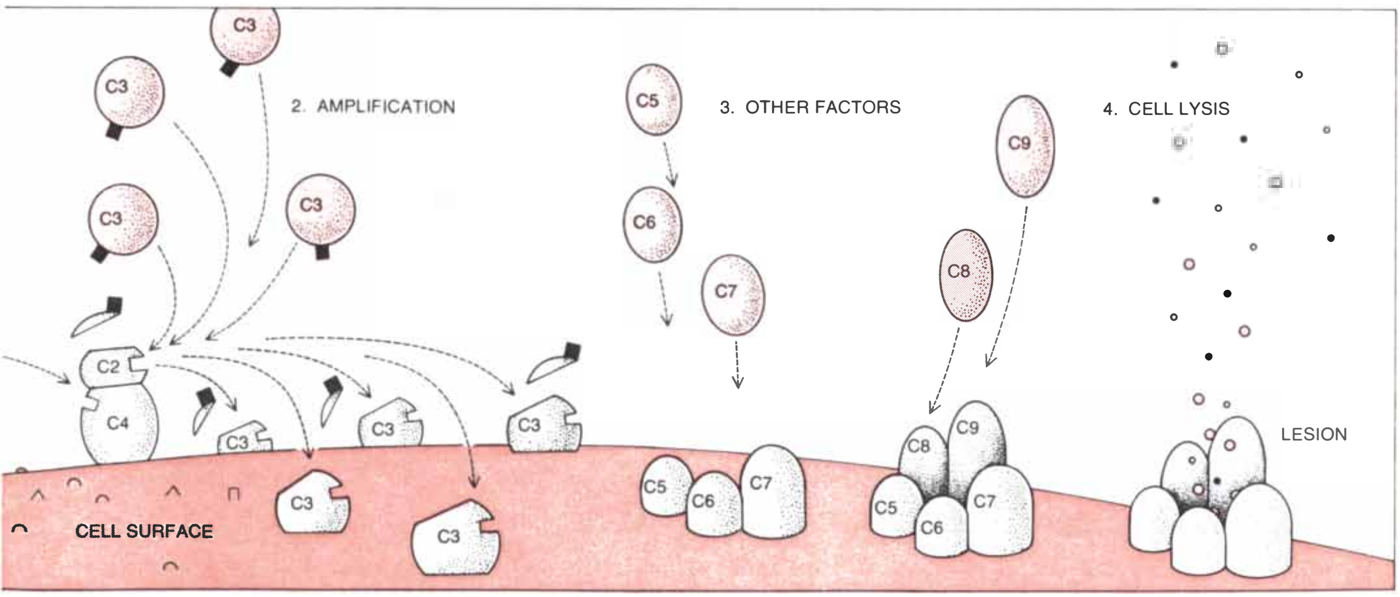
COMPLEMENT SYSTEM, which participates in the immune reaction, consists of 11 proteins; some of them are members of the serine protease family. The system attacks a



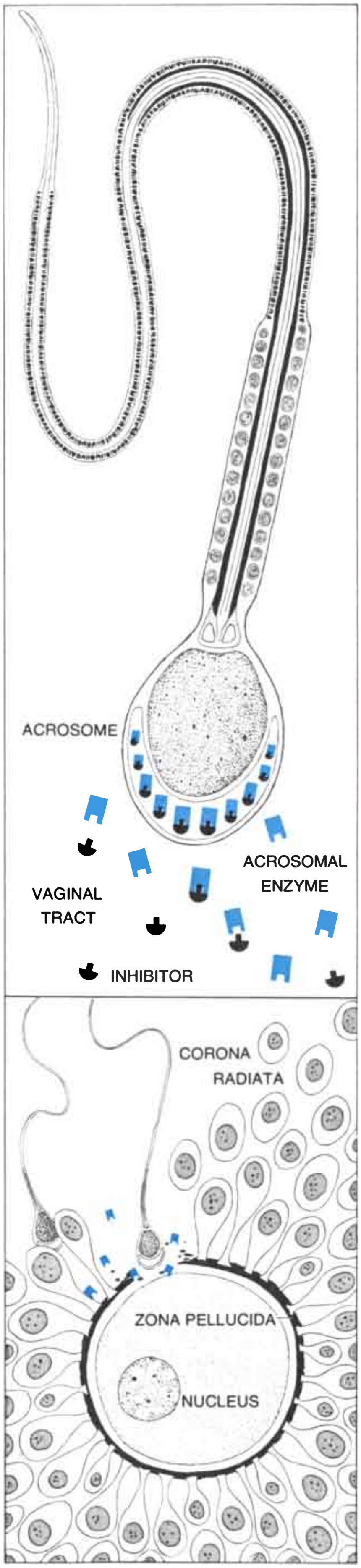
of injury, initiates a "cascade" of enzyme-controlled reactions that ends with the activation of thrombin from prothrombin. Thrombin resembles trypsin but is more specific; it cuts four bonds in fibrinogen, all of them between amino acid units of arginine and glycine. The hydrolysis of these bonds liberates four highly charged peptides and converts the soluble fibrinogen into insoluble fibrin, which forms a fibrous matrix by hydrogen bonding. The fibrin stabilizing factor, activated by thrombin, then bonds the fibrins.



tor and a cofactor, attacks the fibrin matrix at many points. It is much less specific than thrombin, breaking the same kinds of bonds that trypsin attacks: those adjacent to amino acid units of arginine and lysine. The plasmin does not hydrolyze the bonds formed by the fibrin stabilizing factor, but it reduces the fibrin molecule itself to small fragments. These fragments dissolve in the blood serum.



cell when the complex designated C1 binds to one or more antibody molecules already present on the cell surface. C1 is made up of three kinds of proteins, one of which (C1s) is a serine protease. Activated by the binding of the C1 complex, C1s in turn activates C4 and C2 molecules. C4 and C2 become fixed to the cell surface and activate a great many C3 molecules, each of which then binds to a C5 protein. C6 and C7 are bound to C5; this triad migrates to another site, is joined by C8 and C9, and punctures cell membrane.



serine proteases is that in their role as regulatory enzymes they can themselves be precisely controlled.

One powerful means of control is provided by the varying specificity of the enzymes. In their physiological environment the serine proteases catalyze the hydrolysis only of peptide bonds; substances other than proteins, and protein bonds other than the peptide bond, are unaffected. Moreover, the enzymes can be made still more specific, so that they cleave only those peptide bonds that follow a particular amino acid unit; in some cases their activity is confined to a particular bond on a particular molecule. It therefore is not surprising that the serine proteases involved in control and regulation of biological processes resemble trypsin in their specificity, since trypsin, with its electrically charged binding pocket, is the most specific of the digestive enzymes.

The proteases are under still more stringent control because of their susceptibility to enzyme inhibitors, which may be broadly effective or as specific as the enzymes themselves, blocking an entire group of enzymes or only one. A third system of control is embodied in the synthesis of the proteases as inactive precursors that become effectual enzymes only when they are acted on by another molecule. This last mechanism is primarily responsible for ensuring that the enzymes operate only when and where they are supposed to.

Activation of the pancreatic digestive enzymes is initiated by enterokinase, an enzyme secreted in small amounts by the mucous membrane of the stomach. Its primary function is to convert some trypsinogen into active trypsin, which then activates all the proenzymes, including more trypsinogen. In each case the molecule is activated by cutting a single peptide bond, which removes a few amino acid units from the amino terminal end of the proenzyme. Thus the proenzyme serves as a substrate for another proteolytic enzyme [*see illustration on page* 79].

The relatively simple alteration of the molecule represented by the cutting of a single bond results in an extensive modification of its properties. The new amino terminus, which is formed at unit 16, interacts electrostatically with aspartic acid 194, a unit that is adjacent to the reactive serine 195 and is near the binding pocket. The resulting distortion transforms the molecule into the catalytically active configuration.

Kraut and his colleagues, who have analyzed the structure of chymotrypsinogen, have found that the proenzyme appears to have the active site of chymotrypsin but that it lacks a suitable binding pocket for the substrate. Tony Kossiakoff, Lois Kay and I have recently conducted a preliminary investigation of the structure of trypsinogen; most of the molecule is remarkably like the molecule of active trypsin, but there are extensive differences in the substrate-binding region.

It is significant that all these control systems operate not on the active site but on those regions of the enzyme associated with the binding of the substrate to the enzyme. An evolutionary explanation of this regulatory procedure is not difficult to surmise: A highly reactive site might catalyze the desired process with great efficiency, but it would also participate in many undesired reactions. It is advantageous instead to have a moderately efficient catalytic site coupled with a very selective binding requirement.

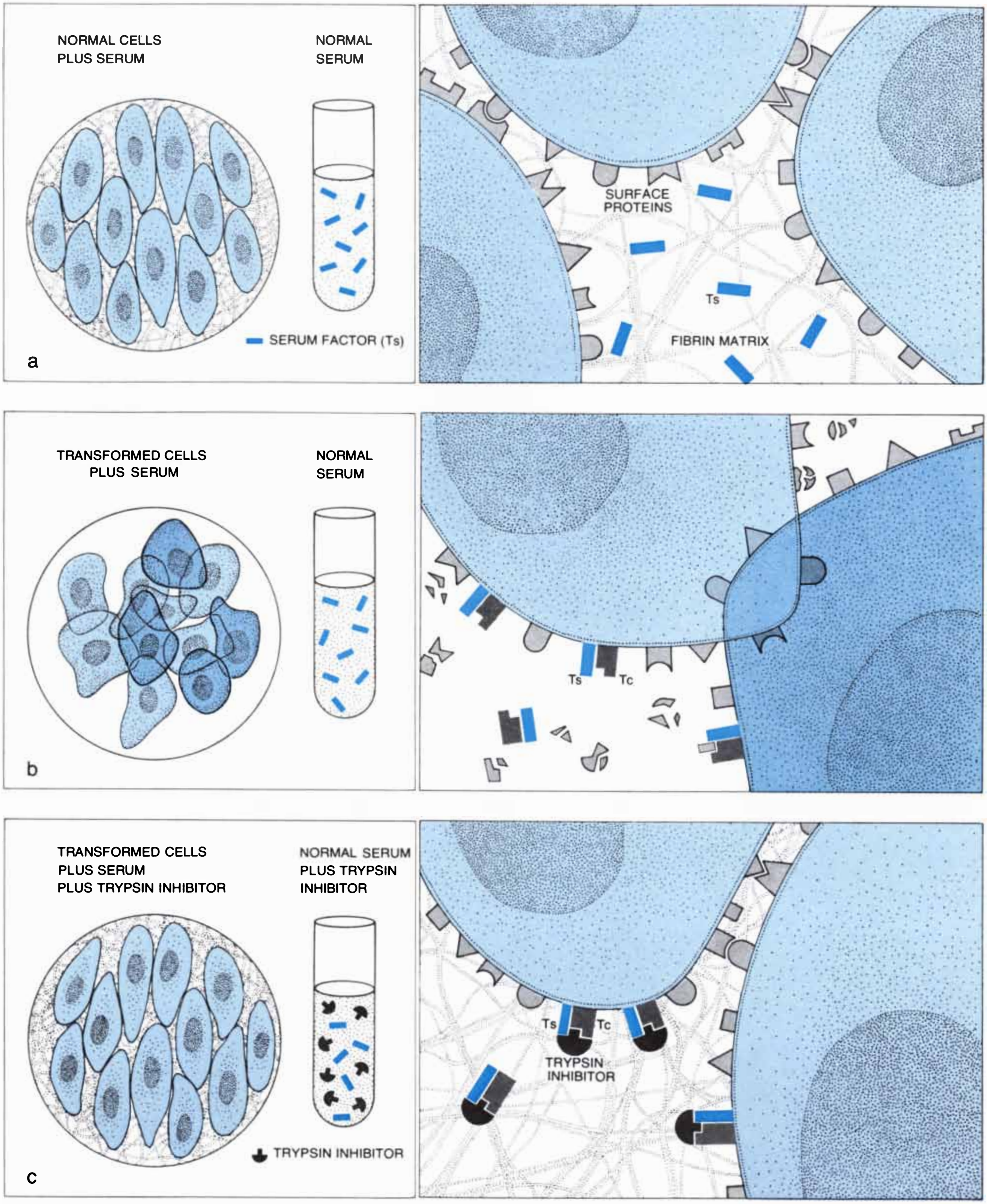
We have so far examined the serine proteases that participate in digestion, but the methods by which these molecules are controlled are more fully elaborated in other enzymes. The function of the digestive enzymes is merely to decompose all the proteins they encounter; the enzymes involved in such functions as blood clotting and tissue repair must be much more selective.

When tissues are broken, the imperative first requirement of the body's defensive systems is to stop the leakage of blood. This process is begun immediately, since the tissue injury itself releases activating agents. One of these agents is Hageman factor.

The release of Hageman factor initiates a cascade of enzyme-controlled reactions. A precursor is activated; the resulting enzyme activates another precursor, and so on, until six enzymes, of which at least three are serine proteases, are activated in sequence. The last enzyme in the cascade is thrombin, a serine protease very much like trypsin.

The natural substrate of thrombin is

FERTILIZATION OF THE OVUM requires the presence of a serine protease supplied by the spermatozoon. The enzyme is carried inside a thin cap, the acrosome, and is called the acrosomal protease. In the acrosome it is combined with an inhibitor, which is removed by some unidentified element in the female reproductive tract. Active enzyme dissolves zona pellucida, a transparent membrane that surrounds the ovum.



GROWTH OF CANCER CELLS in a laboratory culture is influenced by a serine protease. Normal cells, grown on a matrix of fibrin, form an ordered array one cell thick (*a*). The cells are inhibited from further growth when they touch one another, and the fibrin layer is undisturbed. Cancer cells, on the other hand, grow in clumps (*b*). They are unaffected by contact with one another, and their appearance is changed. The peculiar growth characteristics of the transformed cells can be explained in part by the action of a serine protease formed by the association of a factor in the serum (Ts) and a factor secreted by cancer cells (Tc). The enzyme attacks proteins, including those on the cell surface, some of which serve as markers of cellular identity; it also dissolves the fibrin matrix and for this reason is called fibrinolysin. The addition of a trypsin inhibitor to the serum (*c*) prevents the rapid growth of cancer cells. While inhibitor is present cancer cells look and behave as if they were normal, and the fibrin matrix is not attacked.

fibrinogen, a long, soluble, sticklike protein that circulates in the blood serum; the enzyme converts it to fibrin, an insoluble protein with a unique capability: its molecules can cross-link in large numbers to form a solid matrix. In the transformation from fibrinogen to fibrin four short, highly charged peptides are released; it has been suggested that the charged groups on these peptides are responsible for the change in the protein's solubility.

The fibrin matrix is initially held together by hydrogen bonds. An enzyme called fibrin stabilizing factor, which is activated by thrombin, links the fibrin molecules to each other with covalent bonds, joining the acidic side chains of glutamic acid to the basic side chains of lysine. The resulting mass of tangled threads is what we call a blood clot [*see top illustration on pages 82 and 83*].

Thrombin is a close relative of trypsin, but it is much more selective. Thrombin cuts only four bonds in fibrinogen (molecular weight 330,000), whereas trypsin would reduce the same molecule to 150 fragments. The peptide bonds split by thrombin are all between arginine and glycine. Arginine is one of the side chains for which trypsin is specific, and the mechanism of substrate binding is probably very similar in the two molecules. It is known that in thrombin, as in trypsin, aspartic acid 189 plays a role in binding the arginine side chain, and so it is likely that thrombin too has a pocket next to the active site. Thrombin, however, apparently has another binding site to which the substrate must conform, a site that will accept only glycine, the smallest of the amino acids.

Thrombin is clearly much more selective than the digestive enzymes; as a result most natural inhibitors of trypsin are useless against it. A few, however, are known. Bloodsucking creatures, such as leeches and bedbugs, possess proteins that inhibit thrombin and thereby prevent clotting of the blood they drink. Another thrombin inhibitor is important in the treatment of pathological thromboses. Such clots are produced when the clotting mechanism is activated in response to a spurious stimulus; clinical treatment involves the injection of heparin, a polysaccharide derived from the liver, which stimulates the release of a substance that inhibits thrombin and thereby prevents the growth of existing clots or the development of new ones.

The dissolution of blood clots, after the tissue has been repaired, is also mediated by a serine protease. When the fibrin matrix is formed, another serum proenzyme, plasminogen, is incorporated between the filaments. It is activated by the cleavage of a single arginine-valine bond, much as trypsin is activated, and, more like trypsin than thrombin in specificity, it breaks down the fibrin matrix by splitting all the peptide bonds following arginine or lysine units. Known portions of its amino acid sequence closely resemble those of trypsin. People with arterial

ENZYME	SOURCE	FUNCTION	MOLECULAR WEIGHT	NATURAL INHIBITORS	PEPTIDE BOND CLEAVED
TRYPSIN	PANCREAS	DIGESTION OF PROTEINS	24,000	PANCREATIC INHIBITOR	Arg-X OR Lys-X
CHYMOTRYPSIN	PANCREAS	DIGESTION OF PROTEINS	25,000	PANCREATIC INHIBITOR	LARGE HYDROPHOBIC-X
ELASTASE	PANCREAS	DIGESTION OF ELASTIN	24,000		Ala-X OR Gly-X
THROMBIN	SERUM OF ALL VERTEBRATES	BLOOD CLOTTING	33,700	PLASMA ANTITHROMBINS	Arg-Gly IN FIBRINOGEN
FIBRINOLYSIN	SARCOMA CELLS	UNCERTAIN (IMPLICATED IN CELL TRANSFORMATION)		INHIBITORS PRODUCED IN SERA OF INFECTED ANIMALS	
PLASMIN	SERUM OF ALL VERTEBRATES	DISSOLUTION OF BLOOD CLOTS	75,000	PLASMA ANTIPLASMINS	Arg-X OR Lys-X IN FIBRIN
KILLIKREIN	MANY TISSUES AND BODY FLUIDS	GENERATION OF KININS, PAIN SENSING	33,000 TO 36,000	ASPIRIN, SERUM INHIBITORS, KILLIKREINASE	Arg-X
COMPLEMENT C1 AND OTHERS	SERUM	CELL LYSIS IN IMMUNE REACTION	80,000	SERUM C1 INACTIVATOR	Arg-X
ACROSOMAL PROTEASE	ACROSOME OF SPERM	PENETRATION OF ZONA PELLUCIDA OF OVUM	27,000	ACROSOMAL INHIBITOR	Arg-X OR Lys-X
KERATINASE		DIGESTION OF HAIR KERATIN	27,000		
SOME COLLAGENASES	CRAB PANCREAS	DIGESTION OF COLLAGEN			
LYSOSOMAL PROTEASES	ANIMAL CELLS, PLANT ROOTS	CELL DESTRUCTION, DEFENSE AGAINST VIRUSES AND BACTERIA			
COCOONASE	MOTH LARVAE	DISSOLUTION OF COCOON IN METAMORPHOSIS	24,000		
ALPHA-LYTIC PROTEASE	*BACILLUS SORANGIUM*	FUNCTION UNKNOWN, POSSIBLY DIGESTION	20,000		Ala-X OR Gly-X
BACTERIAL TRYPSIN	*STREPTOMYCES GRISEUS*	FUNCTION UNKNOWN, POSSIBLY DIGESTION	21,850		Arg-X OR Lys-X
SUBTILISIN	*BACILLUS SUBTILIS*	FUNCTION UNKNOWN, POSSIBLY DIGESTION	28,000		BROAD SPECIFICITY

DIVERSITY OF SERINE PROTEASES is suggested by the variety of physiological roles they perform. Trypsin, chymotrypsin, elastase, thrombin, plasmin, killikrein, cocoonase, alpha-lytic protease and bacterial trypsin all share a similar amino acid sequence and, by inference, a three-dimensional structure. These enzymes, and all but one of the others listed here, almost certainly have a common evolutionary origin. The exception is subtilisin, which has the same active site and catalytic function as trypsin but evolved independently. In addition to the natural inhibitors listed many of the proteases can be inhibited by relatively small synthetic molecules. Some of the enzymes have not been studied in detail and their molecular weight, inhibitors and specific bond are not yet known.

thromboses or rheumatoid arthritis often have higher than normal levels of plasmin inhibitors, which suggests that the inability to eliminate fibrin may be a factor in maintaining chronic inflammation.

A third serine protease activated by Hageman factor is killikrein. By liberating small peptides called kinins from another protein circulating in the blood plasma, killikrein initiates the dilation of capillaries in the injured region and stimulates other processes important to pain-sensing.

Another group of serine proteases that circulate in the bloodstream is part of the system of proteins known as complement. Complement takes part in the immune reaction directed against foreign tissues or organisms. It consists of 11 proteins that in concert puncture the membrane of an alien cell, allowing the contents to leak out [see "The Complement System," by Manfred M. Mayer; SCIENTIFIC AMERICAN, November, 1973].

The complement system is activated when one of its components, designated C1, becomes bound to antibody molecules that have attached themselves to the surface of the foreign cell. Antibodies ordinarily bind only to unfamiliar cells, and this first step thus provides for the recognition of alien cells or organisms. C1 is an aggregate of three proteins, one of which is a trypsinlike enzyme activated when the C1 complex binds to the cell-bound antibodies [*see bottom illustration on pages 82 and 83*].

The function of the C1 protease is to activate two more components of the system, C4 and C2, which become attached directly to the cell membrane. The complex they form then activates many molecules of C3, a proteolytic enzyme that becomes fixed to the membrane nearby. This stage of the process provides amplification; a single C4,C2 complex triggers the deployment of an army of C3 enzymes. This stage and others also incorporate a safety factor: because all the components of the system except the first will bind to and attack any cell they encounter, including the cells of the host organism, they must not be allowed to escape from the region of the C1-marked cell. To ensure that the host is not attacked, C4 and C3 remain active only briefly unless they bind to a cell membrane.

In succeeding stages the C3 enzyme prepares a three-molecule complex of C5, C6 and C7 on the membrane, and this mediates the final process, in which C8 and C9 join the complex and together puncture the membrane. This seems to be the final step in the destruction of bacteria, and possibly of tumor cells as well.

Several components of the complement system are serine proteases. The most thoroughly studied of the group is C1 protease, which resembles trypsin in its specificity. If it is blocked by a trypsin inhibitor, the entire sequence of proteins is inoperative.

The lysis of bacteria and of red blood cells are the best known processes in which the complement system participates, but the system is involved in several other functions as well, such as bringing white blood cells to the site of an infection. It may also play a role in the rejection of transplanted tissue. The complement system would seem to be a fail-safe instrument of cellular warfare. In fact, serious pathological conditions, such as nephritis and serum sickness, ensue if the complement system is prematurely activated, and a malfunction of the system is associated with another disease: rheumatoid arthritis.

Another trypsinlike enzyme is instrumental in the crucial event in mammalian reproduction: the union of the sperm and the egg. The enzyme is carried in a thin cap called the acrosome that covers the front two-thirds of the sperm head. When the spermatozoon becomes attached to the ovum, the acrosomal enzyme, together with a number of other enzymes, facilitates the penetration of the egg. The specific function of acrosomal protease is to cut open the zona pellucida, a transparent membrane surrounding the egg [*see illustration on page 84*].

In isolation human acrosomal enzyme reacts much like trypsin, and it is blocked by many of the naturally occurring trypsin inhibitors. The method by which it is regulated, however, is unlike that of most other serine proteases. The acrosomal enzyme is present in the sperm not as a proenzyme but as an enzyme bound to an inhibitor. The enzyme is activated by the removal of the inhibitor, a process promoted by some unidentified factor in the female genital tract.

In the control of this enzyme there may be a potential method of contraception. One demonstration of this possibility is that injection of a synthetic trypsin inhibitor into the vagina before copulation prevents fertilization in rabbits. A great many questions of safety and efficacy remain to be answered, however, before the method could be applied to the control of human reproduction.

Another serine protease that may eventually bring humanitarian benefits is one that seems to be involved in the development of malignancies. The most obvious distinguishing properties of cancer cells are that they do not stop growing when they are in close contact with one another, and that they are able to migrate to and grow in alien environments alongside cells of other types. The fundamental components of the system that prevents normal cells from behaving in this way are the antigen molecules on the cell surface, which can be identified by neighboring cells or by the immune system. The antigens on cancer cells have been changed, and as a result the behavior of the cells is also changed. Most of the time tumor-cell antigens are recognized as being different by the immune system, and the cells are destroyed. Occasionally, however, a transformed cell is allowed to grow, perhaps because some of the antigens have been blocked or because the immune system is in some way inadequate to cope with the growth rate.

When normal cells are transformed by cancer-producing viruses or by chemical carcinogens, a trypsinlike enzyme is found to be associated with the cell surface. The enzyme is called fibrinolysin because it will dissolve a fibrin matrix laid down under a culture of cells. Soon afterward the enzyme can be detected, and the cells take on the appearance and behavior of mature cancer cells and begin to grow uncontrollably.

The proteolytic enzyme is found on many types of human and animal cancer cells, but it is not found on normal cells, except in one or two very special cases. Max M. Burger, who was then at Princeton University and is now at the Biozentium in Switzerland, showed that treatment with proteases temporarily releases certain normal cells from contact inhibition and induces the appearance and surface properties characteristic of transformed cells. Fibrinolysin strips off some of the protein-bound antigens, and so modifies the surfaces of neighboring cells. It also controls the growth rate of cancer cells, and therefore it appears to be a crucial link in the chain of events leading to cell transformation and growth.

Edward Reich of Rockefeller University has shown that fibrinolysin consists of two components, one generated by the transformed cell and the other found in the serum of man and many other vertebrates. The serum of cancer patients, but not that of healthy persons, contains an inhibitor of the enzyme developed by the host in response to the tumor. By blocking fibrinolysin the inhibitor retards the growth and spreading of cancer cells; in cell cultures low concentrations of trypsin inhibitors almost totally prevent the growth of transformed cells. The malignant cells look and behave like normal cells as long as the inhibitor is present. Normal cells are often unaffected by these inhibitors, even at much higher concentrations.

The recent discovery of the effect of fibrinolysin inhibitors could prove to be particularly important, since it suggests possible ways to attack cancer cells selectively. Perhaps a specific fibrinolysin inhibitor will be found that can depress the growth rate of cancer cells sufficiently for the immune system to destroy them more quickly than they grow. A test for the presence of fibrinolysin inhibitor could become a method of early diagnosis.

All the serine proteases discussed so far are extracellular enzymes, that is, they are synthesized within the cell but act outside of it. A few intracellular serine proteases are also known; they are among the enzymes contained in the organelle called the lysosome [see "The Lysosome," by Christian de Duve; SCIENTIFIC AMERICAN, May, 1963]. Lysosomes are small sacs of enzymes that are involved in the defense of the cell against viruses and other microorganisms and that participate in the programmed control of embryonic development. The role of the serine proteases, particularly in the latter process, is not clearly understood.

Another serine protease functions in certain moths during metamorphosis; it is called cocoonase, and it facilitates escape from the cocoon after pupation. The synthesis of the enzyme begins abruptly eight or nine days after the larva begins to metamorphose. Enormous quantities of an inactive proenzyme are produced and stored in specialized secretory cells, then activated and released to the cocoon surface, where the enzyme dries as a pure deposit. Later, as the moth prepares to emerge, other glands secrete a potassium bicarbonate solution. The enzyme dissolves, breaks up the sericin, or "silk gum," that holds the fibers of the cocoon together, and the moth escapes.

A group of trypsinlike serine proteases is also found among the bacteria. Of greater interest, however, is another family of bacterial enzymes found in the species *Bacillus subtilis.* These subtilisins are serine proteases, but they differ from the trypsin family in an extraordinary way: they are of independent ancestry. Although the active site of the subtilisins has exactly the same arrangement of catalytically important groups present in trypsin, including serine, histidine and aspartic acid units, Kraut and his colleagues have demonstrated that the overall architecture of the molecule is entirely different. Trypsin and the subtilisins are thus an example of convergent evolution on a molecular scale; two unrelated molecules evolved independently to give rise to the same catalytic mechanism.

The evolution and diversification of all the serine proteases raise intriguing questions. Their functions are so diverse it may be difficult to see how they can be related at all, yet there can be little question that they are descendants of a single ancestral enzyme. For those that have been studied in the greatest detail the similarities of amino acid sequence and three-dimensional structure can be explained only as homologies (features derived from the same evolutionary origin), not as analogies (features developed independently that perform similar functions). Even where the complete sequence and structure are not known, considerable evidence of homology can be obtained by chemical methods. In particular, various inhibitors can be employed to investigate the relationships of the enzymes. The binding of the inhibitors depends not only on the chemistry of the active site but also on the geometry of the binding sites near it, so that inhibition of two enzymes by the same molecule suggests that relatively large areas of the enzymes are very similar.

Although the kinship of the serine proteases is established, the evolutionary pathways by which they came to assume their present forms and functions are not. In this regard the evolution of many enzymes is much more complex than that of most other proteins, since the enzymes could not have evolved independently of the molecules they interact with. Thrombin and fibrinogen, for example, are both complex molecules that must have long been in association; any critical change in one could not be tolerated without a corresponding change in the other. The evolution of the enzyme is inextricably entwined with that of the molecules that activate it from the proenzyme and with that of the molecules that inhibit it; most of all, the evolution of the enzyme is dependent on that of the substrate. Indeed, the enzyme and the substrate evolve as a single system to control life processes; their interdependence is the yin and yang of biochemistry.

What would you do with a Statistics and Number-Crunching Computer that starts at $7,100,* has 16K Hardwired Basic Language and 28 Major Peripherals?

Plenty!

The new Wang System 2200 is a System. It gives you the raw power and the peripherals you must have for a wide range of problem solving. For under $7,200 you get a CPU with 16K bytes of BASIC language instructions hardwired into the electronics . . . plus a 4K operating memory. You also get a big 16 lines (of 64 characters each) CRT display, a console mag tape drive and your choice of either alpha or BASIC Keyword keyboards.

Some Words About Language: The hardwired MOS ROM language in your System 2200 finally ends your dollar tradeoffs . . . economy systems that are costly to program or very expensive systems that are relatively easy to program. Many, if not most, of your people already know BASIC. They'll be solving problems the day your system is delivered (and, we can deliver in about two weeks). Most of your budget will go into problem solving; not system support.

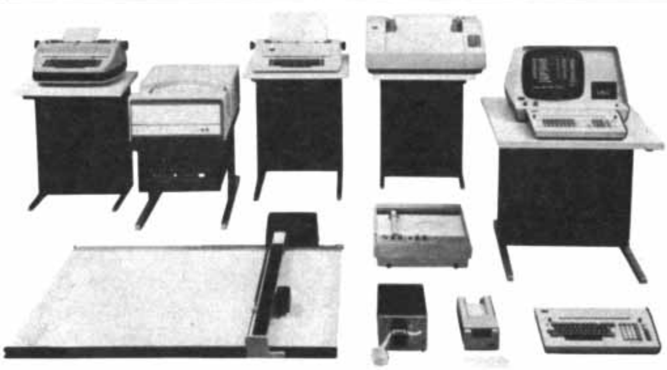
Try To Out-Grow It: Main memory is field expandable in 4K increments (at $1,500 per 4K). Up to 32K. You can choose from three kinds (and 7 price ranges) of printers . . . one even has a stepping motor for very precise 4-quadrant incremental plotting. Speaking of plots, we have a new, very large flatbed (31" × 42") for only $7,500 or a smaller one if you plot small. Both print alphanumerics and plot under full program control. Been appalled lately by disk prices? Starting at just $4,000, we offer you our new "floppy" disk in single, double and triple disk configurations (.25, .50 and .75 MB's). For big disk power, you can have 1, 2 or 5 megabyte fixed/removable disk systems. All peripherals, including punched or mark sense hopper card readers, paper tape readers and on- line BCD or ASCII controllers are easily added-on in the field so your System 2200 will grow with your needs.

The Wise Terminal: If you are now or may soon be getting into terminals, we have several new products that will instantly upgrade your System 2200 for telecommunications with any other System 2200 or a mainframe computer. And, you still have a powerful stand-alone system. Another approach, of course, is to justify it as a powerful terminal and get a "free" stand-alone computer. Wise?

We Do A Lot For You: System 2200 is backed by over 250 factory-trained Wang Service Technicians in 105 U.S. cities. Naturally, we guarantee or warranty everything you buy from us. If you want, there are free programming/operating schools here in Tewksbury, Massachusetts, almost every week. We have a growing program library on a wide range of statistics and math/science applications. Our user group (with the unlikely name of "SWAP") could help you cut programming costs even further. We do a lot for you.

* All prices U.S. List. If you're the entrepreneur type, we've just announced a new 7-module Basic Accounting System software package for the business end of your business like payroll, invoicing, inventory, receivables and some really fancy management reports.

WANG



Even if you call the Wang System 2200 a small system . . . you have to admit it's a big idea.

☐ Please send literature

☐ Please have your representative phone me: ____________

Name ____________ Title ____________

Company ____________

Address ____________

City ____________

State ____________ Zip ____________

SA-7

Wang Laboratories, Inc.,
836 North St., Tewksbury, Ma. 01876, Tel. (617) 851-4111

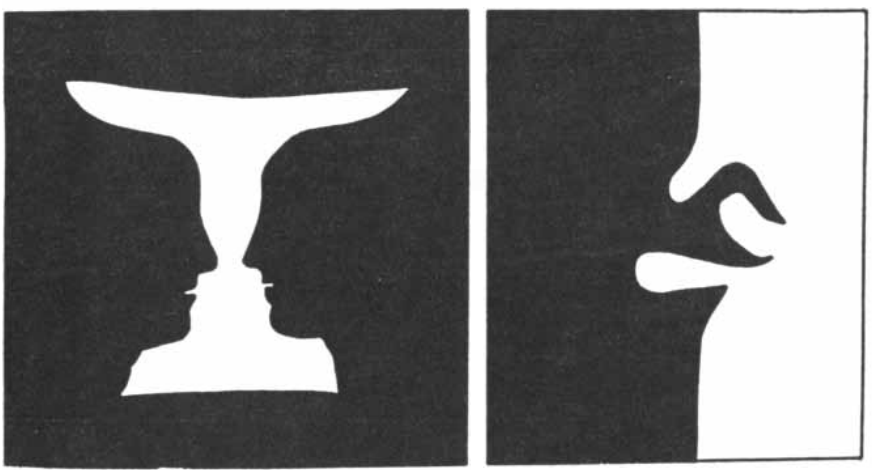
Sources of Ambiguity in the Prints of Maurits C. Escher

The fascinating graphic inventions of the late Dutch artist reflect a strong mathematical and crystallographic influence. Their original inspiration, however, came from experiments on visual perception

by Marianne L. Teuber

If ambiguity is a sign of our time, the late Dutch graphic artist Maurits C. Escher managed to represent it in striking visual terms. In Escher's art there is the ambiguity of figure and ground; the ambiguity of two and three dimensions on the flat surface; the ambiguity of the reversible cube; the ambiguous limits of the infinitely small and the infinitely large. In his prints visual ambiguity goes hand in hand with ambiguity of meaning. Good and evil are contemplated in the figure-ground reversal of black devils and white angels. Day changes into night; sky becomes water; fish metamorphose into fowl.

Escher's regular subdivisions of the surface have been compared to the packed periodic structures of crystals and to the mathematical transformations of topology and non-Euclidean geometry [see "Mathematical Games," SCIENTIFIC AMERICAN, April, 1961, and April, 1966]. The original inspiration for his unusual ambiguous patterns, however, can be traced to contemporary sources more familiar to the student of psychology and visual perception. Psychological studies of the relation of figure and ground—in turn encouraged by the positive and negative forms of Art Nouveau—were for Escher the primary stimulus. Only after he had mastered reversible figure-ground constructions of his own invention did he recognize their similarity to certain principles of crystallography, and his interest in mathematics was aroused.



REVERSIBLE PATTERNS published originally in 1915 by the Danish psychologist Edgar Rubin in a study of the role of the figure-ground relation in visual perception were presumably familiar to Escher at an early stage. The pattern at left can be seen either as a vase in the center or as two profiles facing each other. The more abstract pattern at right can be seen either as a black figure against a white background or vice versa. In each case it is impossible, said Rubin, to see both the black and the white areas simultaneously as figures.

Figure-ground designs appear at an early stage in Escher's work—as early as 1921, when he was only 23 years old. It was not until the late 1930's, however, when Escher "rediscovered" the style that made him famous, that the figure-ground ambiguity clearly became the dominant feature of his art. The well-known woodcuts *Development I, Day and Night* and *Sky and Water I,* among others, date from this period.

From Escher's own commentary on these prints in *The Graphic Work of M. C. Escher* one must conclude that he knew the pertinent psychological literature. In particular he seems to have been familiar with the early experiments on figure and ground by the Danish psychologist Edgar Rubin, with Kurt Koffka's 1935 book *Principles of Gestalt Psychology*, where Rubin's results are summarized, and with the studies of Molly R. Harrower, a student of Koffka's. Patterns related to crystallography and geometry turn out to be a later development.

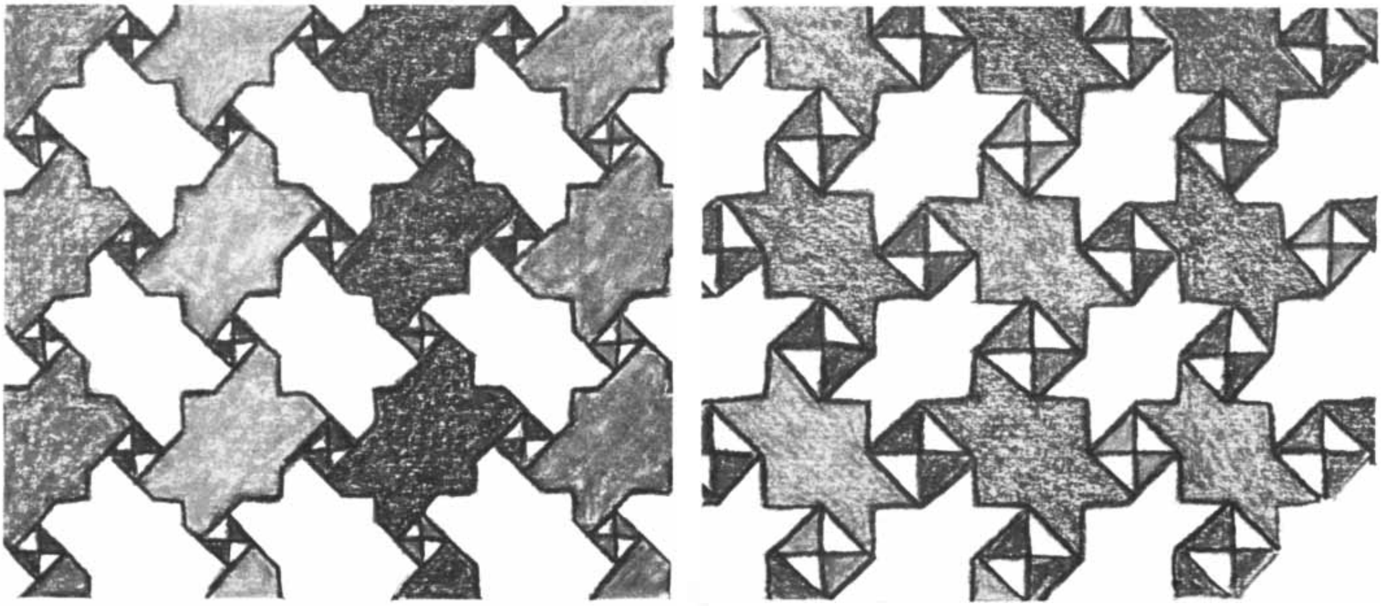
One of Rubin's best-known patterns, presented in his 1915 monograph dealing with visually perceived form as a function of the relation of figure to ground, can be seen either as a vase in the center or as two profiles facing each other [*see illustration at left*]. When the profiles are seen, the vase becomes ground, and vice versa. It is impossible, as Rubin points out, to see vase and profiles simultaneously as figures.

Rubin's book was published in a German translation in Copenhagen in 1921. The following year, when the young Escher was completing his training at the School of Architecture and Ornamental Design in Haarlem, he carved a woodcut called *Eight Heads* [*see illustration below*]. Each head fills exactly the space left between neighboring heads and acts alternately as figure and ground, depending on the viewer's attitude. Escher is quite explicit about his purpose. In *The Graphic Work of M. C. Escher* the artist himself provides an introduction and comments on his prints. These explanations reflect the technical language of his scientific sources. The wording here leaves no doubt about the specific psychological studies that contributed to the formation of his unique style.

One can easily discern the link with Rubin's experiments when Escher classifies *Eight Heads* and his later reversible patterns of fish and bird in *Sky and Water I* or *Day and Night* under the heading "The Function of Figures as a Background." Escher comments: "Our eyes are accustomed to fixing on a specific object. The moment this happens everything round about becomes reduced to background." This description is in keeping with Rubin's analysis of his ambiguous vase-profile pattern.

Whereas Escher insisted on meaningful, if fantastic, creatures for his basic reversible units, Rubin, in his attempt to find general principles of figure formation, actually preferred more abstract designs. In his 1915 book Rubin even cites Wassily Kandinsky's contemporary abstract paintings as good examples of the

"EIGHT HEADS," a woodcut carved by Escher in 1922, the year he finished his training at the School of Architecture and Ornamental Design at Haarlem in the Netherlands, bears a strong resemblance to Rubin's diagrams. Each of the four male and four female heads can act alternately as figure or ground. *Eight Heads* was Escher's first attempt at an infinitely repeating subdivision of a plane surface. This reproduction is from a print in the Escher Foundation collection in the Haags Gemeentemuseum at The Hague.



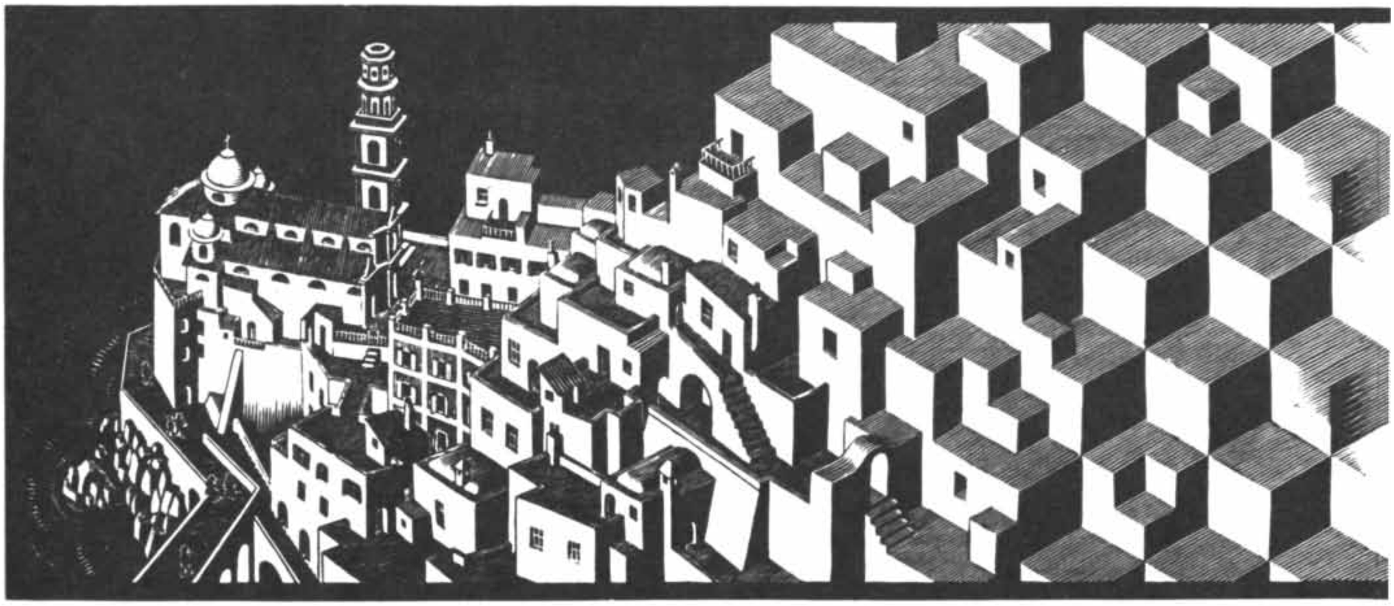
MAJOLICA TILES at the Alhambra in Spain were sketched by Escher during a trip in 1936. The internal symmetry and the ambiguous contours of the Moorish designs appear to have helped revive Escher's early fascination with the figure-ground problem.

equivalence of color fields. By means of abstract test patterns Rubin hoped to isolate basic characteristics of form perception without the distractions caused by figurative images. As he points out, it is always the form with the greater realistic or emotional appeal that tends to attract our attention; in the vase-profile pattern, for instance, the profiles will win out when the reversals are observed over prolonged periods.

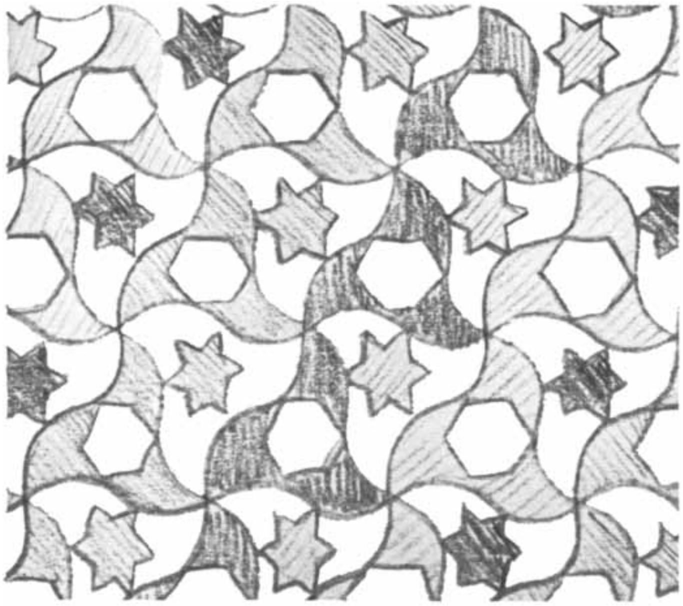
To avoid these pitfalls Rubin derived the main principles of what makes for figure and what for ground from his abstract patterns. According to Rubin, one usually sees the smaller enclosed form as figure by contrast with the larger surrounding expanse of the ground. The figure has "solid-object quality," whereas the ground takes on a "film quality." The figure protrudes; the ground recedes and stretches behind the figure. The contour is seen as belonging to the figure and not to the ground.

In ambiguous patterns, however, the one-sided function of the contour is challenged. Rubin's analysis of contours can be recognized in Escher's description of the difficulties he encountered in drawing his ambiguous creatures. In discussing the borderline between two adjacent shapes having a double function Escher notes that "the act of tracing such a line is a complicated business. On either side of it, simultaneously, a recognizability takes shape. But the human eye and mind cannot be busy with two things at the same moment, and so there must be a quick and continual jumping from one side to the other.... This difficulty is perhaps the very moving-spring of my perseverance."

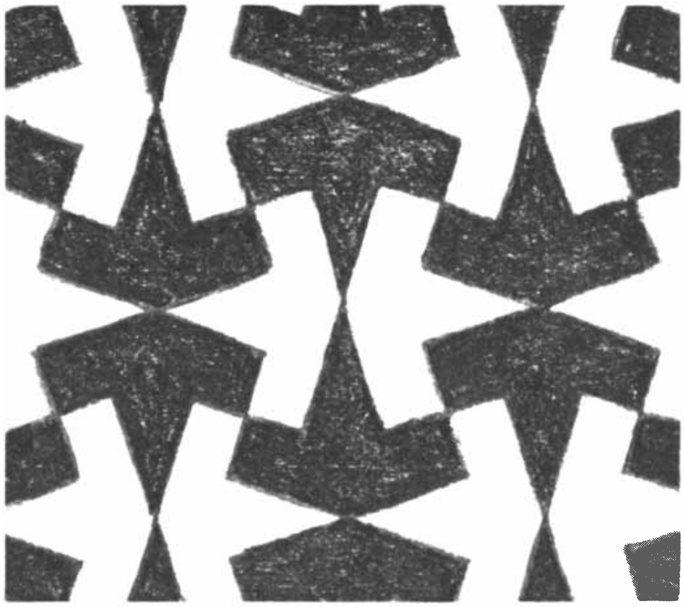
After his early exposure to Rubin's work in 1921 and 1922, Escher left his native Netherlands to live in Italy until



"METAMORPHOSIS I," a woodcut designed by Escher in 1937, represents an abrupt change in his style from the flat periodic subdivisions he had invented up to 1936 to the development of forms from flat to plastic on the two-dimensional plane. This new ap-



Unlike the Moorish artists, however, Escher continued to design his flat tessellations with contiguous human and animal forms.



Escher's original color drawings of the Alhambra tiles are part of the Escher Foundation collection in the Haags Gemeentemuseum.

1934. An extraordinarily skillful craftsman, he created a large series of woodcuts and lithographs; they represent landscapes, architecture and portraiture in brilliantly realistic style, in which the traditional Renaissance picture space prevails. Only once during his Italian period did the figure-ground problem return: in 1926 he designed some interlocking animal shapes, very similar to his later periodic structures on the picture surface. Escher claims that the departure from Italy in 1934 was responsible for his change in style. He felt that landscape in the North did not attract him as it had in Italy; instead, he writes, "I concentrated . . . on personal ideas . . . and inner visions." He lived for more than a year in Switzerland and for five years in Belgium before settling in 1941 at Baarn in the Netherlands, where he remained until his death in 1972, except for brief visits to Britain, the U.S. and Canada.

As I have noted, the major turning point in Escher's style came in the late 1930's after he left Italy. What caused that change? It is evident that at some time between 1935 and 1938 Escher became acquainted with Koffka's *Principles of Gestalt Psychology*. That Escher was indebted to Koffka can be documented from his graphic work and written comments. Koffka, one of the three chief proponents of Gestalt psychology (with Max Wertheimer and Wolfgang Köhler), relied heavily on Rubin's work, although Rubin never counted himself among the Gestalt psychologists. In *Principles of Gestalt Psychology* an entire chapter is devoted to the topic of figure and ground. Thus, through Koffka, an old and fascinating preoccupation of Escher's was revived.

Escher's development can be appre-

proach appears to have been influenced by his reading of Kurt Koffka's 1935 book *Principles of Gestalt Psychology*. The design is reproduced here from an original print in the National Gallery of Art in Washington, D.C. (gift of Cornelius Van Schaak Roosevelt).

ciated by considering drawings the artist made in 1936 on a trip to Spain when he copied majolica tiles at the Alhambra [*see top illustration on preceding two pages*]. These examples of Moorish art with their internal symmetry and obviously ambiguous contours must have attracted him precisely because he had been familiar with the figure-ground problem since 1921. He looked at the Alhambra tiles with eyes trained by Rubin's experiments. As in the case of Rubin's abstract patterns, however, he regretted that the Moorish artists were not allowed to make "graven images" and "always restricted themselves . . . to designs of an abstract geometrical type. Not one single Moorish artist . . . ever made so bold . . . as to use concrete, recognizable, naturalistically conceived figures of fish, birds, reptiles or human beings as elements in their surface coverage."

Instead Escher preferred to design his flat tessellations with contiguous human and animal forms. In the course of the next year or so, however, he abruptly transformed this flat motif into three-dimensional cubes and an entire town in his woodcut *Metamorphosis I* [*see bottom illustration on preceding two pages*]. One of the new aspects noticeable in Escher's post-1937 work, in contrast with the flat periodic subdivisions he had invented up to 1936, is the development of forms from flat to plastic on the two-dimensional plane. In *Principles of Gestalt Psychology* Koffka demonstrates the compelling three-dimensional organization of two-dimensional lines and planes under certain conditions. He shows, among other examples, how a hexagon, depending on the internal arrangement of its lines, can change from a flat pattern to a cube, just as in Escher's *Metamorphosis I*. In summarizing experiments by Hertha Kopfermann and himself, Koffka points out that it is the intrinsic tendency toward simplicity of organization that makes one see one array of forms as two-dimensional and a slightly altered one as three-dimensional.

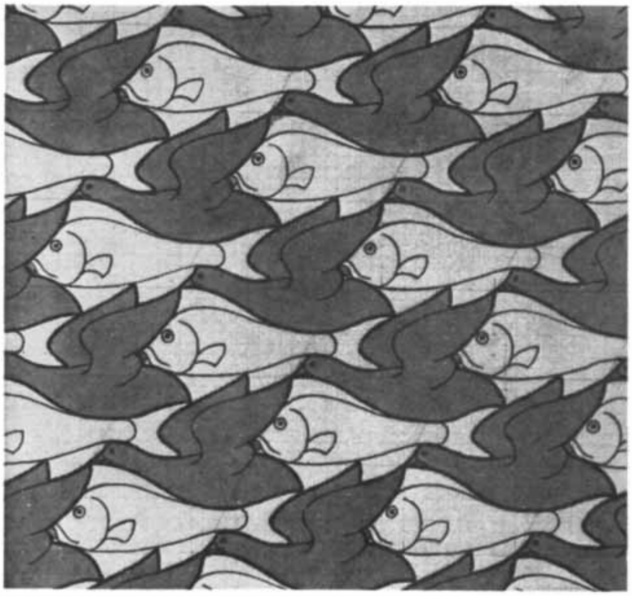
Many of Escher's prints of the following years are based on just such a change of forms from flat to plastic. A noteworthy example is the lithograph *Reptiles,* designed in 1943 [*see illustration below*]. Escher describes this series of prints in words that echo Koffka's text: "The chief characteristics of these prints is the transition from flat to spatial and vice versa." Escher goes on to show how in such designs the individual creatures free themselves from the flat ground in which they are rigidly embedded. He writes: "We can think in terms of an interplay between the stiff, crystallized two-dimensional figures of a regular pattern and the individual freedom of three-dimensional creatures capable of moving about in space without hindrance. On the one hand, the members of the planes of collectivity come to life in space; on the other, the free individuals sink back and lose themselves in the community."

In the same manner Koffka speaks in the last chapter of his book of the embracing ground and the protruding figure as paradigms for the relation between personality and "behavioral social field." The reptiles in Escher's lithograph thus free themselves from the contiguous design on the flat page to venture into three-dimensional space, only to return to the flat surface where their individuality is again submerged. On their way they pass the paraphernalia of smoking and other ephemeral artifacts drawn in hard illusionistic style; they crawl over one of the Platonic solids (a dodecahedron). These forms fascinated Escher because, as he said, geometric shapes were timeless and not man-made.

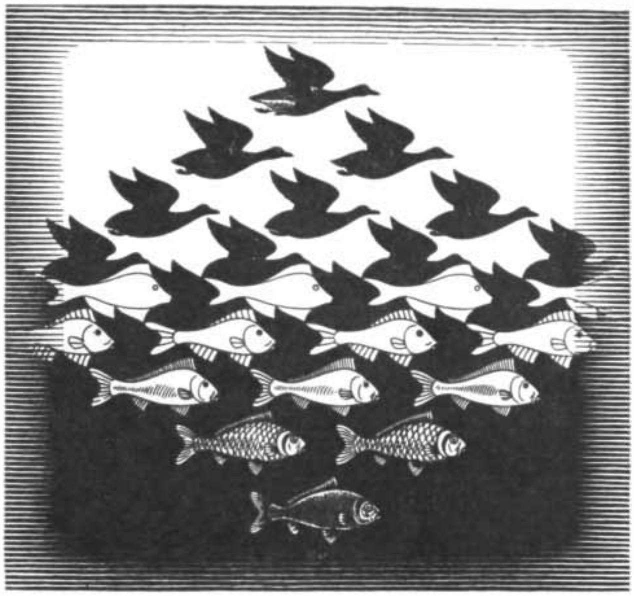
"REPTILES," a lithograph designed in 1943, is a notable example of Escher's increasing preoccupation after 1937 with the transformation of forms from flat to plastic. This reproduction is from a print in the Lessing J. Rosenwald collection at the National Gallery.

During the same period (1936–1938) Escher also became aware of an experimental study by Harrower, who in April, 1936, published an article titled "Some Factors Determining Figure-Ground Articulation" in the *British Journal of Psychology*. She varied Rubin's pattern in the following manner. Several test cards emphasized the outline of the vase and let the profiles recede into the background; other cards emphasized the profiles and allowed the vase to become ground; the center card showed the vase and profiles as being equivalent. Two years later, in 1938, Escher created two of his most striking woodcuts, *Sky and Water I* [*see illustration at top right on opposite page*] and *Day and Night* [*see bottom illustration on opposite page*], according to the same principle; in these works the ground slowly becomes figure and the figure becomes ground; the forms in the center, however, remain equivalent. In his interpretation of *Sky and Water I* Escher employs Harrower's terminology when he says: "In the horizontal central strip there are birds and fish equivalent to each other."

This principle of equivalence, first discussed by Rubin and emphasized by Harrower, is an important ingredient of Escher's many inventive preparatory drawings for his woodcuts. When explaining his compositions, Escher would frequently refer to the fact that his forms had to be "equivalent." The crystallographic terms "distinct" and "equivalent" should not, of course, be confused with the simple notion of equivalence Escher (and Harrower) had in mind. The ingenious basic drawing of *Fish and*



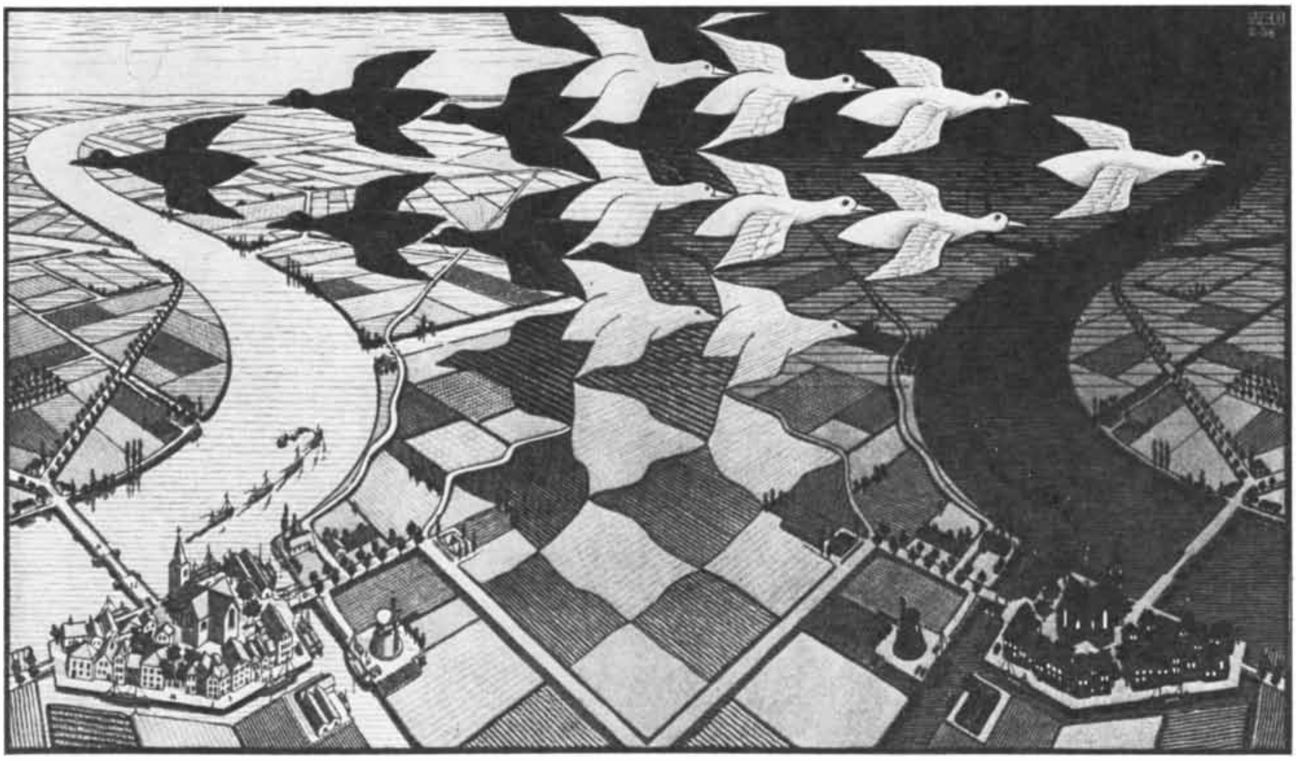
"FISH AND FOWL," a preliminary drawing for the woodcut *Sky and Water I*, is a good example of Escher's interest in the "equivalence" of visual forms, a notion he adapted from Molly R. Harrower, a student of Koffka's. The original is in the Gemeentemuseum.



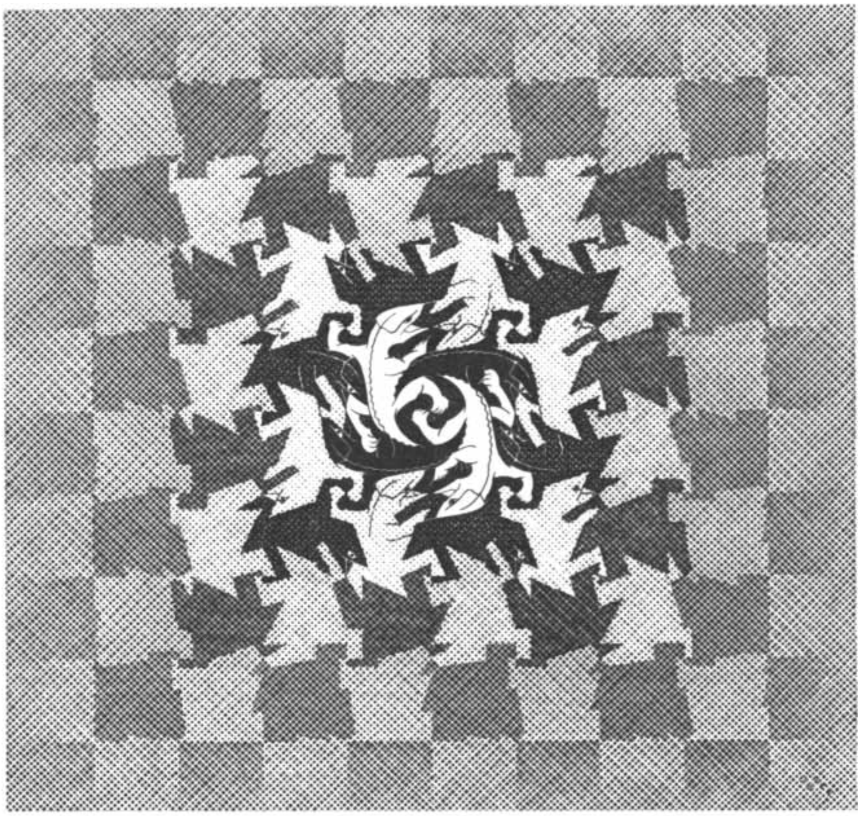
"SKY AND WATER I," carved in 1938, is one of Escher's best-known woodcuts. Unlike the preliminary watercolor drawing at left, the forms of the birds and the fish are equivalent only in the center. This reproduction is from a print in the Gemeentemuseum.

Fowl for *Sky and Water I* is a good example of equivalence; the surfaces of the individual birds and fish are approximately equal in extent, internal design, light-dark contrast and simplicity of contour [*see illustration at top left on this page*]. Such equivalence makes the figures ambiguous, and a rapid reversal is the result.

In her 1936 article Harrower tested the relation of figure to ground by introducing a number of variables, among them increasing and decreasing brightness contrast (or graded grays). Escher's woodcut *Development I,* made in 1937, shows how faint gray squares arranged along the periphery gain in black-and-white contrast as well as distinctness of shape until they become four black and white reptiles in the center [*see top illustration on next page*]. The two "factors"



"DAY AND NIGHT," another 1938 woodcut, represents the same slow transformation of ground into figure and figure into ground, with only the forms in the center remaining equivalent. The principle of transformation is the same as that discussed by Harrower in her 1936 article in the *British Journal of Psychology*. The original print is in the Rosenwald collection at the National Gallery.



"DEVELOPMENT I," a 1937 Escher woodcut, incorporates two basic variables from Harrower's experiments: brightness gradient and development from shapeless ground to distinct figure. This print is in the John D. Merriam collection at the Boston Public Library.

from Harrower's experiments, brightness gradient and development from shapeless ground to distinct figure, are the basic compositional principles of this impressive work. Escher's comment on the print is again couched in the technical language of Harrower's study. He writes: "Scarcely visible gray squares at the edges evolve in form and contrast toward the center."

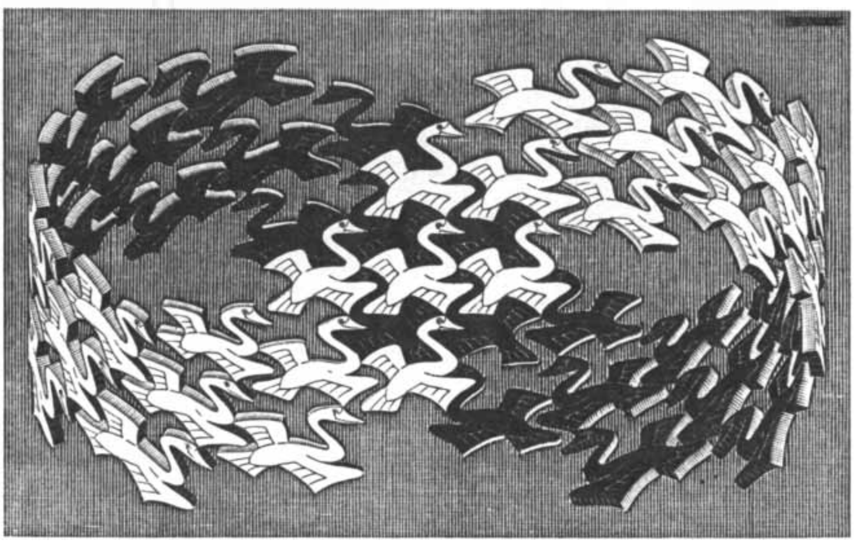
Escher groups several additional prints under the category "Development of Form and Contrast," in keeping with Harrower's analysis. One of these is the lithograph *Liberation,* designed in 1955 [*see illustration on opposite page*]. He describes this print in terms that are reminiscent of Harrower's test cards and Koffka's text: "On the uniformly gray strip of paper that is being unrolled, a simultaneous development in form and contrast is taking place. Triangles—at first scarcely visible—change into more complicated figures, whilst the color contrast between them increases. In the middle they are transformed into white and black birds, and from there fly off into the world as independent creatures, and so the strip of paper on which they are drawn disappears."

In *Liberation* Escher presents us with a surrealist situation; the birds freed from the gray scroll are caught, nevertheless, on the surface on which the lithograph is printed. The artist reflects here on the visual absurdity of his own craft, as he had implicitly in *Reptiles.*

To summarize this important phase in Escher's artistic development, starting in 1937, he transforms his ambiguous figurative patterns in three ways: (1) from flat to plastic, derived from Koffka's *Principles of Gestalt Psychology* of 1935; (2) from shaped form to shapeless ground, derived from Harrower's study of 1936; (3) from strong black-and-white contrast to gray, also derived from Harrower.

Sky and Water I and *Day and Night,* both done in 1938, exhibit these categories of transformation of shape. In *Day and Night* the square gray fields in the foreground gain in articulation of shape and contrast; they become an equivalent pattern of distinct black birds and white birds in the upper center and from there develop into three-dimensional creatures flying off into the "real" world of day or night. In *Sky and Water I* the strongly articulated plastic single bird and single fish, above and below, evolve from the flat equivalent strip in the middle. What was bird becomes watery ground and what was fish becomes sky. Here Escher enhances the individuality, or object quality, of the figure compared with the film quality of the ground, features already emphasized by Rubin in 1915.

It is difficult to reconstruct by what route Escher came in such close contact with the technical aspects of figure-ground experiments. He may have had a mentor. The artist himself belonged to a family where professional and intellectual achievement were the rule, and he may have come across Koffka's and Harrower's experiments because of his own strong interests. The year in the French-speaking part of Switzerland (1936), near the universities of Geneva and Lau-

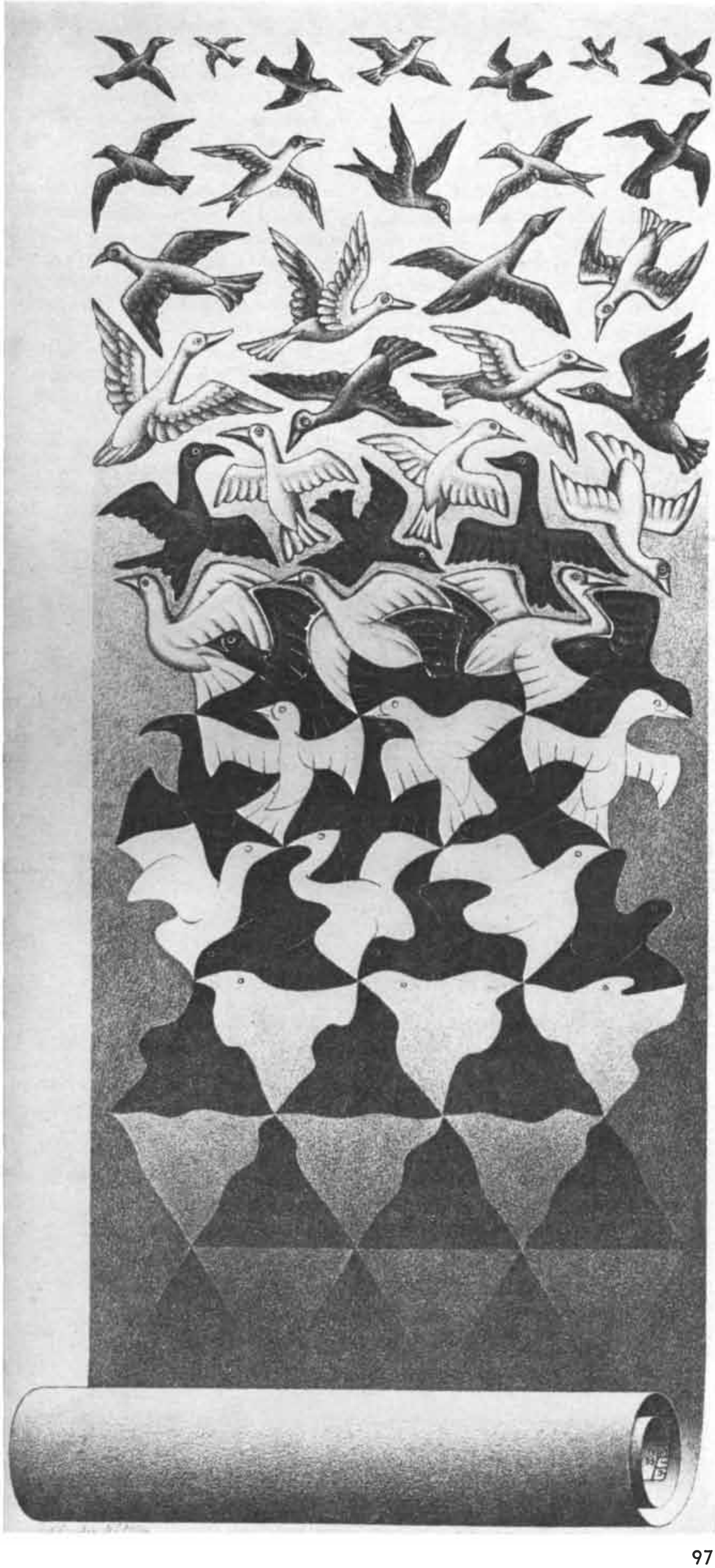


"SWANS," a 1956 woodcut, is a good example of how, in experimenting with space-filling tessellations on a flat surface, Escher often relied on crystallographic rules of transformation. He himself classified this print under the heading "Glide Reflexion." The print used to make this reproduction is in the Roosevelt collection at the National Gallery.

sanne, and the five years in Ukkel, not far from the University of Brussels (1937–1941), were the period of his "conversion," when he made the figure-ground problem a permanent feature of his style. Whatever his contacts may have been, by the 1930's not only was the impact of Gestalt psychology widespread at European universities but also it had become fashionable among intellectuals.

The figure-ground studies of the Gestalt psychologists were not, however, Escher's only source of inspiration. He varied his fantastic tessellations on the picture plane by following the structural principles of periodic packing in crystals. Caroline H. MacGillavry analyzed Escher's inventions in these terms in her 1965 monograph *Symmetry Aspects of M. C. Escher's Periodic Drawings.* In *Color and Symmetry,* published in 1971, A. L. Loeb selects striking instances of form and color symmetry from Escher's work to accompany his text. Escher himself recognized the similarities of his regular subdivisions on the plane to principles of crystallography. They had been pointed out to him by his brother, B. G. Escher, professor of geology at the University of Leyden. By that time, however, the artist had created his own figure-ground patterns based on Rubin's visual analyses and the Moorish tiles at the Alhambra. As the mathematician H. S. M. Coxeter has remarked, the Moors had already made use of all 17 crystallographic groups of symmetry structures, subsequently established by E. S. Fedorov in 1891.

In experimenting with space-filling creatures on the flat surface, Escher arrived at many intriguing compositions that follow crystallographic rules of transformation; a good example is his woodcut *Swans,* designed in 1956 [*see bottom illustration on opposite page*]. Again Escher writes a commentary, as he had done for his figure-ground inventions. He groups these prints under the heading "Glide Reflexion" and acknowledges the "three fundamental principles of crystallography"; they are, in his words, "repeated shifting (translation), turning about axes (rotation) and glide mirror image (reflexion)." Among



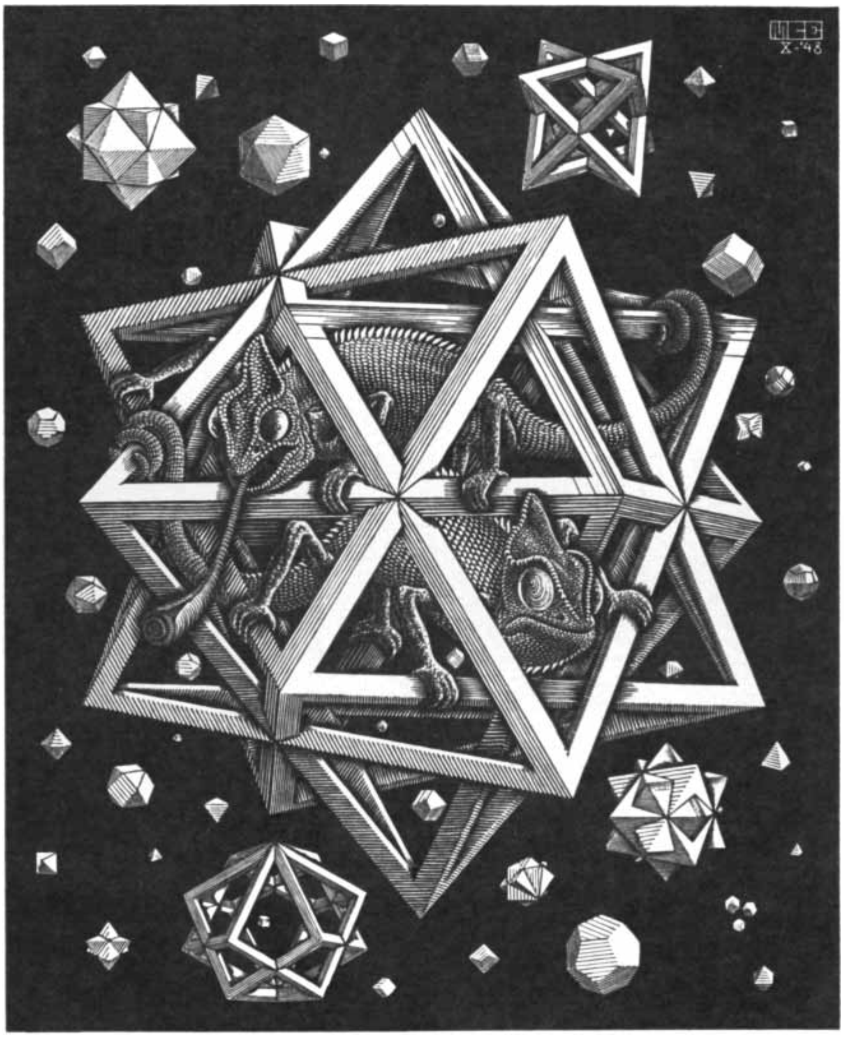
"LIBERATION," a lithograph designed in 1955, was classified by Escher under the heading "Development of Form and Contrast," in keeping with the technical terms of Harrower's analysis. This print is in Merriam collection at Boston Public Library.

scientists this aspect of Escher's graphic work is probably the best known.

Yet the origin of his compositions from playful manipulations of the figure-ground ambiguity has so far been noted only once before—by the art historian E. H. Gombrich in his article "Illusions and Visual Deadlock" (reprinted in his 1963 book *Meditations on a Hobby Horse*). This oversight is understandable, since Escher's later prints suggest mathematical prototypes as a primary source for his work. Such an interpretation is offered, for example, by Coxeter. In his essay "The Mathematical Implications of Escher's Prints" (reprinted in *The World of M. C. Escher*) Coxeter marvels at Escher's ability to extend the theory of crystallographic groups beyond Fedorov's original 17 by anticipating "through artistic intuition" the added principle of color symmetry.

Escher, however, was led to these extensions by his earlier sources from the psychological literature. Thus he knew how to combine both the figurative reversals and the crystallographic rules of regular and semiregular tessellations in one and the same composition on the flat picture surface. In *Reptiles* and in many other drawings he achieved such a feat. The fundamental region of a tessellation is a polygon (triangle, square or hexagon) or a combination of polygons; they must meet corner to corner. In *Reptiles* three heads, three elbows and three toes abut exactly at the corners of a hexagon, which forms the fundamental region of this regular tessellation on the plane. Escher looked at these solutions, some more difficult than others, with a great deal of pride.

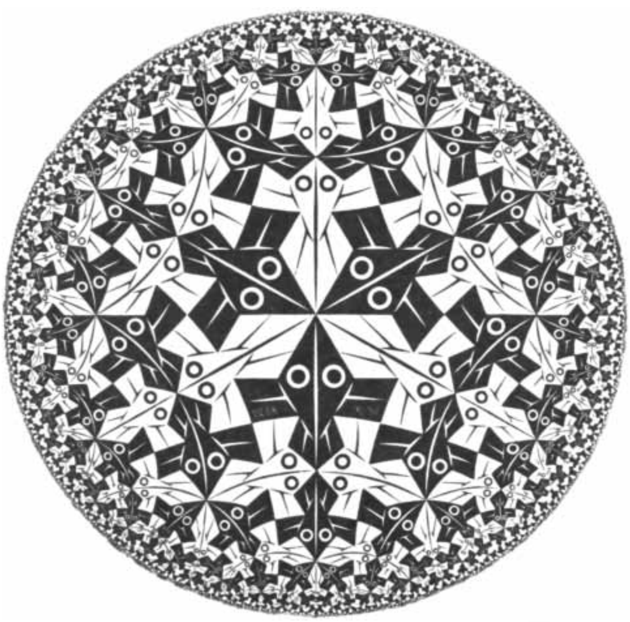
A similar close association between crystallographic principles and the design of densely packed surfaces was recognized by Paul Klee and later by Victor Vasarely. Both painters based certain pictures and diagrams on Johannes Kepler's humorous treatise *De Nive Sexangula (The Six-cornered Snowflake)*, published in 1611. Kepler's neo-Platonic concept of an underlying order or harmony—the belief in a mathematical structure of the universe—was shared by Escher. Occasionally one or another of his graphic works illustrates that idea, for example *Reptiles* or *Stars*, a 1948 wood engraving in the style of the early 17th century, Kepler's period [*see illustration on this page*]. This work depicts a star-studded sky in which the stellar bodies are composed of the Platonic solids cherished by Kepler. In such prints Escher intends to draw a contrast between the permanent laws of mathematics and the incidentals of debris or the changing colors of chameleons. "There is something in such laws that takes the breath away," Escher wrote in his essay "Approaches to Infinity." He continued: "They are not discoveries or inventions of the human mind but exist independently of us." Thus had Socrates explained the intrinsic beauty of geometric forms in Plato's *Philebus*. The abstract laws or principles of simplicity of form that attracted Escher to the perceptual analyses of the Gestalt psychologists were also essentially Platonic in concept.



"STARS," a 1948 wood engraving done in the style of the early 17th century, celebrates Escher's identification with Johannes Kepler's neo-Platonic belief in an underlying mathematical order in the universe. The print is in Roosevelt collection at the National Gallery.

Through his new interest in mathematics and contact with mathematicians, Escher expanded his vocabulary of ambiguous forms. He used the Möbius strip, the Klein bottle, knots and various forms of polygons. *Circle Limit I*, a hyperbolic (non-Euclidian) construction, was developed in 1958 in an exchange of letters with Coxeter [*see illustration at top left on opposite page*]. It gave Escher a chance to represent "the limits of infinite smallness," as he termed it. *Heaven and Hell*, done in 1960, belongs to the same series [*see illustration at top right on opposite page*].

In the 1950's Escher returned to sources from the psychology of visual perception in a group of prints dealing with reversible perspectives. The 1957 lithograph *Cube with Magic Ribbons* combines the reversible Necker cube (a discovery of the 19th-century Swiss mineralogist L. A. Necker) with the crater illusion [*see bottom illustration on page 100*]. In 1938 the Finnish psychologist Kai von Fieandt published a study on apparent changes in depth perception from concave to convex depending on different directions of light. For his experiments he used small knobs shaped



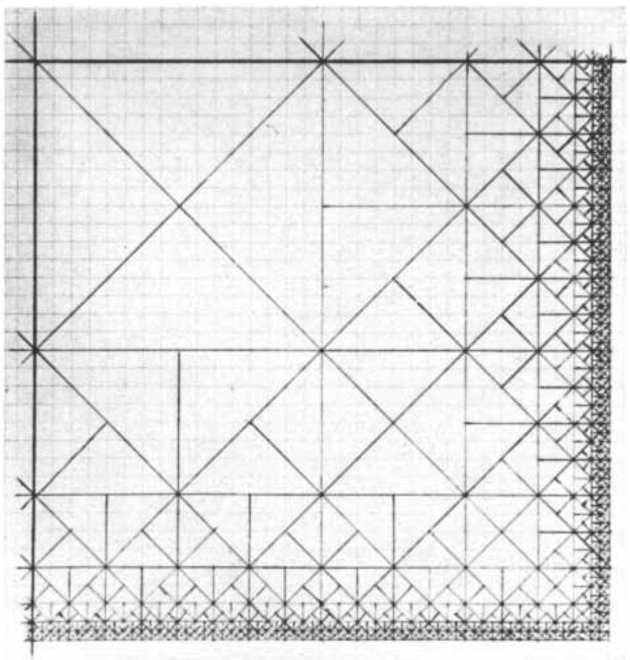
"CIRCLE LIMIT I," a woodcut designed by Escher in 1958, was based on a non-Euclidean mathematical construction developed in an exchange of letters with the mathematician H. S. M. Coxeter. The reproduction is made from a print in the Gemeentemuseum.

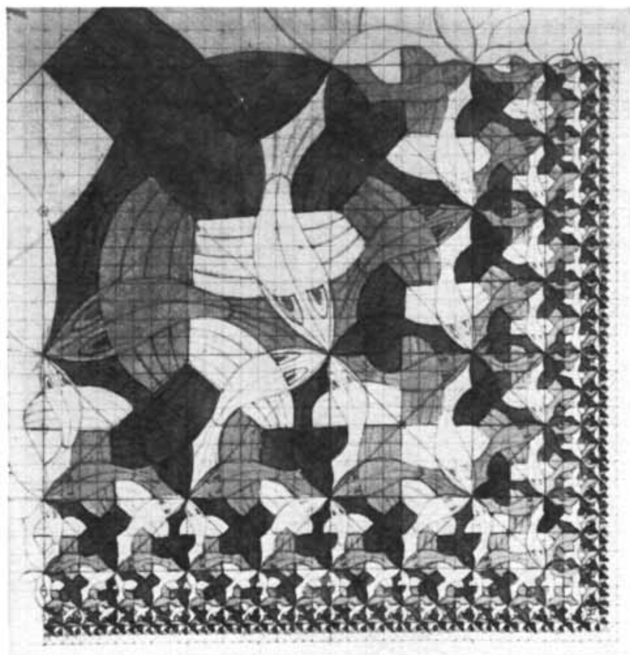
"HEAVEN AND HELL," a 1960 Escher woodcut in which the figure-ground ambiguity mirrors an ambiguity of meaning (good and evil), belongs to the same series of mathematically derived designs. The reproduction is from a print in the Gemeentemuseum.

just like those appearing on Escher's band [*see top illustration on next page*]. Escher must have known von Fieandt's experiments. The artist explains: "If we follow... the strips of buttonlike protuberances... with the eye, then these nodules surreptitiously change from convex to concave."

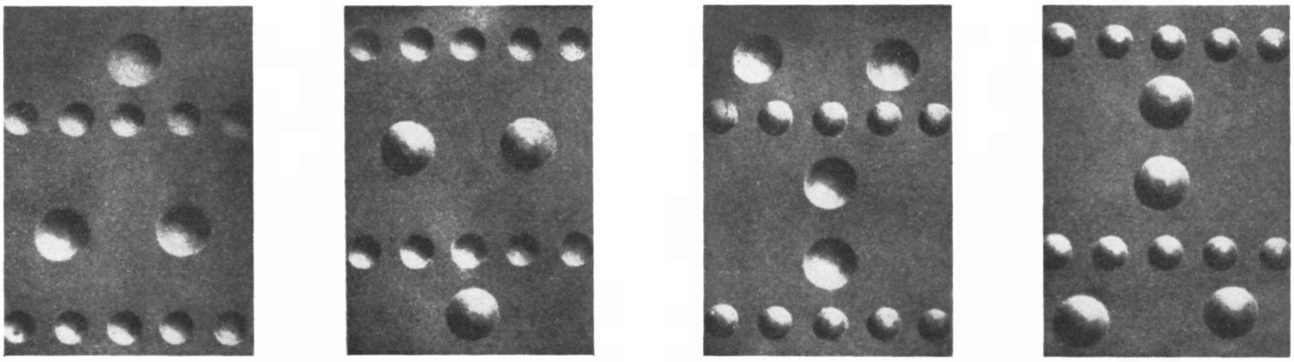
Concave and Convex [*see top illustration on page 101*] belongs to the same group of prints where reversible perspectives, or "inversions," as Escher called them, are the topic. The cluster of cubes on the flag announces the basic visual motif of the composition. In this 1955 lithograph Escher plays with the ambiguity of volumes on the flat picture plane; they switch from solid to hollow, from inward to outward, from roof to ceiling—like the symbol on the flag.

In 1958 Escher created *Belvedere,* an impossible building, also based on the reversible Necker cube [*see bottom illustration on page 102*]. By the end of the 19th century the Necker cube had become one of the most popular and most frequently debated optical illusions

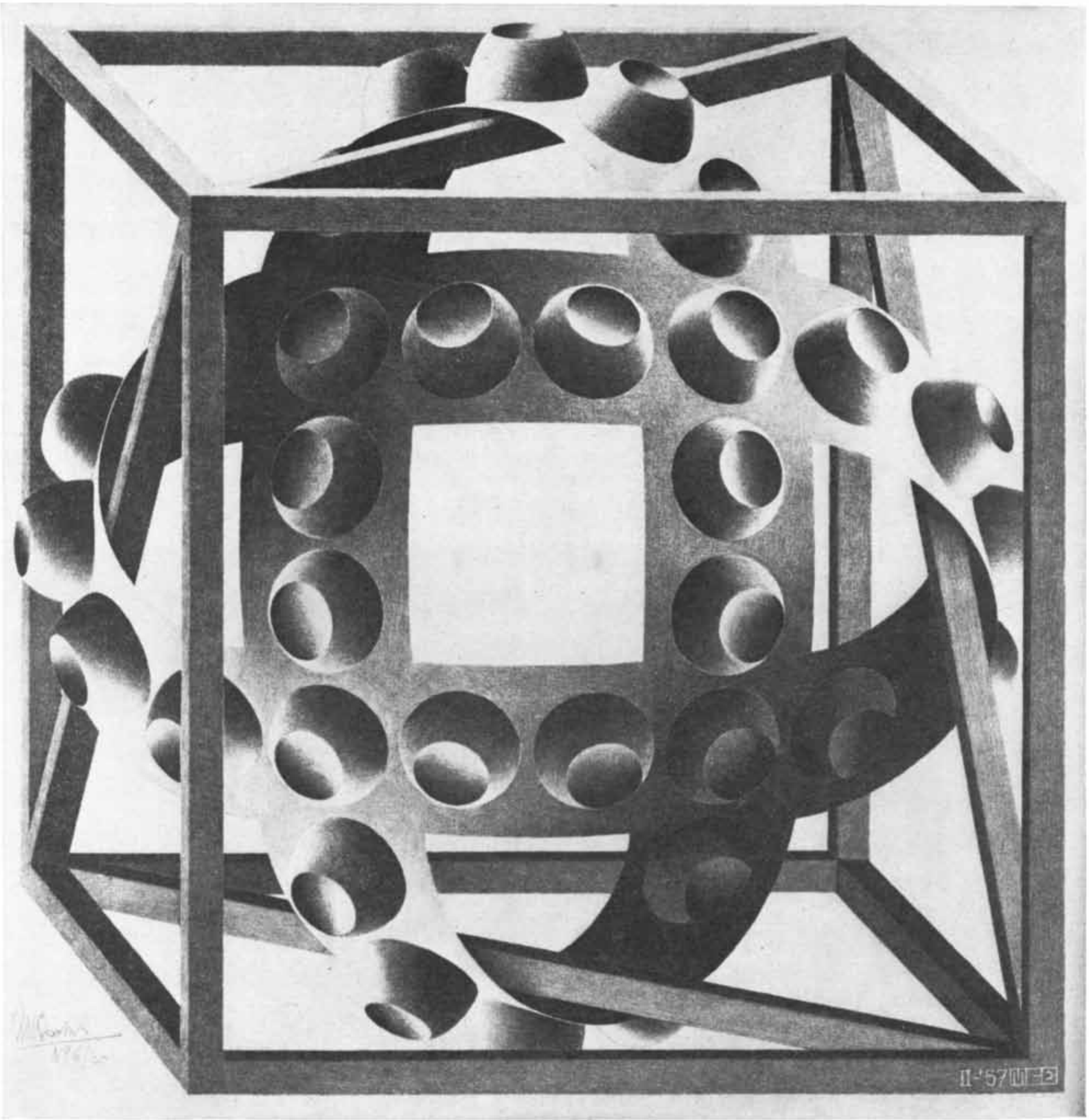




STUDIES for Escher's 1964 woodcut *Square Limit* reveal that this design was carried out by simply dividing surface after surface in half, up to the limit of visibility at the outer edge. This reproduction is from the original drawing in the Gemeentemuseum.



EXPERIMENTAL PATTERNS designed to study the crater illusion were published by the Finnish psychologist Kai von Fieandt in 1938. The patterns test the changes in depth perception from concave to convex, depending on the direction of the illumination.



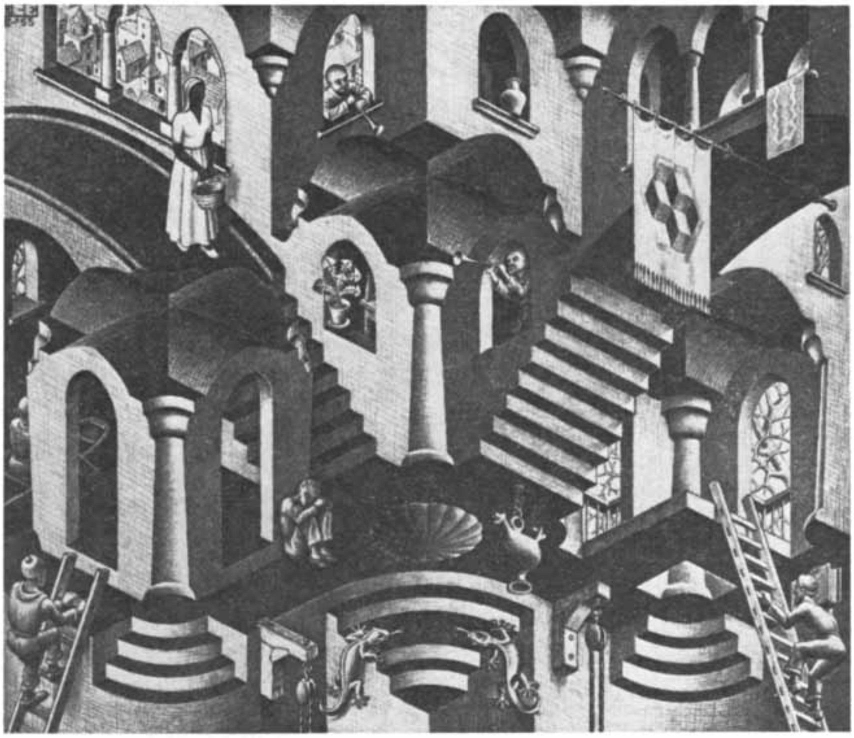
"CUBE WITH MAGIC RIBBONS," a 1957 Escher lithograph in the Roosevelt collection at the National Gallery, uses strips of small "buttonlike protuberances" just like those illustrated in the von Fieandt study to explore "inversions" of concave and convex.

in the psychological literature. To emphasize the theme of the fantastic piece of architecture, the boy on the bench contemplates the reversible cube in his hands and on paper. The corners that are flipping forward and backward during reversals are connected by diagonals, just as in Necker's original 1832 drawing [*see top illustration on next page*]. In *Belvedere,* however, Escher not only uses reversible perspective but also introduces perceptual impossibility, which obstructs the two perceptual interpretations of the cube simultaneously. This technique resembles the constructions of impossible figures published in 1958 by L. S. Penrose and R. Penrose in the *British Journal of Psychology,* acknowledged by Escher as a source for his 1960 lithograph *Ascending and Descending.*

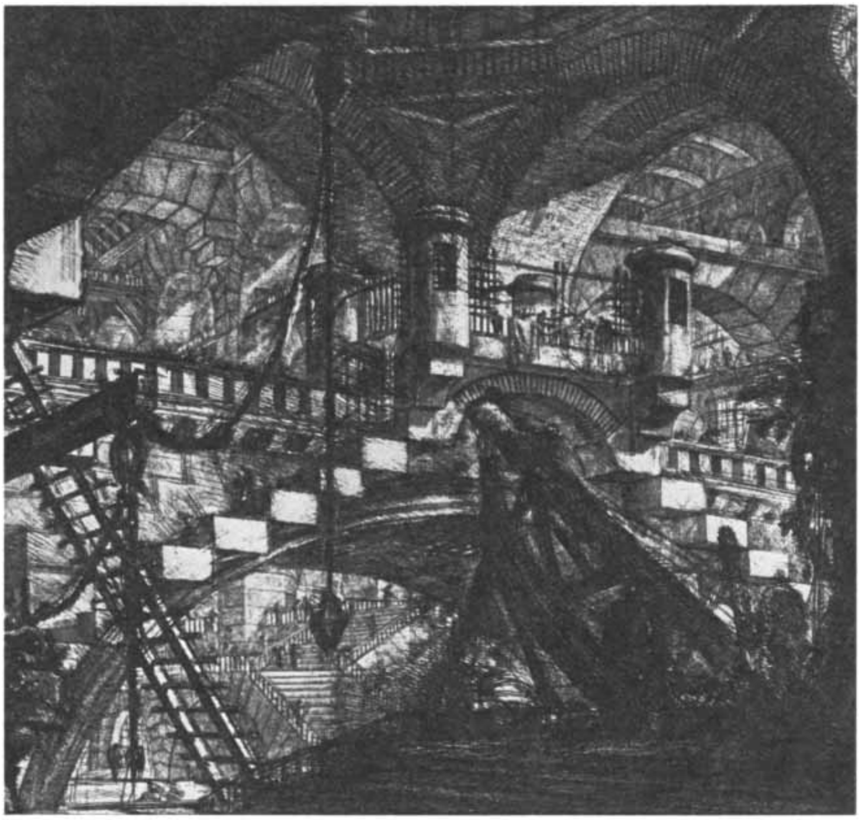
The Schröder stairs, another 19th-century reversible-perspective illusion, first published by H. Schröder in 1858 [*see top illustration on page 103*], is the theme of Escher's 1953 lithograph *Relativity* [*see middle illustration on page 103*]. The stairs show the characteristic shading that facilitates reversals, so that they look either like a staircase going up or an overhang of wall coming down. For the inhabitants of this structure the stairs lead up and down at the same time.

These compositions resemble certain 18th-century engravings, particularly Giovanni Battista Piranesi's *Carceri* (*Prisons*), with their obsessional repetitions and their shifting viewpoints that break up the unity of Renaissance perspective, thus giving a hallucinatory quality to these architectural dreams [*see bottom illustration at right*]. Note in the distance in the upper left quadrant of Piranesi's print a light-shaded underside of an arch. Or is it a walkway leading to a set of stairs? In *Concave and Convex* Escher employs the same motif in both orientations. Escher actually owned a set of Piranesi's engravings, according to his son, George A. Escher, who relates the following revealing story about his father and *Belvedere:*

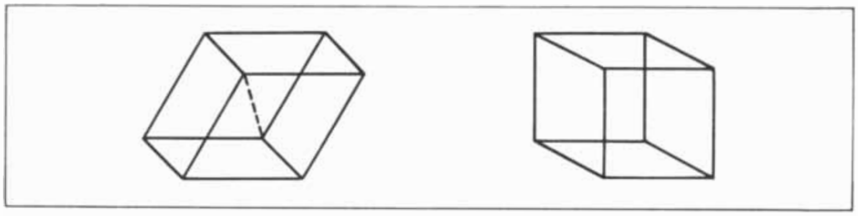
"One evening, it must have been late 1958, we were looking at the *Carceri* by Piranesi, which he greatly admired and of which he owned a posthumously printed set. We had been hunting for the many perspective aberrations of the same nature as occur in *Belvedere* and I asked him whether these had inspired him to make that print. No, he said, he had been aware of these oddities since long, but had always considered them as carelessness due to the reputed furious pace at which Piranesi had produced the prints during an illness. They had



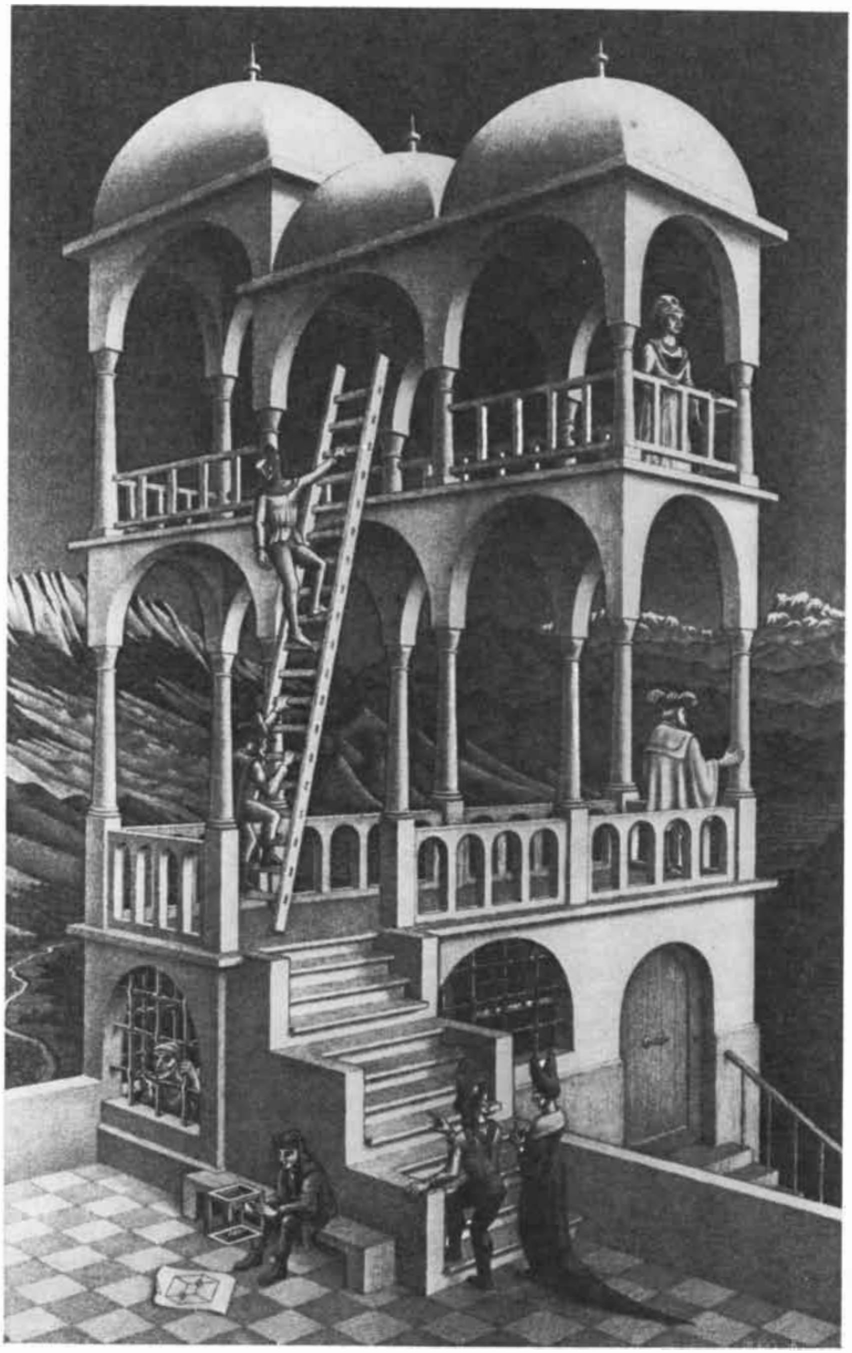
"CONCAVE AND CONVEX," a lithograph designed by Escher in 1955, also makes use of reversible perspectives to bring out the ambiguity of volumes portrayed on the flat picture plane. The original print is in the Rosenwald collection at the National Gallery.



EIGHTEENTH-CENTURY ENGRAVING by Giovanni Battista Piranesi, one of a series titled *Carceri* (*Prisons*), is distinguished by perspective aberrations of a type similar to those in Escher's prints. Escher actually owned a set of Piranesi's *Carceri,* but he discounted their influence on his own work, pointing out that he was much more inspired by experiments on visual perception. This detail of a print catalogued Plate XI, 2nd State is reproduced here by courtesy of the Museum of Fine Arts in Boston (gift of Miss Ellen Bullard).



NECKER CUBE, a famous reversible-perspective illusion, was first described in 1832 by the Swiss mineralogist L. A. Necker, who noticed the reversals in his drawings of crystals.



"BELVEDERE," a 1958 lithograph by Escher, is also based on the ambiguous geometry of the Necker cube. The reversible cube, in which the corners that flip forward and backward during reversals are connected by diagonals, appears in three different forms in this print: in the impossible architecture of the building itself, in the model held by the boy sitting on the bench in front of the building and in the drawing on the piece of paper lying on the floor. The reproduction is from a print in the Roosevelt collection at the National Gallery.

never awakened the particular twist of fantasy which gave birth to *Belvedere*. That, he said, was the direct consequence of noting somewhere . . . a picture of the reversible . . . cube."

Nothing could confirm more closely the essentials of Escher's art. As I have tried to show, the artist was fascinated by certain phenomena from experimental work on vision. These were the intellectual starting points for his inventions. Once gripped by one of his "visual ideas" he would spend sleepless nights, writes his son, "trying to bring some vague concept to clarity. . . . For weeks he would refuse to talk about what he was doing and lock his studio, whether he was there or not." The perspective displacements in Piranesi's *Carceri* or the ambiguities in the reversible patterns of the tiles of the Alhambra were exciting to him because he felt a kinship with these works reaching back over the centuries, but they were not his primary sources.

It is quite apparent that Escher's use of principles derived from the contemporary psychology of visual perception meant much more to him than a new set of themes or artistic techniques. Escher himself described the profound change that occurred in his style between 1936 and 1938 as if it had been the result of a religious conversion: "There came a moment when it seemed as though scales fell from my eyes . . . I became gripped by a desire the existence of which I had never suspected. Ideas came into my head quite unrelated to graphic art, notions which so fascinated me that I longed to communicate them to other people." It is no contradiction that this sudden revelation had been foreshadowed in Escher's much earlier application of Rubin's original ideas in the beginning of the 1920's. Artistic ideas, like scientific ideas, have a way of going underground only to reemerge with full force at a later stage.

Once gained, Escher's insights stayed with him into his final years. Even 30 years after his first contact with Koffka's work, when Escher was 70 years old, he expressed himself entirely in Koffka's terms (in his 1968 essay "Approaches to Infinity"): "No one can draw a line that is not a boundary line; every line splits a singularity into a plurality. Every closed contour, no matter what its shape, whether a perfect circle or an irregular random form, evokes in addition the notions of 'inside' and 'outside' and the suggestion of 'near' and 'far away,' of 'object' and 'background.' "

Here Escher refers not only to the

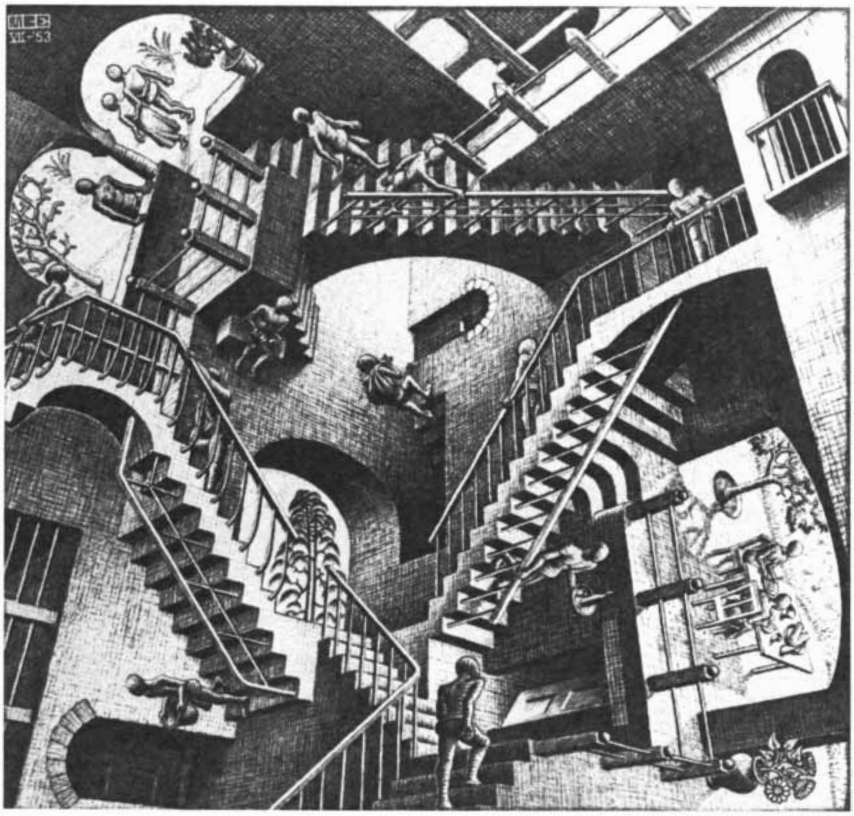
principle of protruding figure and receding ground, and the double function of contours, but also to well-known Gestalt investigations on closed-contour figures and the problem of "*unum* and *duo* organization" of surface subdivisions analyzed by Wertheimer and Koffka. Regular subdivisions, as in Escher's studies for *Square Limit* of 1964 [*see bottom illustration on page* 99], could be carried out by simply dividing surface after surface in half, up to the limit of visibility at the outer edge. Therefore Escher could say with Koffka: "Every line splits a singularity into a plurality."

The figure-ground ambiguity of Rubin, Koffka and Harrower provided the decisive impetus for Escher to give up the traditional Renaissance picture space. Although he held on to the recognizable image of beast or man for his reversible units, he arrived at a new—sometimes surreal—emphasis on the flat picture surface, a development that the great innovators of 20th-century art had reached at a much earlier date. Picasso and Braque painted their first Cubist pictures between 1907 and 1909; Kandinsky's first abstract color compositions date from 1911. The interdigitation of shapes and their symbolic interpretation in Escher's graphic work, however, can be traced to a trend antedating the modern movement, namely the flat positive and negative patterns of Art Nouveau—often equally charged with meaning and in vogue just before and after the turn of the century.

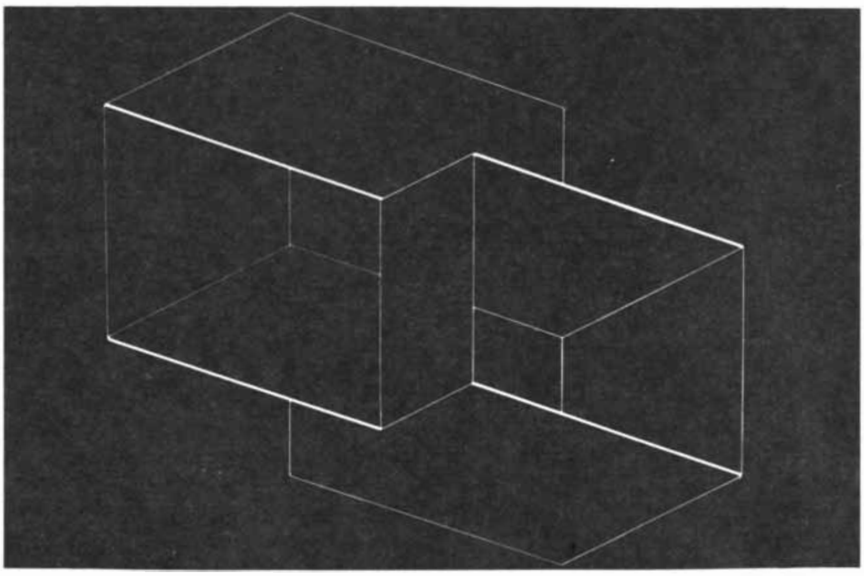
Escher's contact with the visual experiments of the Gestalt psychologists is not an isolated instance. Joseph Albers' striking constructions on laminated plastic and many of his drawings can be traced to similar prototypes from the psychological literature [*see bottom illustration at right*]. Such ambiguous forms had become a focus of renewed interest at the Bauhaus in Dessau, Germany, in 1929 and 1930, when lectures on Gestalt psychology were offered at this influential school of design. Albers, first a student and then a teacher at the Bauhaus from 1920 until its closing by the Nazis in 1933, became fascinated by reversible perspectives. Beginning in 1931, abstract reversible-line constructions have continued to fascinate him throughout his artistic career.

In the 1930's (almost contemporaneously with Escher) Albers and Vasarely, both precursors of the "Op art" movement (which Albers prefers to call "perceptual art"), created paintings and woodcuts displaying the ambiguity of figure and ground. Yet it is apparent that

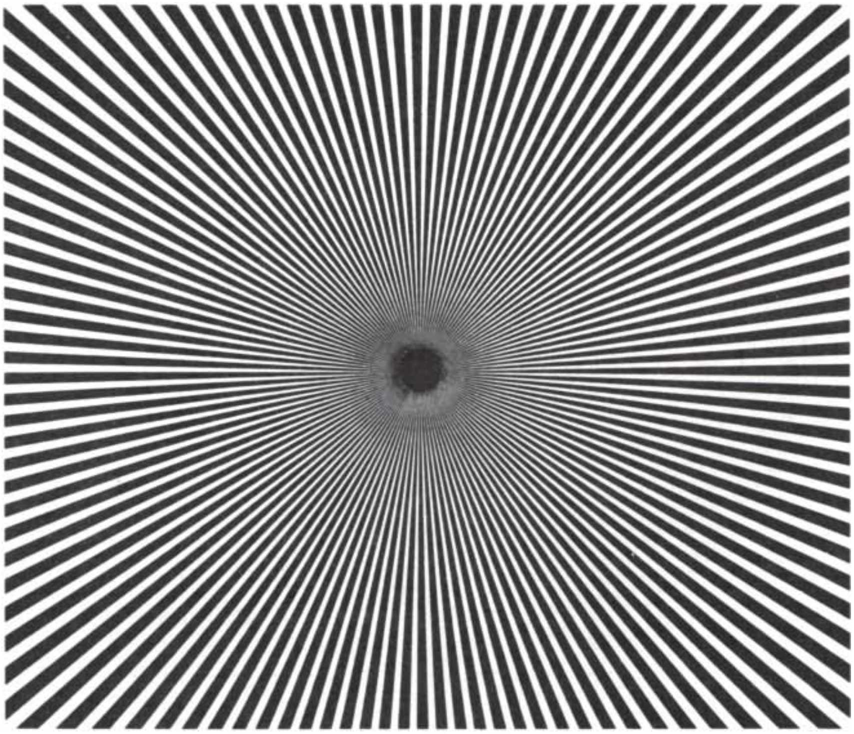
SCHRÖDER STAIRS, another 19th-century reversible-perspective diagram, was first published in 1858 by H. Schröder, who pointed out that the shading facilitates the reversals.



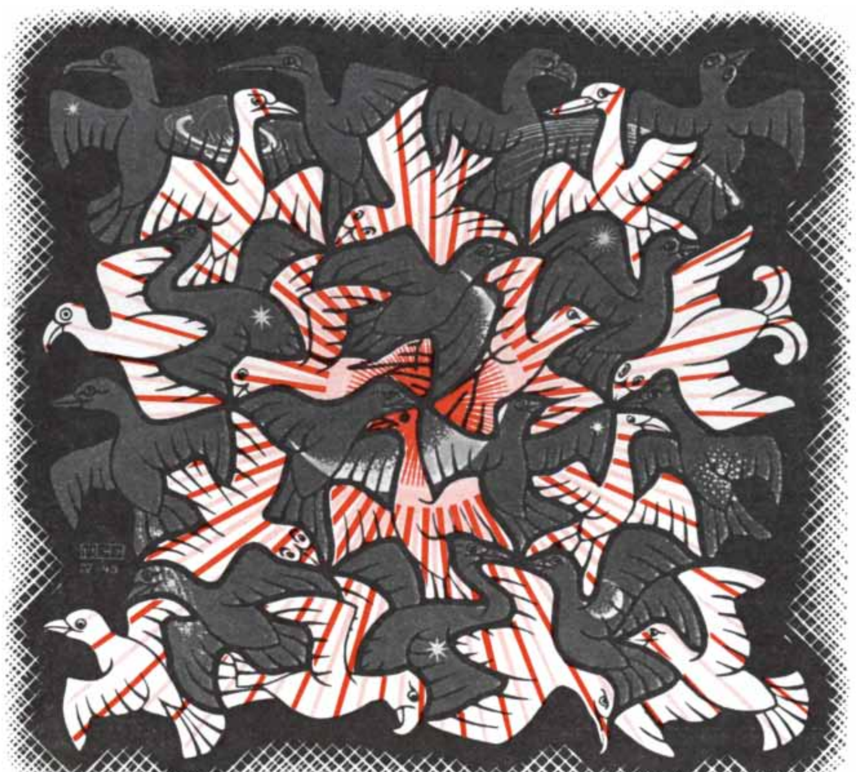
"RELATIVITY," a 1953 Escher lithograph, shows reversals similar to those in the diagram at top. The reproduction is from a print in the Rosenwald collection at the National Gallery.



"STRUCTURAL CONSTELLATION," a laminated-plastic construction by Josef Albers, can also be traced to certain reversible-perspective diagrams in the psychological literature.



"RAY PATTERN," published originally in *Nature* in 1957 by the British information theorist D. M. MacKay, influenced the work of the Op artists of the 1960's. A scintillating pattern of fine lines appears to run at right angles to the rays, indicating that a line-detecting mechanism of form perception complementary to the ray pattern has been stimulated.



"SUN AND MOON," a four-color woodcut designed by Escher in 1948, epitomizes, in the author's view, "the symbolist yearnings of the turn of the century with . . . demonstrations of ambiguity in the visual process." Reproduction is from a print in the Gemeentemuseum.

the intellectual stimulation provided by Gestalt theory manifests itself in very different ways, depending on the artist's choice and predisposition. Albers and Vasarely continue the abstract tradition of modern art by giving it a new direction through insights gained from investigations on vision and visual perception. Escher instead extended the decorative tradition of Art Nouveau coupled with the Symbolist movement. It is perhaps no accident that Art Nouveau patterns are similarly repetitive and crowd the flat surface, just as Escher's inventions do.

Uncovering Escher's sources does not diminish the fascination of his work. Indeed, it underscores how directly the awe we experience before his compositions derives from the perplexing ambiguity of his scientific prototypes. By employing motifs from contemporary attempts at the scientific analysis of form perception, the artist plays with stripped-down mechanisms of perception and reflects on his own visual means.

Similarly, the abstract perceptual artist of the 1960's reflects on the presumed functional property of our visual apparatus by making his patterns vibrate with repetitive line and color stimuli. It is remarkable that such art culminated at the very time when physiologists of the brain began to demonstrate mechanisms for primitive "feature detection" in the cerebral visual pathways [see "The Visual Cortex of the Brain," by D. H. Hubel; SCIENTIFIC AMERICAN, November, 1963]. As the British information theorist D. M. MacKay has pointed out, complementary mechanisms of form perception (similar to complementary color perception) play a role in these scintillating patterns [*see top illustration at left*]. These effects were adopted by the Op-art painters in their provocative arrays of lines. Without any other visual clues to guide us, these patterns make the feature-extracting machinery in the human visual system reverberate *in vacuo*.

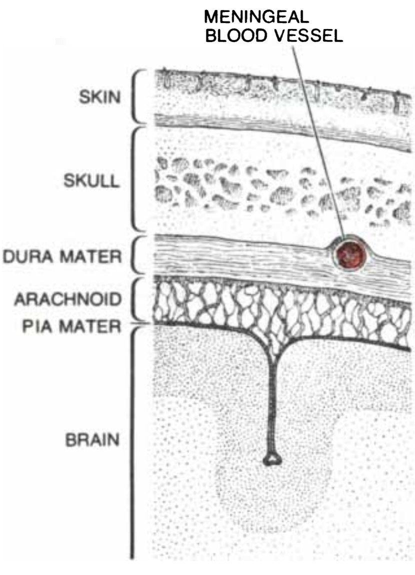
Escher instead clings tenaciously to meaningful, if fantastic, patterns and invites the viewer to repeat the basic figure-ground experiments of the Gestalt school. This can best be seen in the four-color woodcut *Sun and Moon,* which combines the Symbolist yearnings of the turn of the century with demonstrations of ambiguity in the perceptual process [*see bottom illustration at left*]. If you focus on the light birds, the crescent of the moon appears and night prevails; if you focus on the dark birds, the sun will shine.

THE CASTS OF FOSSIL HOMINID BRAINS

The skulls of man and his precursors can be used as molds to make replicas of the brain. These casts indicate that man's brain began to differ from that of other primates some three million years ago

by Ralph L. Holloway

Man is not the largest of the primates (the gorilla is larger), but he has the largest brain. How did this come about? The question is hardly a new one, but a considerable amount of new evidence is now available to those in search of the answer. In brief the evidence suggests that, contrary to what is widely believed, the human brain was not among the last human organs to evolve but among the first. Neurologically speaking, brains whose organization was essentially human were already in existence some three million years ago.



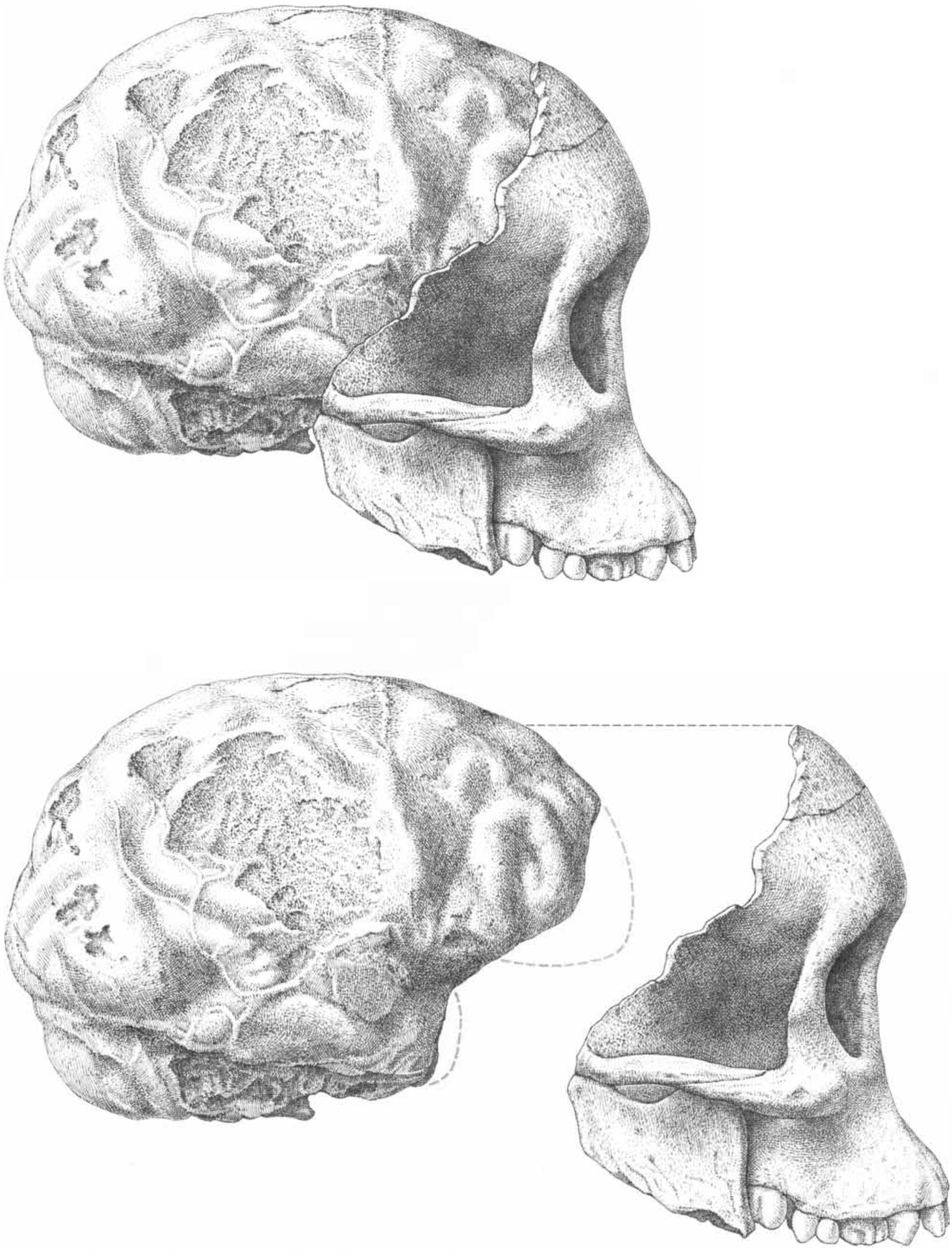
LAYERS OF TISSUE lie between the surface of the brain and the cranium. They include the pia mater, the arachnoid tissue that contains the cerebrospinal fluid, and the thick dura mater. Meningeal blood vessels (*color*) often leave a clear imprint on the inner surface of the cranium. The sutures of the cranial bones may also show, but among higher primates the convolutions of the cerebral cortex are not often apparent.

The brain is the most complex organ in the primate body or, if one prefers, the most complex set of interacting organs. It consists of a very large number of interconnected nuclei and fiber systems, and its cells number in the billions. With certain exceptions, however, the cortex, or upper layer, of one primate brain exhibits much the same gross morphology as the cortex of another, regardless of the animals' relative taxonomic status. Whether the animal is a prosimian (for example a lemur), a monkey, a pongid (an ape) or a hominid (a member of the family that includes the genus *Homo*), what varies from one primate brain to another is not so much the appearance of the cortex as the fraction of its total area that is devoted to each of the major cortical subdivisions. For instance, whereas the chimpanzee is taxonomically man's closest living primate relative and the brains of both look much alike superficially, the chimpanzee brain is significantly different from the human brain in the relative size and shape of its frontal, temporal, parietal and occipital lobes. The differences are reflected in the different position of the sulci, or furrows, that mark the boundaries between lobes, and in the differently shaped gyri, or convolutions, of the lobes themselves.

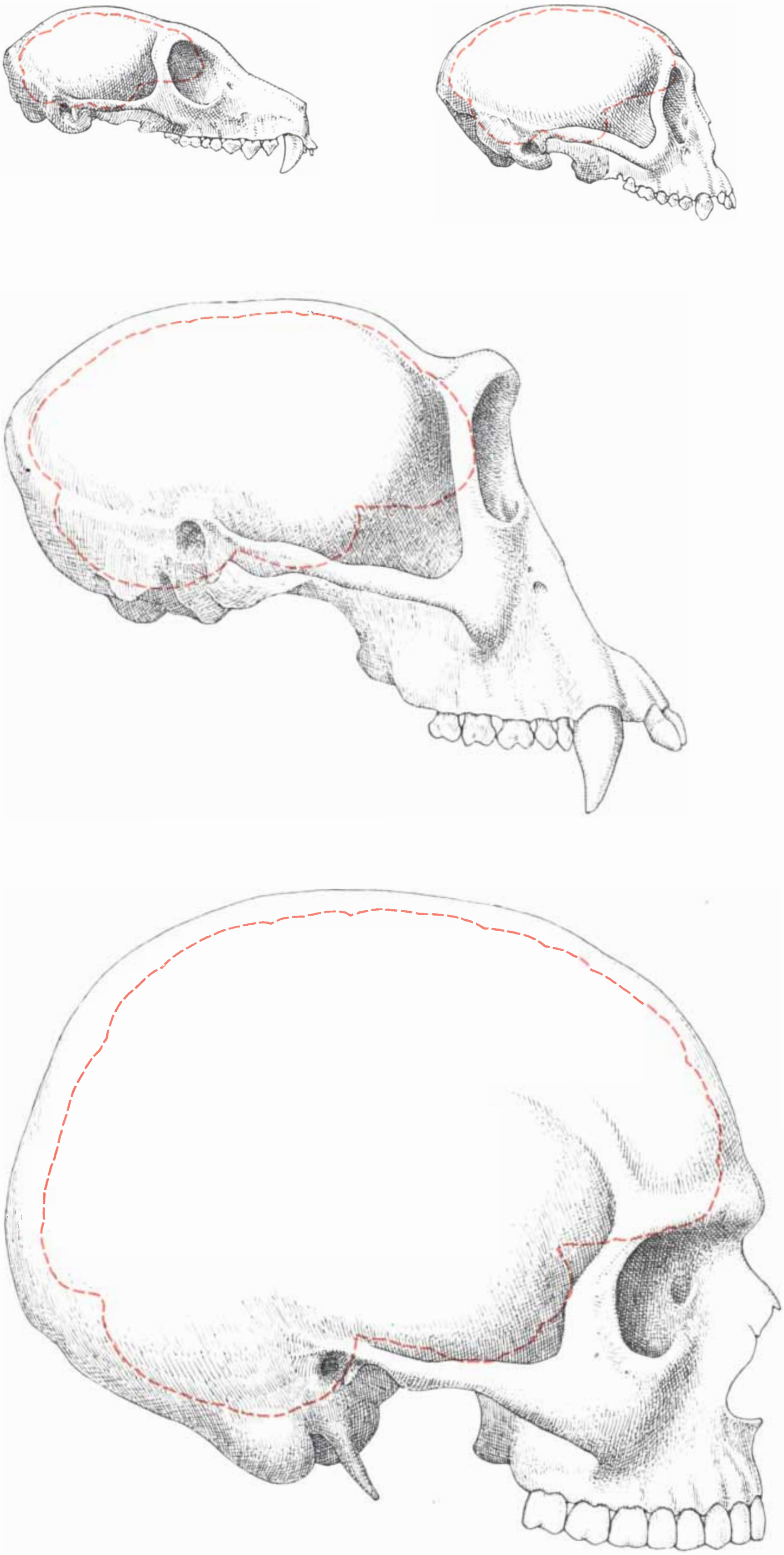
The comparative neurology of living primates can of course cast only the most indirect light on human brain evolution. After all, the living representatives of each primate line have reached their own separate evolutionary pinnacle; their brains do not "recapitulate" the evolutionary pathway that man has followed. The neurological evidence is nonetheless invaluable in another respect. It is the only source of information linking aspects of gross brain morphology with aspects of behavior. The results of a great many primate studies allow some valid generalizations on this subject. For example, it is well established that the occipital lobe of the cortex, at the back of the brain, is involved in vision. The parietal lobe is involved in sensory integration and association, the frontal lobe in motor behavior and the more complex aspects of adaptive behavior, and the temporal lobe in memory. It is the interactions among these gross cortical divisions, and also among the subcortical nuclei and fiber tracts, that organize coordinated behavior.

It is also a valid generalization to say that a gross brain morphology which emphasizes relatively small temporal and parietal lobes and a relatively large area of occipital cortex is neurologically organized in the pongid mode and is thus representative of the apes' line of evolutionary advance. Conversely, a gross morphology that emphasizes a reduced area of occipital cortex, particularly toward the sides of the brain, and an enlarged parietal and temporal cortex is hominid in its neurological organization. It follows that any evidence on the neurological organization of early primates, including hominids and putative hominids, is of much importance in tracing the evolution of the human brain.

Such evidence is of three kinds: direct, indirect and inferential. The only direct evidence comes from the study of endocranial casts, that is, either a chance impression of a skull interior that is preserved in fossil form or a contemporary man-made replica of the interior of a fossil skull. The indirect evidence is of two kinds. The first is a by-product of endocranial studies; it consists of the conclusions that can be drawn from a comparison of brain sizes. To draw conclusions of this kind, however, can involve some degree of acceptance of the



TAUNG JUVENILE, the first specimen of *Australopithecus* to be unearthed, is shown in the top drawing with a portion of the fossilized skull (including the facial bones, the upper jaw and a part of the lower jaw) in place on the natural cast of its brain. The cast is seen separately in the bottom drawing; parts of the frontal and temporal lobes that were not preserved are indicated. The estimated brain volume of an adult of this lineage, once calculated at about 525 cubic centimeters, has now been scaled down to 440 c.c.



INCREASE IN BRAIN SIZE among primates is apparent in the different dimensions of the skull of a prosimian (*top left*), a New World monkey (*top right*), a great ape (*second from bottom*) and modern man (*bottom*). The brains (*color*) of the latter three are illustrated on the opposite page; skulls are reproduced at approximately half actual size.

questionable premise that there is a correlation between brain size and behavioral capacity. The second kind of indirect evidence comes from the study of other fossil remains: primate hand and foot bones, limb bones, pelvises, vertebral components, jaws and teeth. From these noncranial parts conclusions can be drawn about body size and behavioral capabilities such as bipedal locomotion, upright posture, manual dexterity and even mastication. Various patterns of musculoskeletal organization, of course, reflect matching variations in neurological organization.

Finally there is inferential evidence, particularly with respect to early hominid evolution. This can be described as fossilized behavior. The inferential evidence includes stone tools that exhibit various degrees of standardization (suggesting "cultural norms") and bone debris that reveals what animals the hominids selected as their prey. Any evidence of activity provides some grounds for inference about the general state of the individuals' neurological organization. Our concern here, however, will be with the direct endocranial evidence and with a part of the indirect evidence.

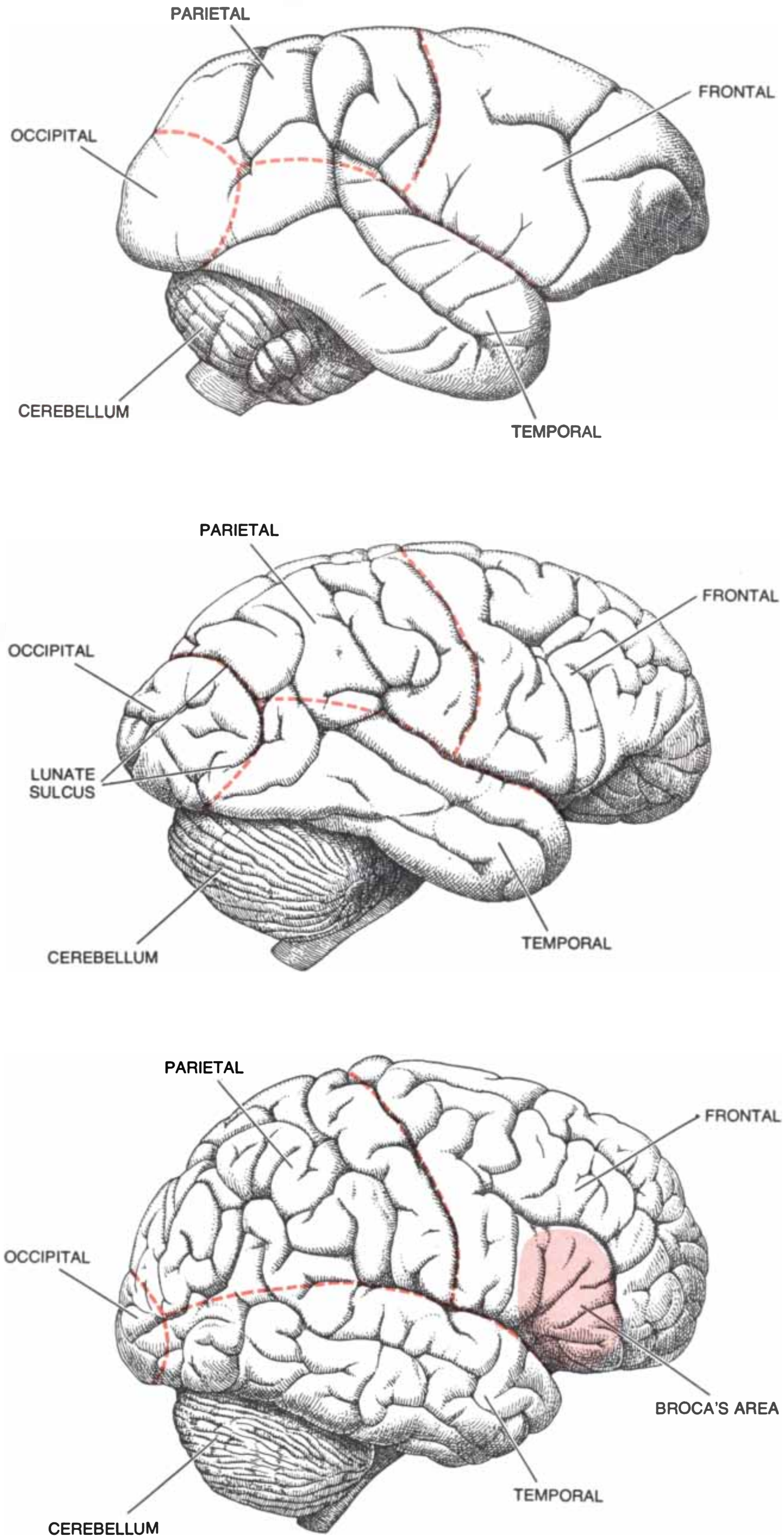
The convolutions of the cerebral cortex and its boundary furrows are kept from leaving precise impressions on the internal bony table of the skull by the brain's surrounding bath of cerebrospinal fluid and by such protective tissues as the pia mater, the arachnoid tissue and the dura mater, or outer envelope. In mammals the extent of masking varies from species to species and with the size and age of the individual. The skull interiors of apes and men, both living and fossil, are notable for bearing only a minimal impression of the brain surface. In virtually all cases the only detailed features that can be traced on the endocranial cast of a higher primate are the paths of the meningeal blood vessels. Depending on the fossil's state of preservation, however, even a relatively featureless cast will reveal at least the general proportions and shape of the brain. This kind of information can be indicative of a pongid neurological organization or a hominid one.

Just how far back in time can the distinction between pongid and hominid brains be pursued? There is a barrier represented by the key Miocene primate fossil *Ramapithecus:* no skull of the animal has been discovered. *Ramapithecus* flourished some 12 million to 15 million years ago both in Africa and in Asia. Elwyn L. Simons and David Pilbeam of Yale University have proposed that it is a hominid, but its only known remains are teeth and fragments of jaws. For the present we must make do with more numerous, although still rare, primate skulls that are many millions of years more recent.

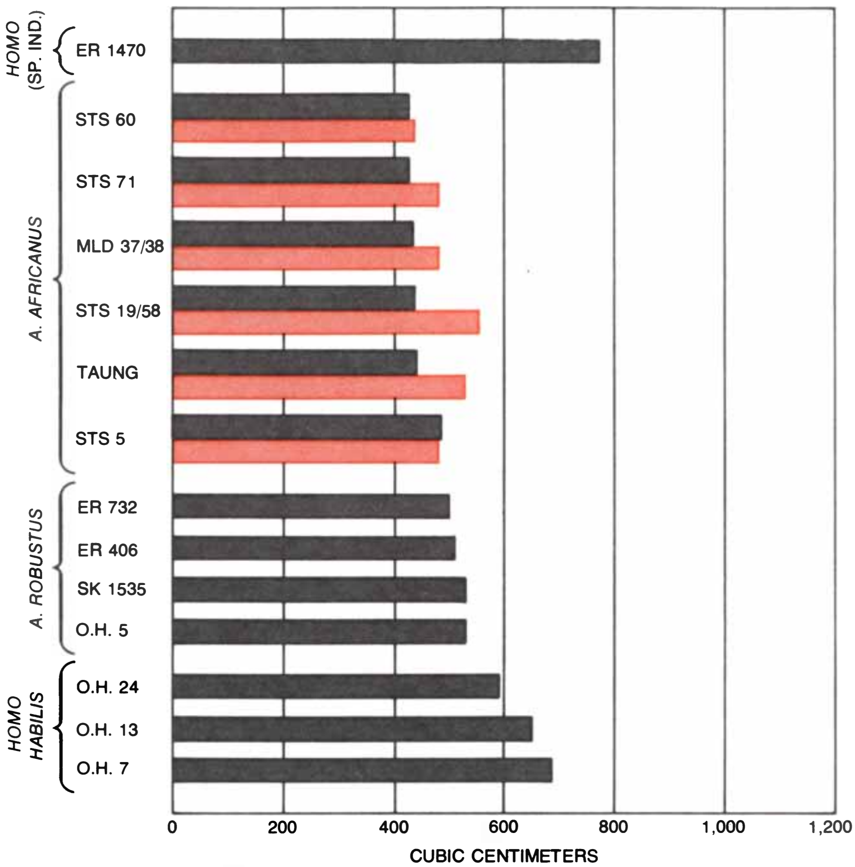
The first of these skulls became known in 1925, when Raymond A. Dart of the University of Witwatersrand described the fossilized remains of a subadult primate that had been discovered in a limestone quarry at Taung in South Africa. The find consisted of a broken lower jaw, an upper jaw, facial bones, a partial cranium and a natural endocranial cast [*see illustrations on page 107*]. In the half-century since Dart named the specimen *Australopithecus*, or "southern ape," seven other skulls of *Australopithecus* (six of them internally measurable) have been found in South Africa and from three to six more in East Africa.

"Three to six" refers not to any uncertainty about how many skulls have been found but to how they are to be assigned to one or another genus or species of hominid. For example, the genus *Australopithecus* consists of two species: *A. africanus*, the "gracile," or lightly built, form to which the Taung fossils belong, and *A. robustus*, a larger, more heavily built species [*see illustration on page 113*]. Of the eight South African skulls, six are gracile, one is robust and the eighth (a specimen from Makapansgat designated MLD 1) is not definitely assigned to either species. In the same way three of the East African skulls are unanimously assigned to the robust species of *Australopithecus*. The other three, designated *Homo habilis* by their discoverers, Louis and Mary Leakey, are classed with the gracile species by some students of the subject, but they are still generally known by the name the Leakeys gave them. As will become apparent, there are grounds for preserving the distinction.

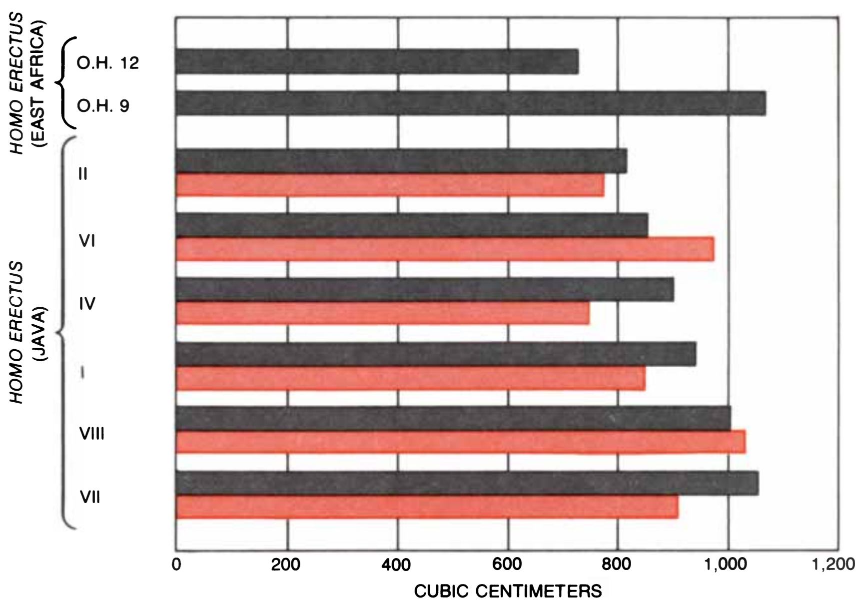
Fortunately, no matter what controversy may surround the question of how these early African hominids are related to one another, it has very little bearing on the question of their neurological development. The reason is that in each instance where an endocast is available, whether the skull is less than a million years old or more than two million years old, the brain shows the distinctive pattern of hominid neurological organization. Let us review the cortical landmarks that distinguish between the pongids and the hominids. Starting with the frontal lobe, the *Australopithecus* endocasts show a more hominid pattern in the third inferior frontal gyrus, being larger and more convoluted than endo-



GROSS DIFFERENCES in the neurological organization of three primate brains are apparent in the size of the cerebral components of a ceboid monkey (*top*), a chimpanzee (*middle*) and modern man (*bottom*). The small occipital lobe and the large parietal and temporal lobes in man, compared with the other primates, typify the hominid pattern. Lunate sulcus, or furrow (*color*), on the chimpanzee's brain bounds its large occipital lobe.



ABSOLUTE BRAIN SIZES varied both among and between the four earliest African hominid species. The 14 specimens are shown here in order of increasing brain volume, except for the earliest of all: the three-million-year-old specimen from East Rudolf, ER 1470, at the head of the list. The other abbreviations stand for Sterkfontein (STS), Makapansgat (MLD), Olduvai hominid (O.H.) and Swartkrans (SK). Double bars show former (*color*) and present calculated brain sizes of six specimens; the Taung volumes are adult values.



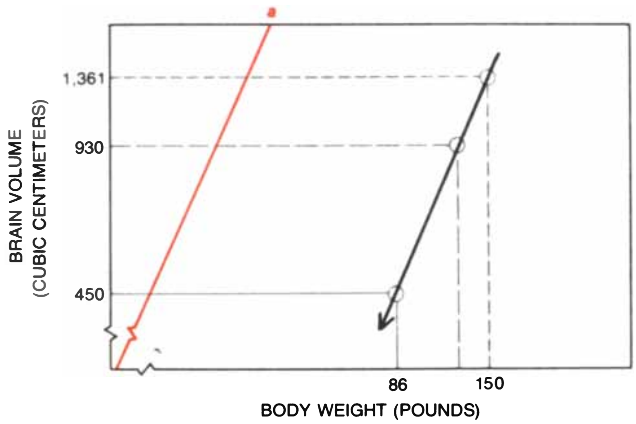
LATER HOMINID'S BRAINS were larger both on the average and absolutely, with the exception of one doubtful specimen from Olduvai Gorge. Shown here are the brain volumes for the seven certain specimens and one doubtful specimen of *Homo erectus;* two are from East Africa and six are from Java. Double bars show former (*color*) and present calculated brain sizes of the Java specimens. Casts of other *H. erectus* brains are not available.

casts of pongid brains of the same size. The orbital surface of their frontal lobes displays a typically human morphology rather than being marked by the slender, forward-pointing olfactory rostrum of the pongid. The height of the brain, from the anterior tips of the temporal lobes to the summit of the cerebral cortex, is proportionately greater than it is in pongids, suggesting an expansion of the parietal and temporal lobes and a more "flexed" cranium. Moreover, the temporal lobes and particularly their anterior tips show the hominid configuration and not the pongid one.

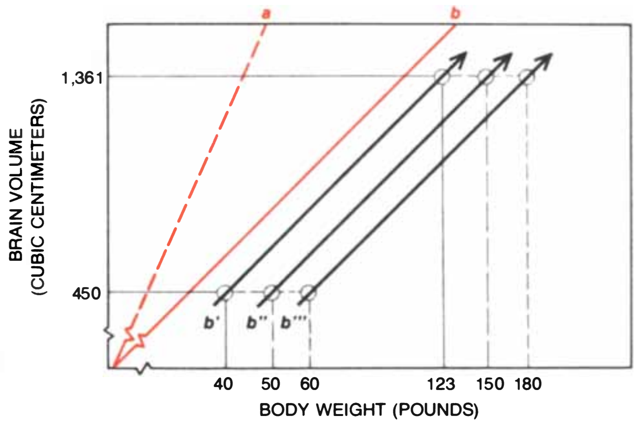
Another landmark of neurological organization provides further evidence, in part negative. This is the lunate sulcus, the furrow that defines the boundary between the occipital cortex and the adjacent parietal cortex. In all ape brains the lunate sulcus lies relatively far forward on the ascending curve of the back of the brain. The position is indicative of an enlarged occipital lobe. In modern man, when the lunate sulcus appears at all (which is in fewer than 10 percent of cases), it lies much closer to the far end of the occipital pole. On those *Australopithecus* endocasts where the feature can be located, the lunate sulcus is found in the human position, which indicates that the (perhaps associative) parietal lobes of the *Australopithecus* brain were enlarged far beyond what is the pongid norm.

It may seem surprising that *Australopithecus,* a genus that has only in recent years been granted hominid status on other anatomical grounds, should have an essentially human brain. It has certainly been a surprise to those who view the human brain as a comparatively recent product of evolution. The finding is not, however, the only surprise of this kind. The most remarkable new fossil primate discovery in Africa is the skull known formally as ER 1470, found by Richard Leakey and his colleagues in the region east of Lake Rudolf in Kenya in 1972. The fossil is nearly three million years old.

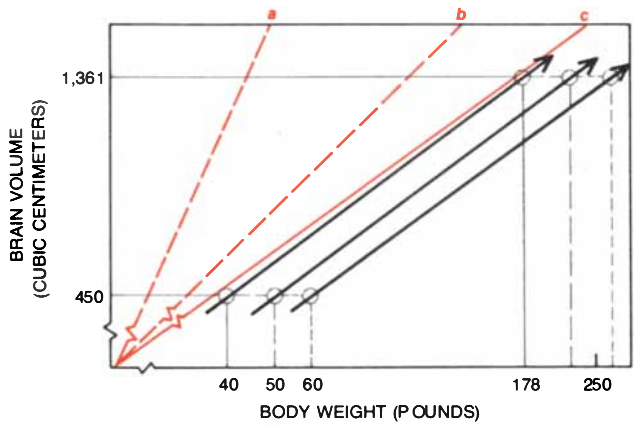
Through the courtesy of its discoverer I recently made an endocranial cast of ER 1470. Two facts were immediately apparent. Not only had the skull contained a brain substantially larger than the brain of either the gracile or the robust species of *Australopithecus* (and that of *Homo habilis* too) but also this very ancient and relatively large brain was essentially human in neurological organization. Leakey's find pushes the history of hominid brain evolution back in time at least as far as the shadowy



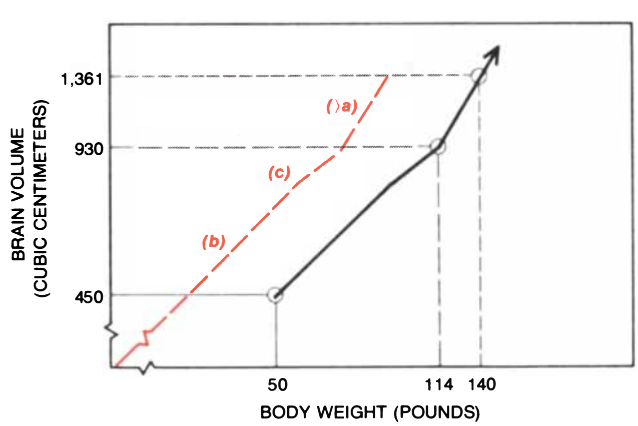
ALLOMETRIC GROWTH EQUATIONS with different exponential slopes are tested against known data in this group of graphs. A known average body weight of 150 pounds for modern man, combined with a known average brain volume of 1,361 cubic centimeters, is tested against a slope of 1.9 (*a, color*); the slope appears as a straight line on the double-log plot. At a brain volume of 450 c.c., the *Australopithecus* average, a projected body weight of 86 pounds is heavier than estimates of *Australopithecus* weight allow.



LESS RADICAL SLOPE, 1.0 (*b, color*), is tested in this graph against three different estimates of the average body weight of *Australopithecus:* 40 pounds, 50 pounds and 60 pounds. Where the parallel exponential slopes intersect the average cranial capacity for modern man (*b′, b″, b‴*) only the 50-pound body weight estimated for *Australopithecus* yields a projected body weight for man that agrees with the known average. An *Australopithecus* body weight much above or below 50 pounds thus appears improbable.



STILL A THIRD SLOPE, .66 (*c, color*), which is usually the most favorable allometric rate for mammals, is tested against the same three *Australopithecus* body-weight estimates. The .66 slope proves clearly unsuited to hominids. When even the minimum weight for *Australopithecus* is projected, the predicted weight for modern man is excessive; added weight yields even more grotesque results.



COMBINATION OF SLOPES tests the assumption that hominid rates of growth differed at different times. The zigzag predicts a plausible 114-pound body weight for *Homo erectus* with an average-size brain (930 c.c.) and a body weight for modern man not far below average. The implication is that rates of growth probably did vary, thereby varying selection pressures for changes in brain size.

boundary between the Pliocene and Pleistocene eras.

To summarize the direct evidence of the endocasts, it is now clear that primates with essentially human brains existed some three million years ago. Let us go on to see what details, if any, the indirect evidence of brain size can add to this finding. Here it is necessary, however, to state an important qualification. So far as modern man is concerned, at least, no discernible association between brain size and behavior can be demonstrated. On the average, to be sure, *Homo sapiens* is a big-brained primate. What is sometimes forgotten, however, is that the human average embraces some remarkable extremes. The "normal" sapient range is from 1,200 cubic centimeters to 1,800, but nonpathological brains that measure below 1,000 c.c. and above 2,000 are not uncommon. Moreover, this range of more than 1,000 c.c. is in itself greater than the average difference in brain size between *Australopithecus* and *H. sapiens.*

As to how such variations may affect behavior, it is probably sufficient to note that the brains of Jonathan Swift and Ivan Turgenev exceeded 2,000 c.c. in volume, whereas Anatole France made do with 1,000 c.c. Clearly the size of the human brain is of less importance than its neurological organization. This conclusion is supported by studies of pathology. For example, microcephaly is a disease characterized by the association of a body of normal size with an abnormally small brain. The brain may have a volume of 600 c.c., which is smaller than some gorilla brains. (The gorilla average is 498 c.c., and brains almost 200 c.c. larger have been measured.) Humans with microcephaly are quite subnormal in intelligence, but they still show specifically human behavioral patterns, including the capacity to learn language symbols and to utilize them.

I have recently made complete or partial endocasts of 15 early fossil hominids from South and East Africa. These

SQUIRREL MONKEY	1:12
PORPOISE	1:38
HOUSE MOUSE	1:40
TREE SHREW	1:40
MODERN MAN	1:45
MACAQUE	1:170
GORILLA	1:200
ELEPHANT	1:600
BLUE WHALE	1:10,000

WEIGHT OF THE BRAIN, expressed as a proportion of the total body weight, varies widely among mammals. In spite of his large brain modern man is far from showing the highest proportional brain weight. The proportional weight for man, however, is greater than that of other higher primates. So too, it appears, was the proportional weight for the hominid predecessors of modern man.

endocasts, together with natural ones, have allowed me to calculate the specimens' brain size [*see top illustration on page 110*]. Leaving aside for the moment two specimens from Olduvai Gorge (one of them certainly *Homo erectus* and the other possibly so), my findings are as follows. First, the brains of most of the South African specimens were substantially smaller than had previously been calculated. Of the six gracile specimens of *Australopithecus* from South Africa none had brains that exceeded 500 c.c. in volume, and most were well below 450 c.c. In contrast, none of the four specimens of the robust *Australopithecus* species had a brain volume of less than 500 c.c. Moreover, two of the four, the one from South Africa and one from East Africa, had brains measuring 530 c.c. Of the three representatives of *Homo habilis*, all from Olduvai, the one with the smallest brain was hominid No. 24, with a possibly overestimated cranial capacity of 590 c.c. The brains of the other two respectively measured 650 c.c. and 687 c.c.

Now, one reason—perhaps the main reason—students of human evolution were so slow to accept *Australopithecus africanus* as a hominid, even though Dart had emphasized the manlike appearance of the Taung endocast from the first, was that its estimated cranial capacities were so small. The brains of some apes, particularly the brains of gorillas, were known to be larger. In fact, as we now know, the brains of the gracile species of *Australopithecus* are even smaller than was originally thought. Since it is now also evident that the *Australopithecus* brains were essentially human in neurological organization, what are we to make of their surprisingly small size? The answer, it seems to me, is that in all likelihood the size of *Australopithecus'* brain bore the same proportional relation to the size of its body that modern man's brain does to his body.

This contention cannot be proved beyond doubt on the basis of the *Australopithecus* fossils known today. One can estimate body weight on the basis of known height, but the height of both the gracile and the robust species must also be estimated, on the basis of an imperfect sample of limb bones. Even though guesswork is involved, however, it is still possible to estimate various body weights and relate each of these guesses to the species' known brain size. One can then see how the results compare with the known brain-to-body ratio among the other mammals, particularly the primates.

The exponential relation between variations in overall body size and in the size of a specific organ, known technically as allometry, has been studied for nearly a century, so that the brain-to-body ratio is well known for a number of animals [*see illustration on this page*]. The brain-to-body ratio of modern man is not first among mammals; various marine mammals, led by the porpoise *Tursiops*, rank higher than man. So does one small primate, the ceboid squirrel monkey. However, among the hominoids, that is, the subdivision of primates that includes both the apes and man, modern man's brain-to-body ratio does rank first. Depending to some extent on what weight estimate one accepts for *Australopithecus*, this early hominid appears to have enjoyed a position similar to modern man's.

Thanks to Heinz Stephan and his colleagues at the Max Planck Institute for Brain Research in Frankfurt, an assessment of the relation between brain size and body size has become available that is subtler than simple allometry. Stephan's group has been collecting quantitative data with respect to animal brains for many years. The size and weight of various parts of the brain are measured, and these measurements are related to similar measurements of the animals' entire brain and body. One outcome of the work has been the development of what Stephan calls a progression index. It is the ratio between an animal's brain weight and what the brain weight would have been if the animal had belonged to another species with the same body dimensions. As their standard in this comparison Stephan and his colleagues use the brain-to-body ratio of "basal insectivores" such as the tree shrew, which are representative of the original stock from which all the primates evolved.

When the progression index is calculated for modern man, assuming an average body weight of 150 pounds and an average cranial capacity of 1,361 c.c., the resulting index value is 28.8. If different body weights are used with the same average cranial capacity and vice versa, the range of progression indexes for modern man extends from a minimum of 19.0 to a maximum of 53.0. It is interesting to compare the range of the human progression indexes with the indexes of other primates, including those based on estimated brain-to-body ratios for the two species of *Australopithecus*. For example, the maximum progression index for chimpanzees is 12.0. If one works the formula backward and assumes a chimpanzeelike progression index for the gracile species of *Australopithecus*, using an average cranial capacity of 442 c.c., the required body weight turns out to be about 100 pounds. That is nearly twice the maximum estimated weight for the species.

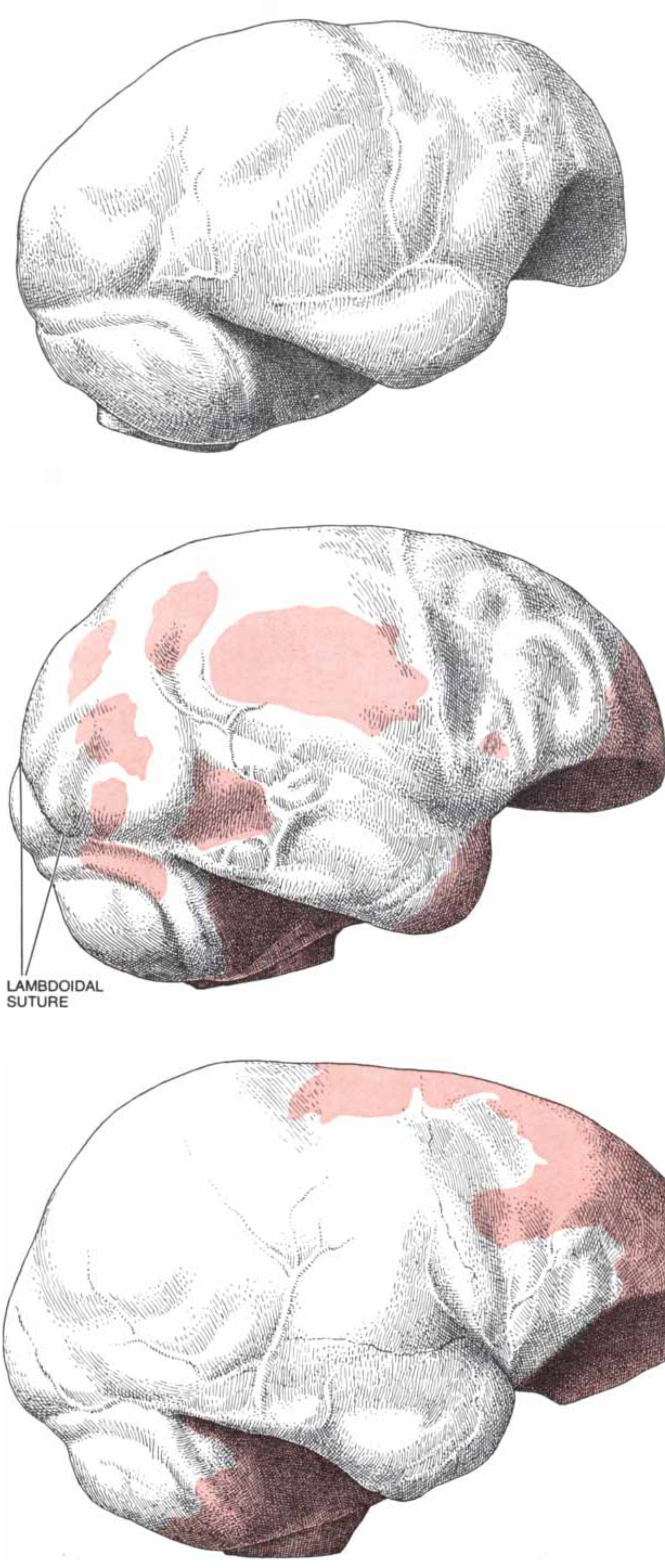
If one uses the same average cranial capacity and assumes that the body weight of the gracile species was only 40 pounds, the progression index is 21.4, which is well within the human range and comfortably close to the human average. When the body weight of the gracile species is estimated at 50 and 60 pounds and the body weight of the robust species is estimated at 60 and 75 pounds, the resulting progression indexes also fall close to or within the human range.

To recapitulate, both the direct evidence of neurological organization and the indirect evidence of comparative brain size appear to indicate that *Australopithecus* and at least one other African primate of the period from three million to one million years ago had brains that were essentially human in organization and that *Australopithecus* was also probably within the human range of sizes with respect to the proportion of the brain to the body. That the brain was small in absolute size, particularly in the gracile species of *Australopithecus*, therefore seems to be without significance. The ratio of brain size to body was appropriate. So far as the subsequent absolute cranial enlargement is concerned, the major mechanism involved, although surely not the only one, appears to have been that as hominids grew larger in body their brains enlarged proportionately. Quite possibly this expansion progressed at different rates at different times [*see illustrations on page 111*]. In any event, for at least the past three million years there has been no kind of cranial Rubicon waiting to be crossed.

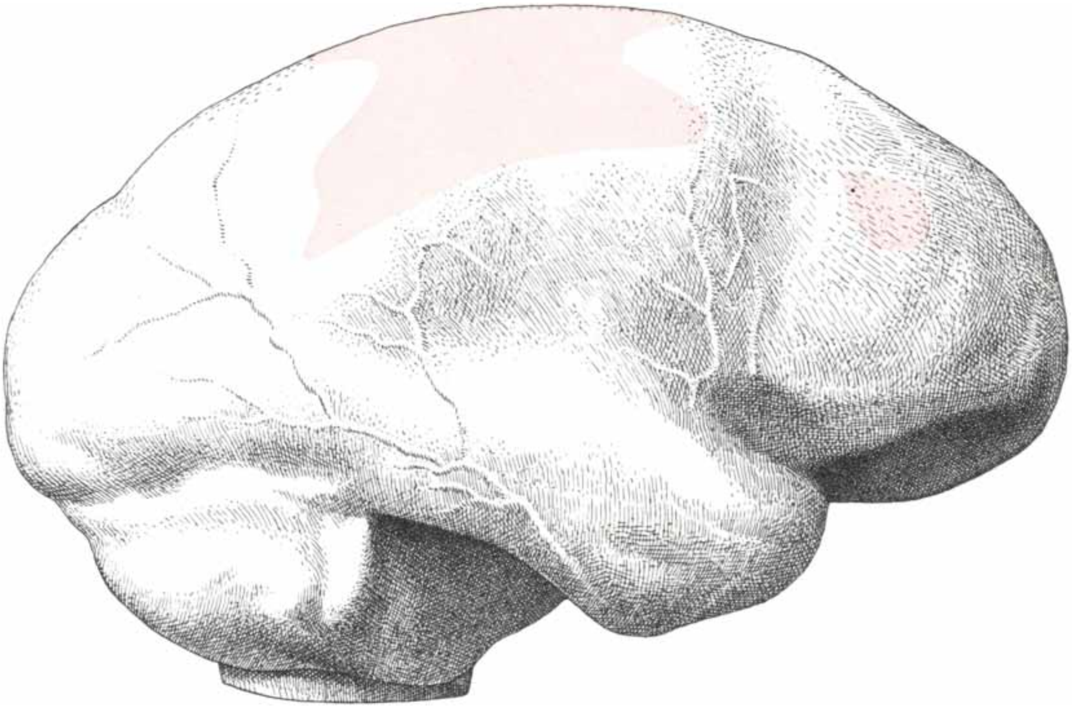
Considering the gaps in the primate fossil record, it is remarkable that so many specimens of a hominid intermediate between the earliest African fossil forms and modern man have been unearthed. The intermediate form is *Homo erectus.* The skulls and fragmentary postcranial remains of that species have been found at various sites in China and in Java, and more fossils are steadily being turned up in both areas. There is also one specimen (and possibly a second) from Olduvai and, depending on one's choice of authorities, one from southern Africa, one from eastern Europe and perhaps two from northern Africa. With the possible exception of the Olduvai and Java fossils, no *H. erectus* specimen is much more than 500,000 years old, and one or more may be a great deal younger.

I have made endocasts of five of the six available *H. erectus* specimens from Java and of both the certain Olduvai specimen and the doubtful one. The *erectus* fossil with the largest cranial capacity is Olduvai hominid No. 9; its brain measures 1,067 c.c. Omitting the doubtful Olduvai specimen, the *erectus* fossil with the smallest brain is one of those from Java. Its cranial capacity is 815 c.c., only 40 c.c. larger than the capacity of the three-million-year-old East Rudolf skull. The average *erectus* cranial capacity, based on the endocasts I have made, is 930 c.c.

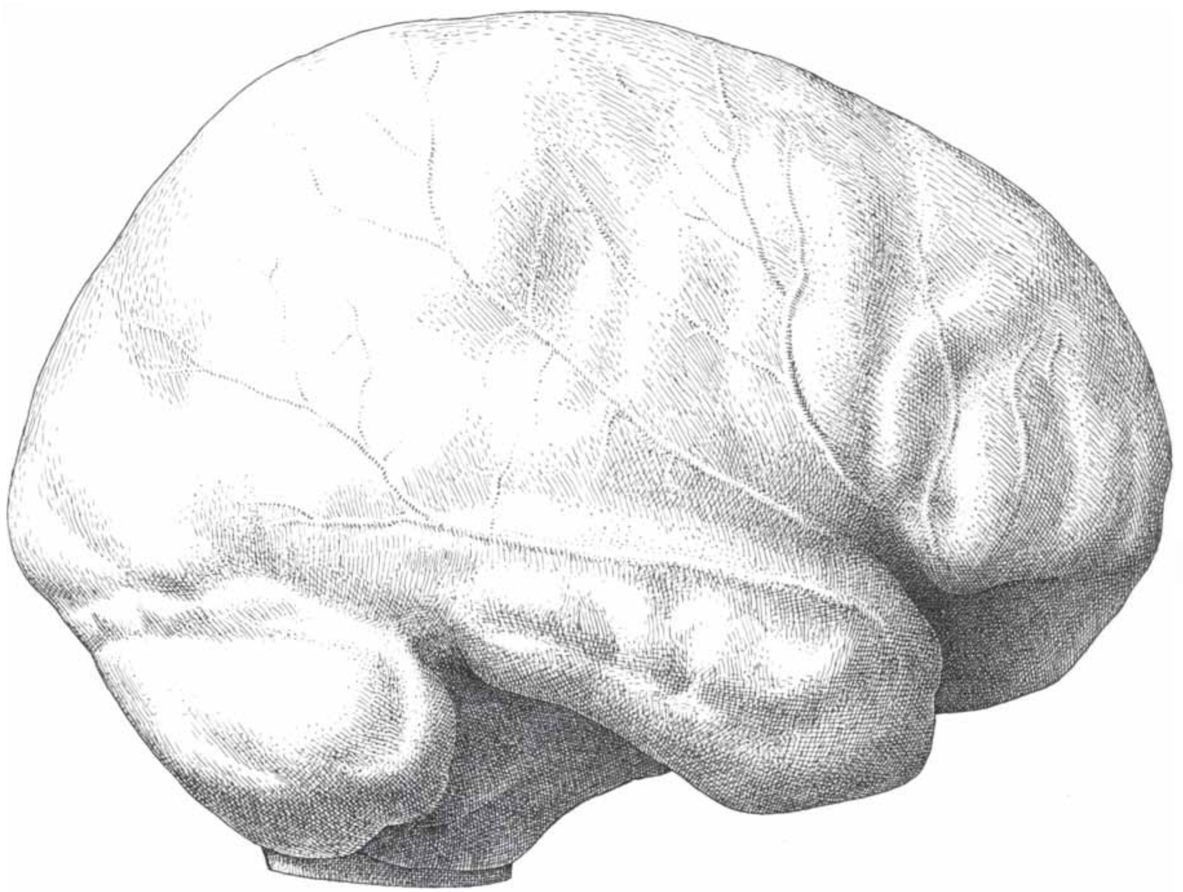
Working with this average brain size and estimating the body weight of *H. erectus* to have been 92 pounds, one finds that the Stephan progression index for the species is 26.6. That is remark-



ENDOCRANIAL CASTS are those of (*top*) a chimpanzee, *Pan troglodytes,* (*middle*) a gracile *Australopithecus africanus,* and (*bottom*) the other species, *A. robustus* (*color indicates restored areas*). In all three casts details of the gyral and sulcal markings of the cerebral cortex are minimal. A differing neurological organization, however, can be seen. Both of the hominid brains are higher, particularly in the parietal region. Orbital surface of their frontal lobes is displaced downward in contrast to chimpanzee's forward-thrusting olfactory rostrum. The location of the hominids' lunate sulcus, indicated (*middle*) by suture markings, implies a far smaller occipital lobe than the ape's.



BRAIN CAST OF HOMO ERECTUS shows similar evidence of human neurological organization. As is true of most human cranial casts, the position of the lunate sulcus cannot be determined but the expansion of the temporal lobe and the human shape of the frontal lobe are evident. This is a cast of Java specimen VIII (1969); it reflects the flat-topped skull conformation typical of the fossil forms of *H. erectus* found in Indonesia. Like endocranial casts on preceding page and below it is shown 90 percent actual size.



BRAIN CAST OF HOMO SAPIENS was made from a cranium in the collection at Columbia University. The height of the cerebral cortex, measured from its summit to the tip of the temporal lobe, and the fully rounded, expanded frontal lobe, showing a strong development of Broca's area (*see illustration on page 109*), typify the characteristic *H. sapiens* pattern of neurological organization.

ably close to modern man's average of 28.8. Even if the body-weight estimate is raised by some 30 pounds, the progression index falls only to 22.0. This being the case, it is difficult to escape the conclusion that, like *Australopithecus*, *H. erectus* possessed a brain that had become enlarged in proportion to the enlargement of the body. Although the average *erectus* brain is smaller than the average brain of modern man, it nonetheless conforms in gross morphology to the species' status as a recognized member of the genus *Homo* [*see top illustration on opposite page*].

I am not suggesting that any simple, straight-line progression connects the earliest African hominids, by way of *Homo erectus*, to modern man. As others have noted, it is the investigator's mind, and not the evidence, that tends to follow straight lines. The data are still far too scanty to trace the detailed progress of the human brain during the course of hominid evolution. For example, it would not even be safe to assume that, after attaining the stage represented by the earliest-known hominid endocasts, the brain thereafter consistently increased in size. Perhaps there were times when on the average brain sizes decreased simply because body sizes also decreased. One might even go so far as to speculate, and it would be pure speculation, that we are seeing something like this when we observe that the specimens of *Homo habilis* at Olduvai had substantially smaller brains than the far older East Rudolf hominid did.

Some generalizations are nonetheless possible. First, both the direct evidence of the endocasts and the indirect evidence of comparative cranial capacities indicate that the human brain appeared very much earlier than the time when *H. erectus* emerged, perhaps 500,000 years ago. Second, it can be inferred that the emergence of the human brain was paralleled by the initiation of human social behavior. It is not appropriate here to review the evidence of relations between nutrition and behavioral development or of those between the endocrine system and brain growth. Still, it is obvious that brains do not operate in a vacuum and that a part of the nourishment the brain requires is social as well as dietary. Much of the humanness of man's brain is the result of social evolution. The weight of the inferential evidence today suggests that the genesis has been long in the making. It may well predate such elements of fossil behavior as the systematic use of stone tools and the large-scale practice of hunting.

MATHEMATICAL GAMES

On the patterns and the unusual properties of figurate numbers

by Martin Gardner

Ancient Greek mathematicians, particularly the Pythagoreans, were entranced by figurate numbers: numbers that could be represented by arranging points in regular patterns on a plane or in space. Among the plane figurate numbers the most studied were the polygonal numbers. The illustration below shows how the first four polygonal numbers—triangular, square, pentagonal and hexagonal—are built up as the partial sums of simple arithmetic progressions. Triangular numbers are partial sums of the counting numbers $1 + 2 + 3 + 4 + \ldots$. Square numbers are formed by successive addition of the consecutive odd numbers $1 + 3 + 5 + 7 + \ldots$. Pentagonal numbers derive from the progression $1 + 4 + 7 + 10 + \ldots$ and hexagonal numbers from the progression $1 + 5 + 9 + 13 + \ldots$. The respective differences are $1, 2, 3, \ldots$.

The study of figurate numbers belongs to a branch of number theory called Diophantine analysis, which has to do with finding integral solutions of equations. An enormous effort was made by the great pioneers of number theory in studying the properties of polygonal numbers. Most of this work is crisply summarized in the second volume of Leonard E. Dickson's *History of the Theory of Numbers.*

Let us start with a classic problem that was solved by Leonhard Euler in 1730. How can we find all the numbers that are both square and triangular? The formula for the nth triangular number is $\frac{1}{2}(n^2 + n)$. If this expression is also square, we have the Diophantine equation $\frac{1}{2}(n^2 + n) = m^2$. Learning the technique of solving this equation is an excellent introduction to Diophantine analysis. The initial step is to manipulate the equation to get a simpler equation that will be a key to the solution. One way to do it is:

1. Express the equation as $n^2 + n = 2m^2$.
2. Multiply each side by 4: $4n^2 + 4n = 8m^2$.
3. Add 1 to each side: $4n^2 + 4n + 1 = 8m^2 + 1$.
4. Factor $(2n + 1)(2n + 1) = 2(4m^2) + 1$.
5. Let $y = (2n + 1)$ and $x = 2m$.
6. Substituting these terms in the preceding equation produces $y^2 = 2x^2 + 1$.

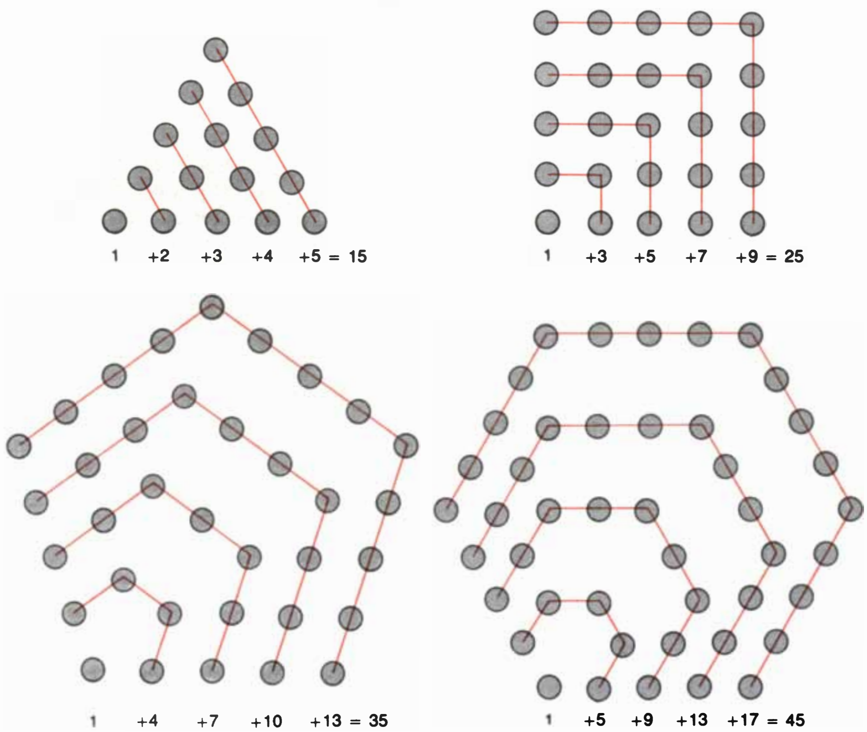
This is the simplest form of what is called the Pell equation, about which more below. If we can find an integral solution for it, we can easily work backward to find integral values for n and m in the original equation. A standard algorithm for cracking a Pell equation is to express the square root of the coefficient of x (in this case 2) as a continued fraction, then explore its convergents for values of x and y that satisfy the Pell. The technique is too involved to explain here, but interested readers will find a good introduction to it in Chapter 22 of Albert H. Beiler's *Recreations in the Theory of Numbers* (a Dover paperback) or in any good textbook on Diophantine analysis.

It turns out that whenever the coefficient is not a square the Pell has an infinity of solutions. Since 2 is not a square, there is an infinity of square triangles. The sequence begins 1, 36, 1225, 41616, 1413721, ... The recursive procedure for continuing this sequence is to multiply the last square triangle by 34, subtract the preceding square triangle, then add 2. A nonrecursive formula for the nth square triangle is

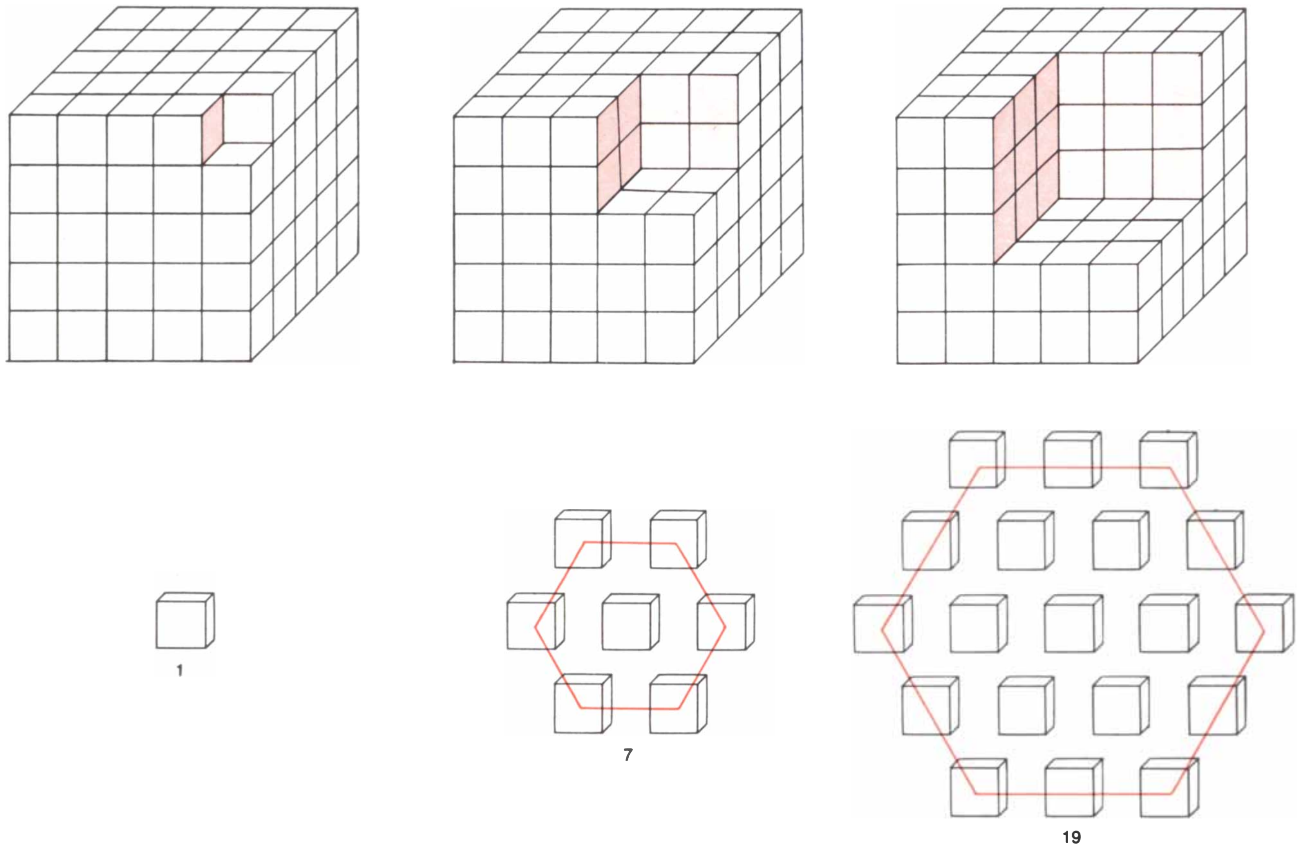
$$\frac{[(17 + 12\sqrt{2})^n + (17 - 12\sqrt{2})^n - 2]}{32}.$$

The irrational numbers in the formula might lead one to suppose that rounding up or down is necessary, but this is not the case. The formula is exact. Substitute any positive integer for n, and the irrationals drop out to give an integral value for the expression. It is surprising how often the problem of finding this formula turns up in the problem departments of mathematical journals, even though it has been shown that the formula goes back to Euler.

Square triangles have many unusual properties. One of the most surprising is that when a simple algorithm is applied, each square triangle gives the sides of an integral right triangle with one leg ex-



Construction of order-5 polygonal numbers



Hexagonal numbers as the difference between consecutive cubes

actly one unit longer than the other. Let the sides of the right triangle be x and $x + 1$ and z the hypotenuse. Let v be the square root of a square triangle and u its side when represented as a triangle. The procedure is merely to solve these two simultaneous equations:

$$u = z - x - 1$$
$$v = \tfrac{1}{2}(2x + 1 - z)$$

For example, if we take the second square triangle, 36, then $v = 6$, $u = 8$. The above equations give x a value of 20 and z a value of 29. The Pythagorean triplet therefore is 20, 21, 29. Had we used the first square triangle, 1, the algorithm would have provided the familiar 3, 4, 5 right triangle. The third square triangle gives the triplet 119, 120, 169. In this way all Pythagorean triangles with consecutive legs can be obtained from the square triangles, and of course we can go the other way and derive all the square triangles from consecutive-legged Pythagorean triangles. The simple procedure, or one equivalent to it, explains how Beiler was able to construct his table (pages 328 and 329 in his book) of the first 100 Pythagorean triangles with consecutive legs. The 100th such triangle has legs that are each expressed by a 77-digit number. No Pythagorean triangle can have equal legs, but this monstrosity is so nearly isosceles that, as Beiler graphically points out, if its smaller leg is a light-year long, the other leg would be longer by an amount so infinitesimal that the difference between the two legs would be millions of times less than the diameter of a proton.

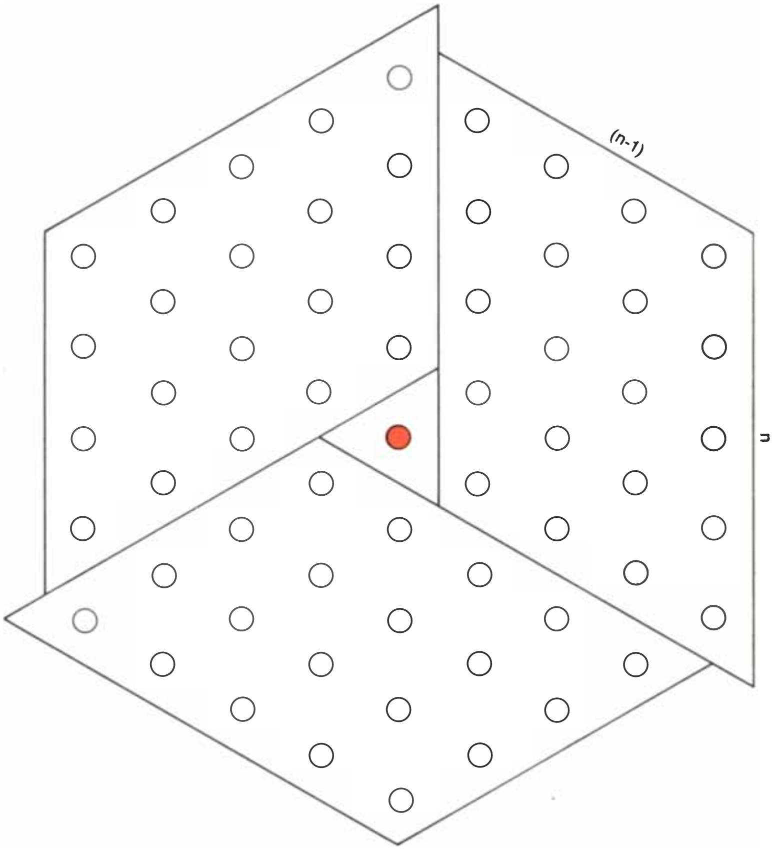
We turn now to two planar figurate numbers that are not polygonal in the classic sense. The first has received scant attention in the past. The second, so far as I know, has not previously been recognized as a figurate number.

If we arrange points as shown in the top illustration on the next page, we have what are known as centered hexagonal numbers. Let us call them "hexes" for short. As the illustration makes clear at a glance, the formula for the nth hex is $3n(n - 1) + 1$. It is the sum of three rhombuses, each of sides n and $(n - 1)$, plus the single spot in the center. The bottom illustration on the next page shows that a hex is also the sum of six triangles plus the central spot. The sequence begins 1, 7, 19, 37, 61, 91, 127, 169, ... The recursive procedure is to multiply by 2, subtract the preceding number and add 6.

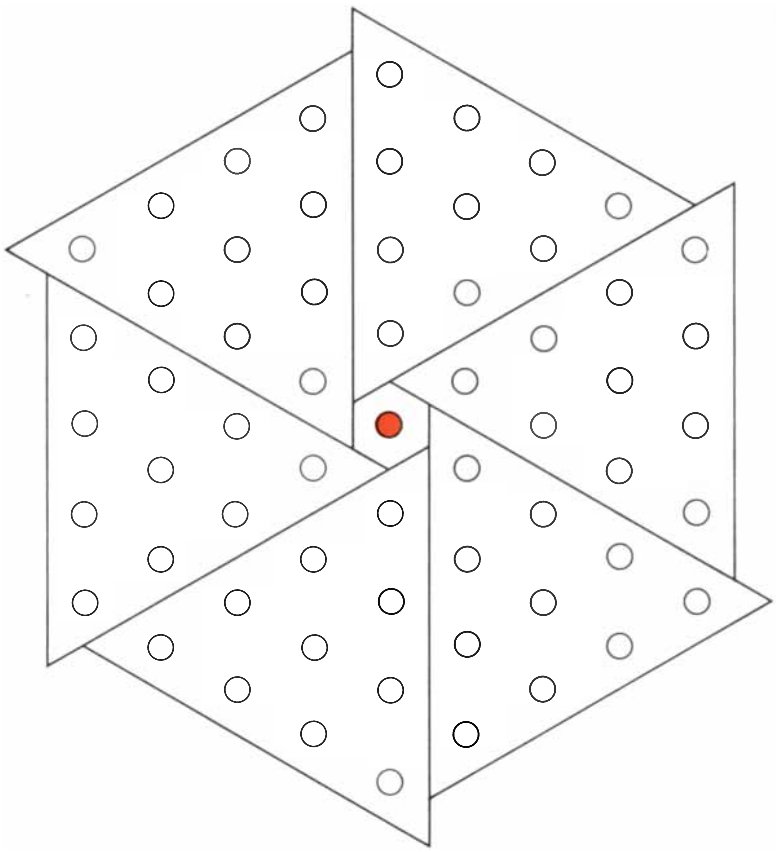
Suppose we build a hex pyramid of coins, starting with a hex that has 100 coins on the side. On top of this we put a hex of 99 coins on the side, then one of 98, and so on, until finally we cap the pyramid with a single coin on the central stack. The pyramid is 100 layers high. How many coins are in it? To answer this we need to know the formula for the sum of the first n hexes. The answer is unexpectedly simple. It is n^3. There are therefore $100^3 = 1{,}000{,}000$ coins in the hex pyramid.

It follows from this formula that every hex is the difference between two consecutive cubes. We can demonstrate this elegantly by building a cube, say a $5 \times 5 \times 5$, out of unit cubes. Remove one cube (the first hex) from a top corner. It leaves a $1 \times 1 \times 1$ hole. Around that hole are 7 cubes (the second hex). Removing these 7 leaves a $2 \times 2 \times 2$ cubical hole. Surrounding this hole are 19 cubes (the third hex). Removing the 19 cubes leaves a $3 \times 3 \times 3$ cubical hole. And so on [*see illustration above*].

Apart from a hex of 1, the first triangular hex is 91 and the first square hex is 169. Readers who know the Pellian technique may enjoy searching for recursive procedures that will generate each of these infinite sequences of num-



Formula for **n***th hex number is* $[3 \times n(n-1)]+1$



Six triangles plus the center make a hex

bers and for their nonrecursive formulas. The Pellian for square hexes is $3x^2+1=y^2$, which is solved by finding the convergents of the continued fraction for the square root of 3. The next square hex after 169 is 32761, and the next is 6355441. Are there hexes that are both square and triangular? Is there a cubical hex?

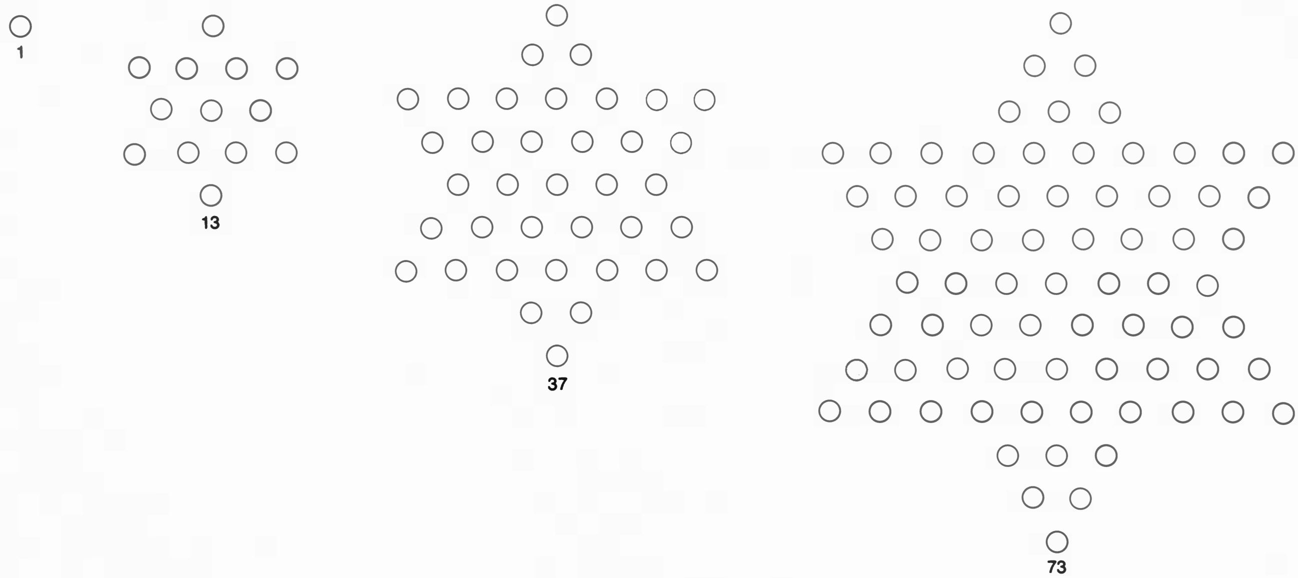
Closely related to hexes are numbers that I have not seen recognized before as figurate although one often sees them as patterns for drainage holes, and game enthusiasts know them as the patterns of boards for playing Chinese checkers. Let us call them "star" numbers. The top illustration on the opposite page shows the first four stars and the bottom illustration is a "look-see" proof of the formula for the nth star. Clearly a star consists of six rhombuses, each n by $(n-1)$, plus the central spot, or $6n(n-1)+1$. A star is also the sum of 12 triangles plus the central spot, as is shown in the top illustration on page 120.

The star sequence begins 1, 13, 37, 73, 121, 181, 253, 337, 433, 541, ... Adding $12n$ to the nth star produces the next star. A hex contains six triangles. Adding six more triangles to its six sides produces a star; consequently any hex number becomes a star number if we double it and subtract 1. The first n stars add up to $2n^3-n$. Is this sum ever a square? Yes, but only when $n=1$ or 169. This was not established until 1942. According to Louis J. Mordell, who mentions it on page 261 of his *Diophantine Analysis*, "the proof is exceedingly complicated."

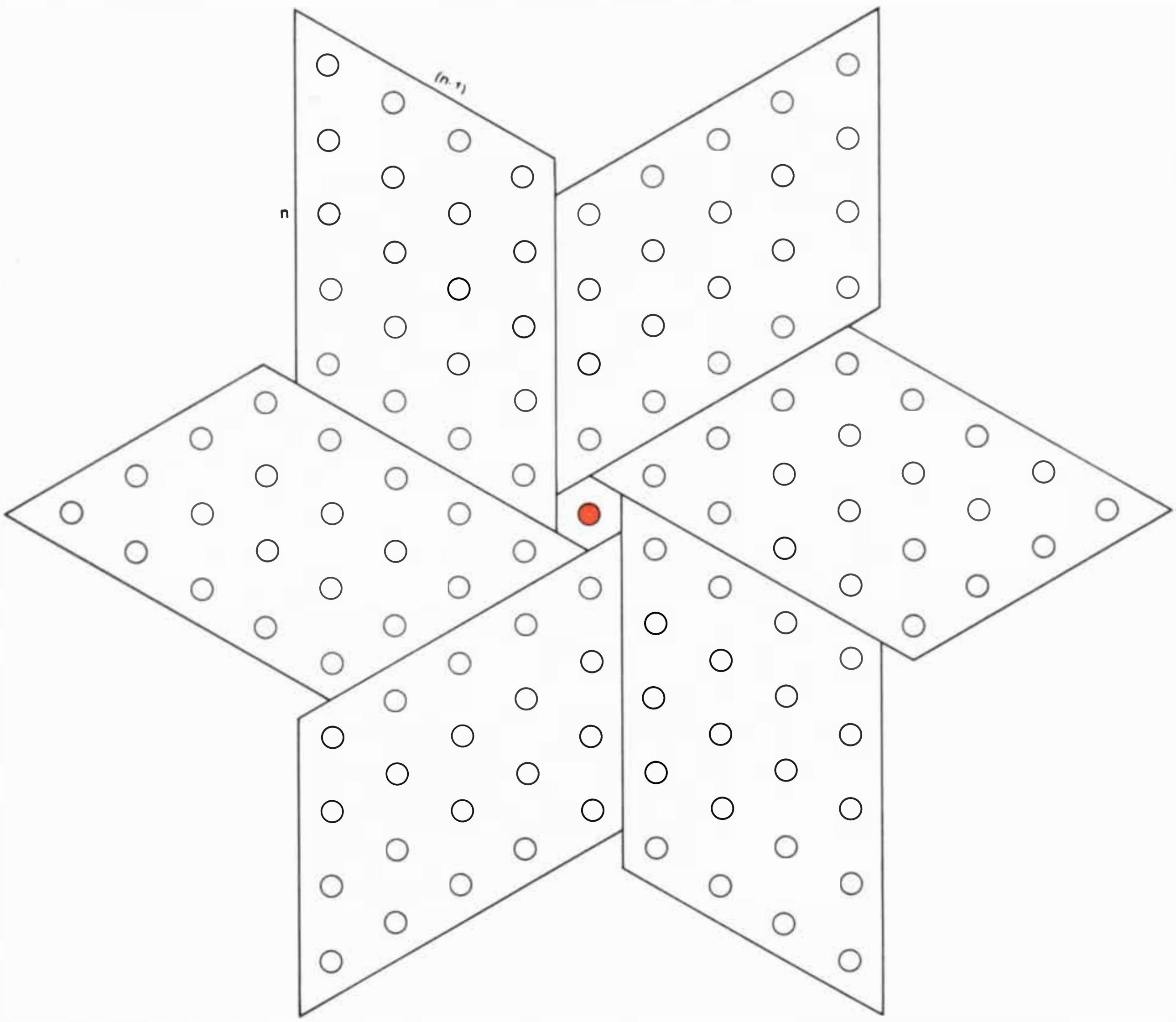
The first triangular star after 1 is 253. The recursive procedure is to multiply a triangular star by 194, add 60 and subtract the preceding triangular star. The infinite sequence begins 1, 253, 49141, 9533161, 1849384153, ... The nonrecursive formula for the nth triangular star is

$$\frac{3[(7+4\sqrt{3})^{2n-1}+(7-4\sqrt{3})^{2n-1}]-10}{32}.$$

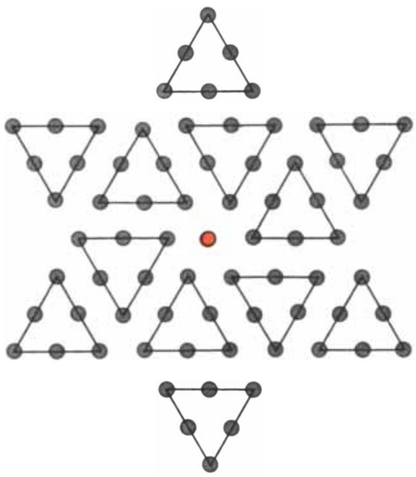
The first square star after 1 is 121. It is the number of holes on the standard Chinese checkers board. The recursive procedure is to multiply a square star by 98, subtract the preceding square star and add 24. The infinite sequence begins 1, 121, 11881, 1164241, 114083761, ... For readers who care to work through the solution of the equation for square stars, $6n(n-1)+1=m^2$, I shall say only that this reduces to a solution of $2x^2+1=3y^2$, where x is the square root of a square star. It can be solved by finding alternate convergents of the square



The first four stars



Formula for **nth** *star is* $[6 \times n(n-1)] + 1$

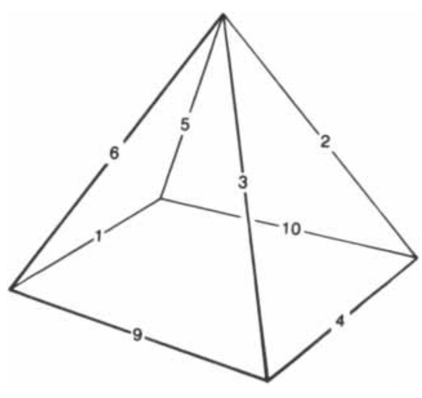


Twelve triangles plus the center make a star

root of 3/2. The exact nonrecursive formula for the *n*th square star is

$$\left[\frac{(5+2\sqrt{6})^n(\sqrt{6}-2)-(5-2\sqrt{6})^n(\sqrt{6}+2)}{4}\right]^2$$

The digital roots of square stars (their value modulo 9) must be 1 or 4, and it is not hard to prove that square stars must begin and end with 1. One of their most remarkable properties is that they provide a simple algorithm for producing every number that can be expressed as the sum of two consecutive squares and also as the sum of three consecutive squares. The smallest such number is 365 (the number of days in the year), which equals $13^2 + 14^2$ and also equals $10^2 + 11^2 + 12^2$. The procedure is simply to take any square star greater than 1, triple it and add 2. The smallest square star greater than 1 is 121. Three times 121, plus 2, is 365. The next square star, 11881, leads to the number 35645, which equals $133^2 + 134^2$ and also equals $108^2 + 109^2 + 110^2$. The third case is $3(1164241) + 2 = 3492725 = 1321^2 + 1322^2 = 1078^2 + 1079^2 + 1080^2$. In each case the middle term of the triplet of consecutive squares is the original square star.



Dr. Matrix' magic pyramid

It is a pleasant exercise, demanding no special skills in number theory, to show that the algorithm always works. Can the reader find a simple proof before I give one next month?

I have been told by Victor Meally that Matila Ghylea, a Romanian mathematician, published some studies of hexes, but I do not know the references. I was unable to find any discussion of stars as such, although their formula turns up in connection with many Diophantine problems. One can raise all kinds of questions about stars that may be easy or difficult to answer. I do not know, for example, if there are stars that are both square and triangular. Since a star's digital root is 1 or 4, and a triangle's digital root must be 1, 3, 6 or 9, we can say that a square triangular star must have a digital root of 1, but that is not of much help.

The general Pell equation, a key to so much of this kind of number analysis, is $ax^2 + 1 = y^2$, where a is a positive integer. It has an infinity of positive integral solutions for x and y unless a is a square, in which case there are no solutions. As we have seen, when $a = 2$ we have the key to finding square triangles, and when $a = 3$, the key to finding square hexes. The equation was mistakenly named for John Pell, a 17th-century number theorist, because of a false impression on Euler's part. Pell had nothing to do with the equation. It was known to early Greeks and Hindus, but Pierre Fermat was the first to propose advanced work on it and the general solution was obtained by John Wallis and others. The classic reference is *The Pell Equation* (Columbia University Press, 1912), by E. E. Whitford, a book unfortunately out of print. The tables in this book can save vast amounts of tedious calculations with continued fractions. As J. A. Lindon has put it elegantly in one of his unpublished mathematical Clerihews:

To equations simultaneously Pellian
My approach is Machiavellian.
Anything goes, rather than resort to
such actions
As covering the walls with continued
fractions.

I wish to thank Meally for providing the nonrecursive formulas for square stars and triangular stars, as well as other things in this column, and also to thank John Harris and John McKay for additional assistance. I also found N. J. A. Sloane's *A Handbook of Integer Sequences* (Academic Press, 1973) to be an invaluable tool. I shall say no more about this marvelous reference except that every recreational mathematician should buy a copy forthwith. (See Philip Morrison's enthusiastic review in the April issue of *Scientific American*.)

Last month's problem was to label the eight edges of a square-based pyramid with different numbers from 1 through 10 so that the sum of the edges meeting at each of the five vertexes is 16. Five vertexes, each summing to 16, make a total of $5 \times 16 = 80$. Because each edge number contributes to two vertexes, the sum of the eight numbers must be $80/2 = 40$. Only three sets of eight different positive integers, each no greater than 10, add to 40. They are:

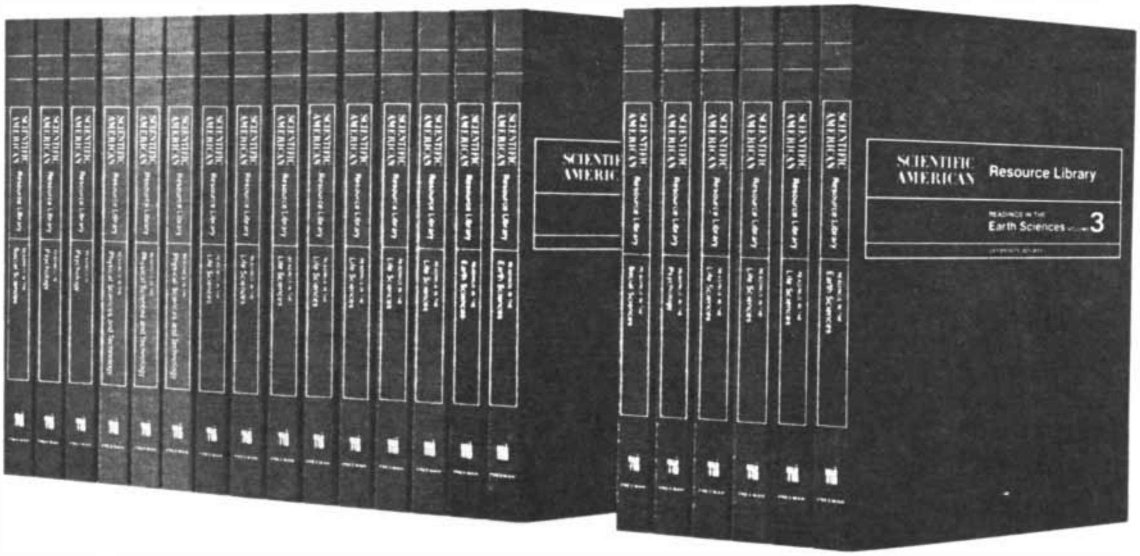
(1) 1, 2, 3, 4, 5, 7, 8, 10.
(2) 1, 2, 3, 4, 6, 7, 8, 9.
(3) 1, 2, 3, 4, 5, 6, 9, 10.

Consider the first set. The number 10 must label either a base edge or an edge on the pyramid's side. If it is a base edge, then at each end it must meet a pair of edges with a sum of 6. The only possible doublets are 1, 5 and 4, 2. By trying them in their four possible arrangements, it is easy to see that there is no way to complete the labeling to make the pyramid magic. If 10 labels a side edge, the three edges meeting at the apex must add to 6. The only possible triplet is 1, 2, 3. Considering the symmetry, there are just three arrangements (the number opposite 10 can be 1, 2 or 3). None allows completion of the labeling.

Consider the second set. If 9 is a base edge, the pairs meeting at the two ends must be 6, 1 and 4, 3. Trying them in their four arrangements does not lead to a solution. If 9 is a side edge, the other three edges meeting at the apex must be 1, 2, 4, in three arrangements. Again no solution is possible.

Consider the third set. If 10 is a side edge, the triplet meeting 10 at the top must be 1, 2, 3. It allows no solution. If 10 is a base edge, the doublets at the ends must be 1, 5 and 4, 2. This allows only the solution shown in the illustration at the left. It is unique except for rotations and reflections.

Is it possible to construct a different magic pyramid, drawing from the same set of numbers but with a different constant? Yes, one other construction is possible. The constant is 18 and the numbers are 2, 3, 4, 5, 6, 7, 8, 10. Readers who are interested should have little difficulty finding the unique labeling.



Announcing the first expansion of the SCIENTIFIC AMERICAN Resource Library

Six new volumes have just been added to what *The New York Times Book Review* has called "The Intelligent Layman's Guide to the Golden Age of Science."* They are:

READINGS IN THE EARTH SCIENCES, Volume 3
Offprints 875-899, $10.00

READINGS IN THE LIFE SCIENCES, Volumes 8, 9, and 10
Offprints 1121-1247, $30.00

READINGS IN PSYCHOLOGY, Volume 3
Offprints 516-542, $10.00

READINGS IN THE SOCIAL SCIENCES, Volume 2
Offprints 638-675, $10.00

The preexisting volumes are:

READINGS IN THE EARTH SCIENCES
With an Introduction by
William H. Matthews III, Lamar State College of Technology
Two volumes, Offprints 801-874, $20.00

READINGS IN THE LIFE SCIENCES
With an Introduction by Marston Bates, the University of Michigan
Seven volumes, Offprints 1-199; 1001-1120, $70.00

READINGS IN THE PHYSICAL SCIENCES AND TECHNOLOGY
With an Introduction by
Isaac Asimov, Boston University School of Medicine
Three volumes, Offprints 201-326, $30.00

READINGS IN PSYCHOLOGY
With an Introduction by Harry F. Harlow, the University of Wisconsin
Two volumes, Offprints 401-515, $20.00

READINGS IN THE SOCIAL SCIENCES
With an Introduction by Harry F. Harlow, the University of Wisconsin
One volume, Offprints 601-637, $10.00

COMPLETE LIBRARY
Twenty-one volumes in all, $210.00 the set

*"Their basic contents are hundreds of articles which explain simply, vividly, and interestingly what modern science is all about . . . most of the vast riches that have appeared monthly since the late 1940's in the *Scientific American,* this country's and perhaps the world's outstanding forum for communication between scientists and the intelligent public." —*The New York Times Book Review,* April 26, 1970

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, CA 94104
58 Kings Road, Reading, England RG 1 3AA

THE AMATEUR SCIENTIST

The polariscope as a measuring instrument and as a means of creating objects of art

Conducted by C. L. Stong

A polariscope is an instrument that polarizes light, that is, it alters the orientation of the planes in which light waves vibrate. One effect of optical polarization is the disappearance of reflected glare when one dons Polaroid sunglasses. With a simple polariscope one can also detect stresses in transparent substances and the concentration of chemicals in solution and can create objects of art that have the quality of stained-glass windows.

A polariscope for doing such experiments can be made at home for about the cost of an inexpensive camera. According to Leslie Holliday (The Knoll, Cley, Norfolk, England), the instrument can display phenomena of sufficient variety to fascinate every member of the family, including people more interested in art than in optical effects. Holliday describes the construction of the polariscope and some of his experiments.

"A polariscope consists of a light source, a diffuser plate, a polarizer, an analyzer and a framework to support these parts in alignment [*see top illustration on page 124*]. The polarizer and the analyzer are optically identical. They are made with sheets of Polaroid or an equivalent material. The light source is an incandescent bulb of from 25 to 40 watts. It is installed in an enclosure designed to provide natural convective cooling. Rays from the lamp are diffused by a sheet of opalescent glass or acrylic plastic.

"Both the polarizer and the analyzer are preferably formed with sheets of Polaroid Type HN-22, a high-extinction material that polarizes transmitted light strongly. In the U.S. the material is available from the Polaroid Corporation (549 Technology Square, Cambridge, Mass. 02139) and from the Marks Polarized Corporation (153–16 Tenth Avenue, Whitestone, N.Y. 11357). It comes in sheets 12 inches square and .03 inch thick. The sheets transmit about 22 percent of the incident light. The material costs about $20 per square foot.

"Material of another type transmits more than 30 percent of the incident light but polarizes less strongly. It is available from the Edmund Scientific Co. (300 Edscorp Building, Barrington, N.J. 08007). Sheets 20 inches square are currently priced at $15.50 each. For protection and mechanical support both the polarizer and the analyzer should be sandwiched between thicker sheets of clear plastic or glass.

"The dimensions of the instrument can be altered according to preference. Other features of the design can also be modified. For example, some experimenters may prefer to enclose the lamp in a lighttight box. If so, care must be taken to provide ventilation ports fitted with light shields that do not seriously restrict the flow of cooling air.

"Objects are placed for examination in polarized light in the space between the polarizer and the analyzer, which I shall call the working space. This space should be at least two inches thick and six inches or more square. In general the versatility of the instrument increases with the size of the working space. The main limitations on the size are the maximum area of commercially available Polaroid and the budget of the experimenter.

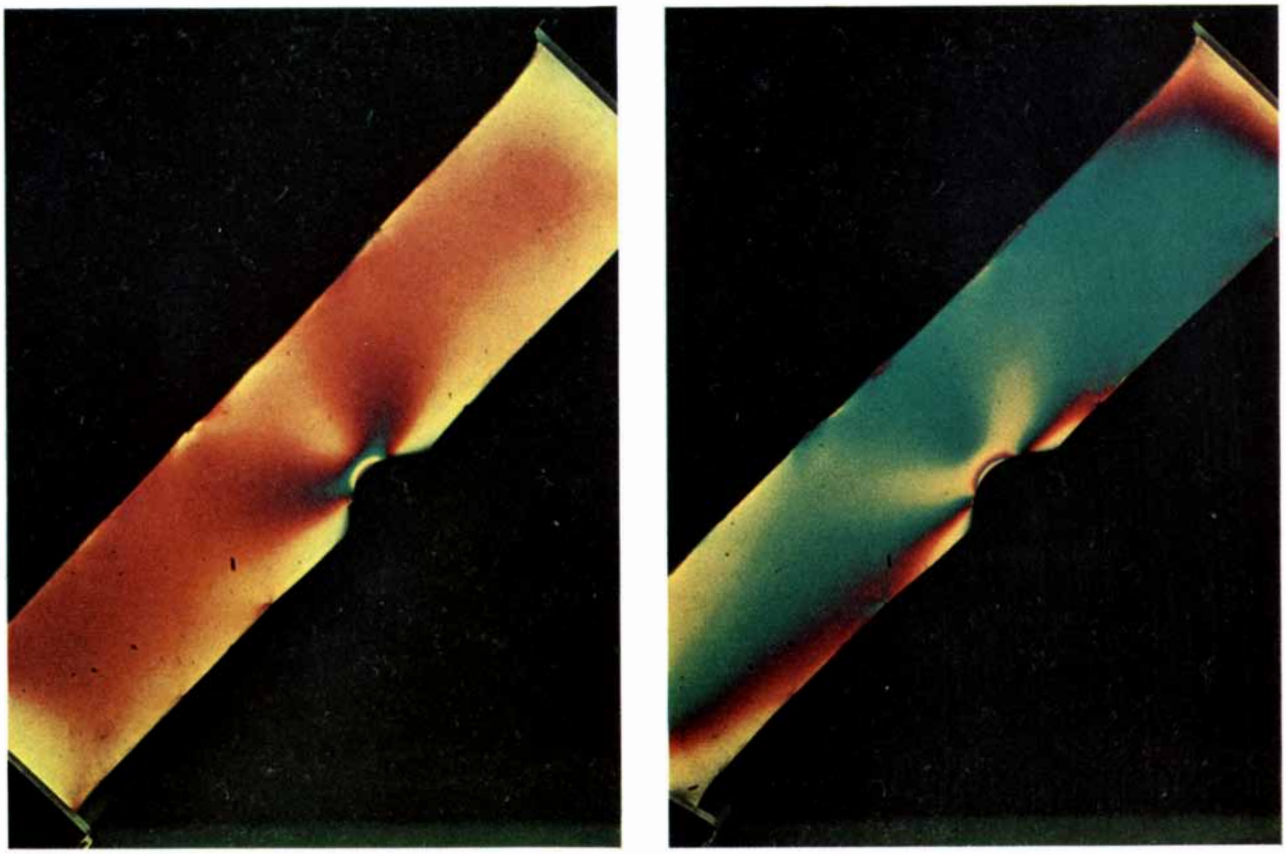
"Polariscopes of the most versatile type, which can make quantitative measurements, are built in a cylindrical form that enables the experimenter to rotate the analyzer in its plane relative to the fixed polarizer and to read the angle of rotation through at least a quarter of a turn by reference to a dial calibrated in degrees of arc. Square instruments should be made so that the analyzer can be removed easily from its supporting frame and rotated as desired through a quarter of a turn. The importance of adjusting the relative angle between the polarizer and the analyzer can be appreciated by considering the nature of the polarizing phenomenon.

"In theory one can imagine a ray of unpolarized light as consisting of myriad photons moving in a slender beam. If the finest details of the unpolarized beam could be observed end on, photons would be found vibrating parallel to every radial plane. Photons in a beam of polarized light vibrate in a single ribbonlike plane. The orientation of the vibrations can be vertical, horizontal or at any intermediate plane.

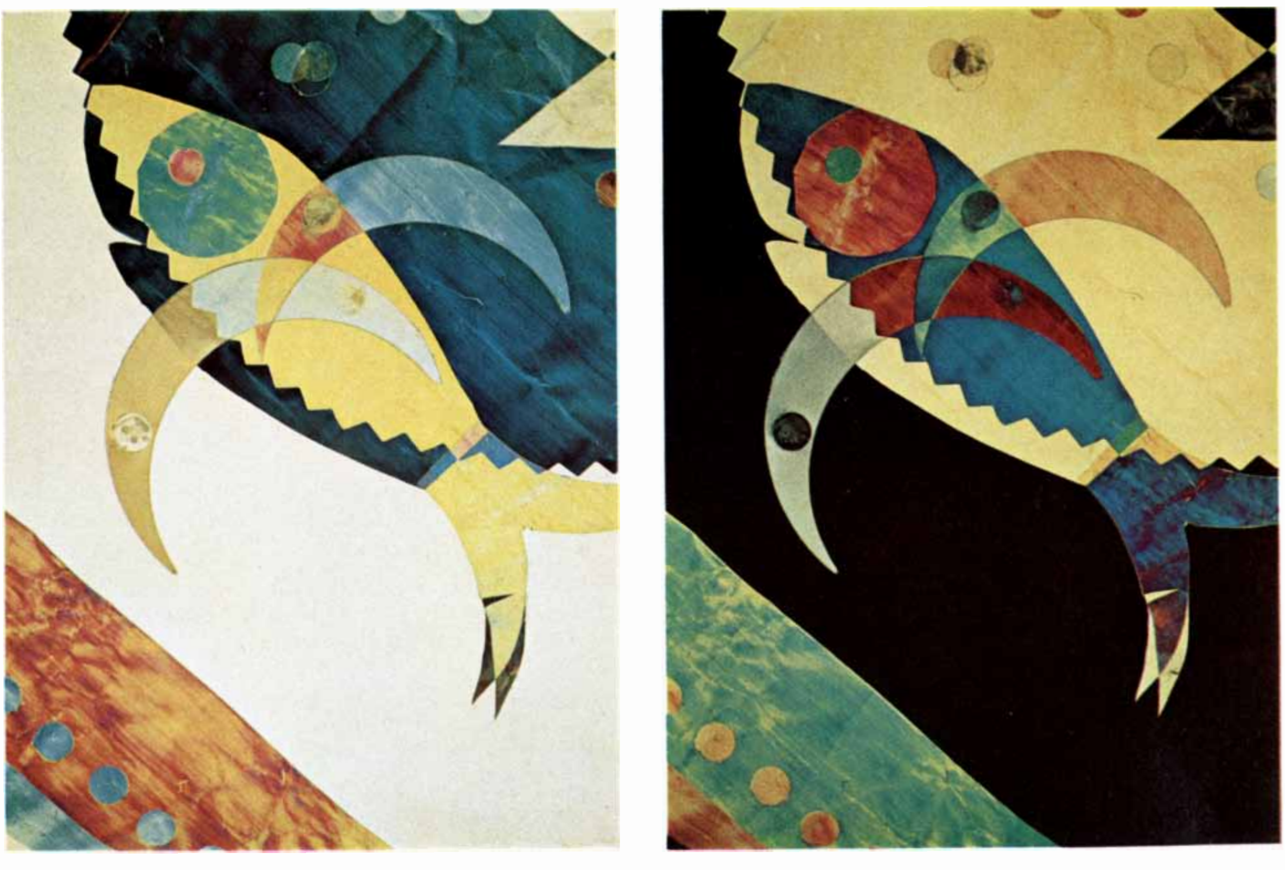
"Ordinary light, which vibrates in all planes simultaneously, can be polarized by several methods. The simplest method employs selective absorption. It is based on the discovery in 1852 by the British physician William B. Herapath that synthetic crystals made from a salt of quinine tend to absorb light increasingly as the plane in which the light waves vibrate departs from the optical axis of the crystals. Herapath attempted unsuccessfully to grow large crystals for polarizing wide beams of light. In 1932 Edwin H. Land packed Herapath's needlelike crystals into a soft sheet of nitrocellulose, stretched the plastic to align the needles in a gridlike array and let the plastic harden. He named the product Polaroid. It absorbs light waves vibrating in planes that make an angle with respect to the optical axes of the crystals and transmits waves that vibrate in the plane of the optical axes.

"Land later developed high-extinction Polaroid by similarly aligning organic molecules impregnated with iodine in a matrix of polyvinyl alcohol. Other varieties of polarizing materials have since been compounded. To visualize the mechanism of polarization it is convenient to imagine that grids of molecular particles filter out light waves that make angles with respect to the direction of the grids, much as only vertical waves can propagate along a rope stretched through a picket fence.

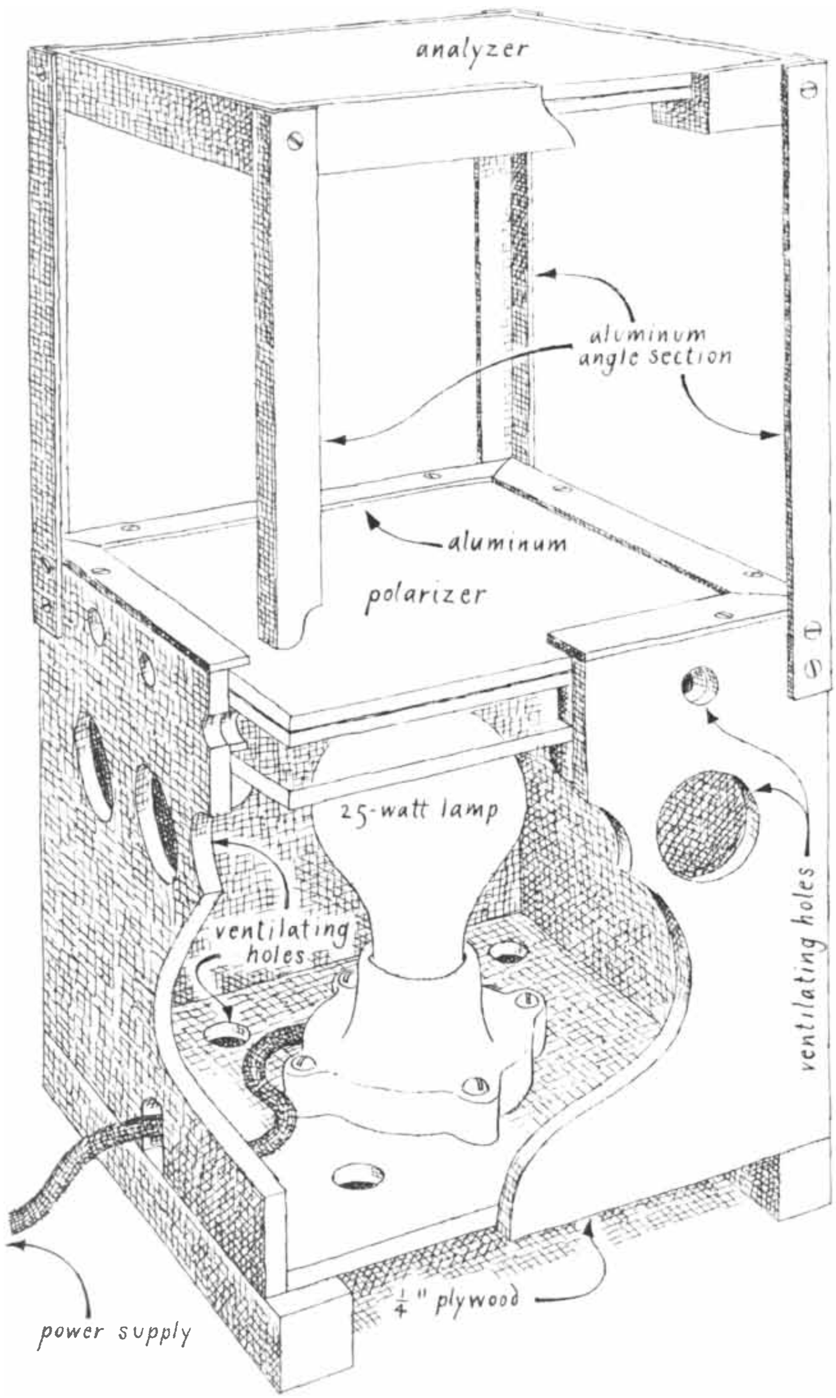
"The polariscope transmits maximum light when the optical axis (the imaginary grid) of the analyzer is rotated into parallel alignment with the polarizer. In general the intensity of the light transmitted by the combination varies in proportion to the trigonometric cosine of the angle between the optical axes of



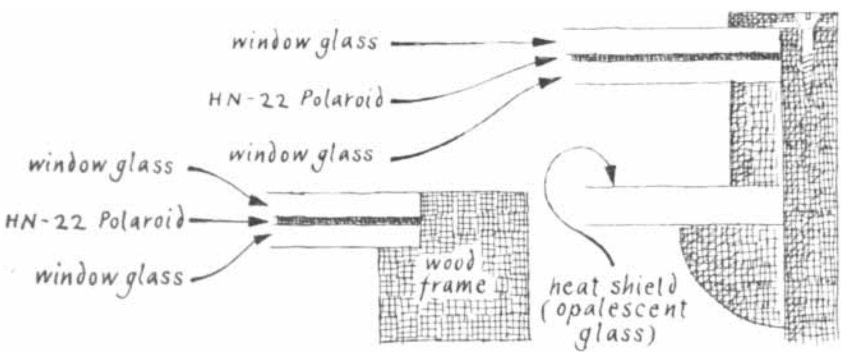
Strip of rubber seen in polariscope under weak stretching (left) *and strong stretching* (right)



Cellophane fish in complementary colors created by rotating polariscope 90 degrees



Leslie Holliday's polariscope



Details of the analyzer (left) *and the polarizer* (right)

the analyzer and the polarizer. The addition of a protractor enables the experimenter to read the angle directly and thus transforms the device into a quantitative instrument. A diameter of eight inches is recommended for circular polariscopes. A shaded lamp is helpful in instruments designed primarily to demonstrate colorful effects of polarization.

"In my opinion the most interesting effects are observed when the optical axis of the analyzer is adjusted to make a right angle with respect to the optical axis of the polarizer. With the analyzer and polarizer so crossed the instrument transmits minimum light (1 percent or less, depending on the kind of polarizing material). Many transparent substances polarize transmitted light more or less, depending on their nature, and polarize reflected light more or less, depending on the angle at which unpolarized rays fall on the reflecting surface.

"Light that is reflected upward from a wet surface at an angle of about 37 degrees becomes polarized. The reflected rays vibrate most strongly in the horizontal plane. For this reason Polaroid sunglasses are assembled with the optical axes of the crystals in the vertical plane. It is in this 'crossed' orientation that they tend to absorb the polarized glare.

"The effect can be demonstrated by examining with Polaroid sunglasses light reflected at various angles from ordinary window glass. The reflections from ordinary glass will be absorbed most strongly when the rays make an angle of about 33 degrees with respect to the surface of the glass. The angle of reflection at which the polarization is strongest is related to the refractive index of the transparent reflector (glass, water or some other substance).

"It is possible to make a rough estimate of the unknown refractive index of a substance with a pair of sunglasses and a protractor. Measure the angle at which reflected rays are absorbed most strongly. Subtract the angle from 90 degrees. The trigonometric tangent of the resulting angle is equal to the index of refraction.

"With ordinary glass the rays reflected at an angle of about 33 degrees are the ones most strongly absorbed by the Polaroid. The complement of this angle is 90 – 33, or 57, degrees. The tangent of 57 degrees is 1.539, which is the refractive index of ordinary glass. The refractive index of water ranges from 1.333 at a temperature of 14 degrees Celsius to 1.317 at the boiling point.

"Many substances have no effect on the polarization of light. Well-annealed glass is an example. A sheet of this glass vanishes (except possibly for the edges) when it is put between the polarizer and the analyzer. Many other substances put in the same place appear brightly lighted and many take on patterns of dazzling color.

"All materials of this type are said to be birefringent. They transmit light in two distinct planes at velocities that are unique and characteristic of each plane but less than the velocity of light in a vacuum. In other words, such materials have two indexes of refraction.

"The phenomenon can be observed by placing on the polarizer of the polariscope a glass microscope slide and a strip of cellophane. The microscope slide will not be seen through the analyzer, but the cellophane will appear brightly lighted and possibly colored. The glass slide will remain invisible at any orientation, but various regions of the cellophane brighten, darken and change color when the strip is rotated.

"Cellophane is one of the more active of the readily available birefringent materials. Its appearance in the polariscope ranges from white through the hues of the rainbow, including both saturated and pastel colors, depending on the thickness of the specimen (or the number of layers) and on its birefringent property. All colors are present as a mixture in the white light of the lamp.

"After polarization the cellophane splits the white mixture into two parts that vibrate in planes at right angles to each other and that travel at different velocities. Having passed through the cellophane, the rays combine to form a single white beam. Because of the relative difference in velocity, however, the waves in the two planes emerge more or less out of step. The resulting effect is equivalent to rotating the planes in which waves of each color vibrate.

"A striking feature of the phenomenon is that the imagined rotation is not equal for each color. The velocity of short waves is retarded more than the velocity of long waves. This phenomenon accounts for the observed colors. One area of a specimen of cellophane may retard by 5,000 angstroms waves that are perceived as green. This retardation would be equivalent to rotating the plane in which these waves vibrate out of alignment with the optical axis of the crystals of the analyzer. Accordingly the analyzer would not transmit wavelengths that are perceived as green. The transmitted portion of the remaining mixture would be perceived as the complementary color: red.

"Still other areas of the specimen might introduce retardations of less than 3,000 angstroms. Waves of this length lie in the ultraviolet region of the optical spectrum, beyond human perception. The eye is also insensitive to total retardations of less than 3,000 angstroms. Hence these areas of the specimen would appear white.

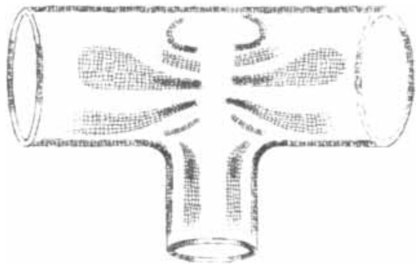
"Rotating the analyzer 90 degrees causes an observed color to be replaced by its complement. A sheet of cellophane can be rotated to positions of maximum and minimum transmission of light. In general colors are observed best when the cellophane occupies an intermediate position between these extremes, being at an angle of about 45 degrees with respect to the optical axes of both the polarizer and the analyzer.

"The characteristic amount by which light must be retarded to display each color has been measured [*see bottom illustration on next page*]. For example, if the crests and troughs of the waves emerge from the two paths of cellophane within 2,000 angstrom units of the same relative positions at which they entered the material, the specimen will appear white. In areas of the specimen where the waves of the two paths fall out of step by 7,000 angstroms the specimen appears green because the complementary color, red, has been absorbed.

"Retardation up to 7,000 angstroms results in bright, saturated colors. At still larger retardations colors continue to appear, but they are less brilliant. 'Higher order' colors are seen at higher retardations when two hues are extinguished simultaneously. The colors overlap. At a retardation of 14,000 angstroms the second-order red ($2 \times 7,000$ angstroms) and the third-order blue ($3 \times 4,500$ angstroms) are extinguished simultaneously, leaving a pale yellow. At retardations above the third order, which begins at 14,000 angstroms, one observes a repetition of white, pink and green with each succeeding multiple of 7,000 angstroms. The colors become still more dilute with increasing retardation. Ultimately they degenerate into off-white.

"Experiments with the polariscope might begin with clear plastic. As I have mentioned, cellophane (a regenerated cellulose) is an excellent material. Certain other packaging films are birefringent; they include polyester and polypropylene. Before the era of plastics slices of mica and tourmaline were popular for specimens.

"Cellophane of the thickness normally found in packaging causes a relative retardation of about 3,000 angstroms per layer. In other words, an unstressed layer of the material appears white when it is



Patterns of strain frozen in glass

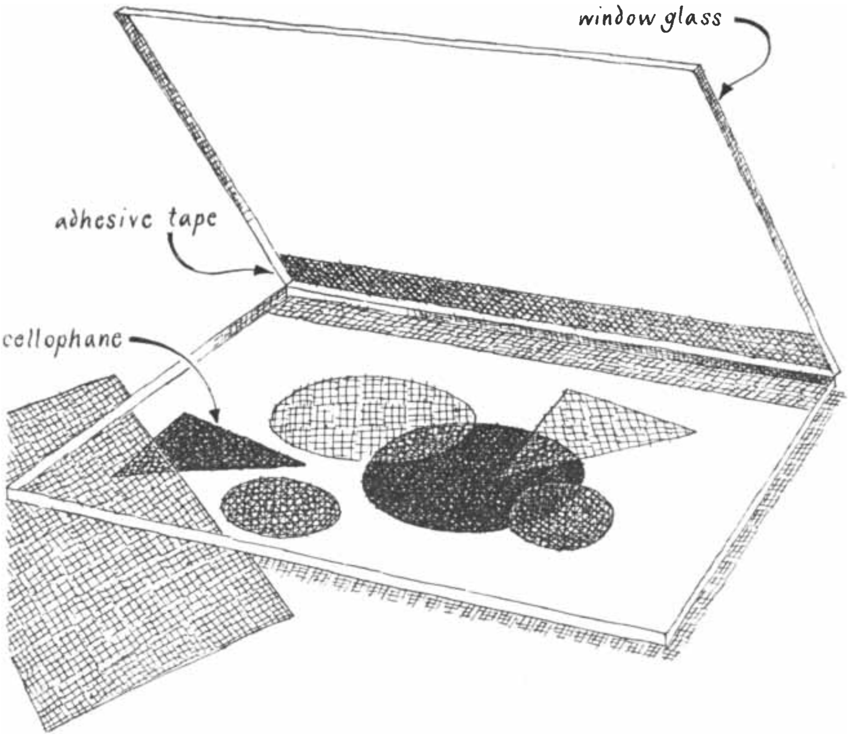
placed between the polarizer and the analyzer of the polariscope at an angle of about 45 degrees. Some samples exhibit more birefringence than others. Two layers aligned in the same direction appear blue. Three layers show orange as a second-order color.

"Since the amount of retardation introduced by a layer varies from specimen to specimen, combinations that exhibit intermediate colors can be found. For example, a double layer combining 2,000 and 3,000 angstroms of retardation yields a retardation of 5,000 angstroms and appears green. I assembled a reference device by superposing 10 layers of cellophane in the form of a book, so that each lower 'page' of film was somewhat longer than the films above it. The exposed, multilayered edges served as a reference for combining layers of cellophane to display any desired color. For protection the layers are bound between sheets of glass that are hinged with adhesive tape. The sheets measure four by six inches.

"The experimenter will discover that only certain kinds of transparent plastic film exhibit birefringence. The property usually appears when the polymer molecules are aligned to a greater or lesser extent during manufacture. Polyethylene film, which has a soft, waxy feel and does not emit a crackling sound when it is crushed, is not birefringent. It can be made birefringent, however, by stretching.

"Place a strip of polyethylene about eight inches long and an inch wide at an angle of 45 degrees in the polariscope. (Some plastic bags of the kind sold for protecting sandwiches are made of polyethylene.) Grasp the ends and stretch the strip. A reversible color will appear when the material is stretched lightly. A range of first-order colors can be observed by applying additional stress. Permanent molecular orientation and birefringence can be induced by stretching the film beyond its elastic limit.

"Photoelasticity can be demonstrated by a specimen of translucent rubber or glass. I cut a strip one by eight inches in



Cellophane patterns with protective glass

size from a surgical glove of polyvinyl chloride and stretched it in the polariscope at an angle of 45 degrees with respect to the optical axis of the polarizer and the analyzer. Even though I stretched the specimen quite forcefully, it remained white, indicating a retardation of less than 2,000 angstroms.

"I then laid a square of cellophane on the polarizer at the 45-degree orientation. From previous experiments I knew that the cellophane introduced a retardation of 3,000 angstroms. I again stretched the strip of polyvinyl chloride. Its color passed from yellow through purple to blue, corresponding to a total retardation of 6,500 angstroms, including the contribution by the specimen of a stress-induced retardation of 3,500 angstroms.

"The coefficients of stress and retardation of plastics are much higher than those of glass. For studying photoelastic effects the space between the polarizer and the analyzer may be limiting. If so, remove the analyzer and view the specimen through a pair of Polaroid sunglasses.

"In engineering circles it is well known that regions of high concentration of stress exist around notches and holes in a structure. The effect can be demonstrated with the polariscope and a strip of translucent rubber. At the center and at one side of a six-inch strip about an inch wide and .03 inch thick I cut a notch with a radius of about .25 inch. Then I put in the polariscope the sheet of cellophane previously employed to introduce a retardation of 3,000 angstroms.

"When I stretched the strip, its color (in combination with the cellophane) was orange, indicating a total retardation of 4,500 angstroms, or a net contribution by the specimen of 1,500 angstroms. At this point the notch just began to show a blue color, indicating a total retardation of 6,500 angstroms, or a net increase of 3,500 angstroms of retardation by the rubber at the site of the notch. Hence a stress concentration of 3,500/1,500, or 2.3, is indicated. That is close to the figure predicted by theory. Higher stress concentrations would be created by sharper or longer notches.

"This experiment raises the question of stresses that are frozen into certain materials, as represented by nonuniform birefringence and color patterns that can be observed in substantially all injection-molded objects of transparent plastic. These objects include dishes, medicine vials and the handles of toothbrushes. Nearly all of them display a profusion of colors in the polariscope.

"The point of mold injection can be ascertained by the higher orders of color (pink and green) that usually surround it. The retardation results from stresses that formed when the article cooled in the mold. The stresses can be relieved by immersing the plastic in boiling water for a few minutes and letting it cool in air; the colors then vanish.

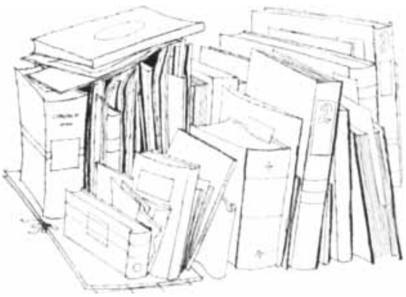
"The polariscope holds much fascination for people who are artistically inclined. Many interesting patterns appear when cellophane packets are viewed between the crossed Polaroids. (Cellophane is difficult to buy alone now, but it can be obtained from packaging materials.) Interesting abstract designs can be made by cutting pieces of cellophane to a desired shape and superposing them in layers. During this work the reference device I have mentioned is helpful in compounding colors. It is possible to cement layers together before cutting them and thus prevent slipping.

"This art form has produced some impressive works. A schoolboy in a community near mine based a large picture on the brass figure of a Crusader in a nearby church. The entire figure was made of multilayered cellophane segments carefully executed to create the desired forms and colors. The effect resembles a stained-glass window.

"Large pictures can be illuminated by polarized light from a 35-millimeter slide projector fitted with a Polaroid slide. The pictures are viewed with either Polaroid glasses or a small square of Polaroid held in the hand. Slides of convenient size can be made for larger projectors that can be assembled at home. In this scheme the slide carrier is sandwiched between the polarizer and the analyzer in the optical train. The lamp housing must include a blower to cool the polarizer if the lamp exceeds 100 watts."

RETARDATION (ANGSTROMS)	COLOR OBSERVED	COLOR EXTINGUISHED	ORDER
2,000	WHITE	———	1ST
4,000	YELLOW	VIOLET	1ST
4,500	ORANGE	BLUE	1ST
5,000	RED	GREEN	1ST
5,900	PURPLE	YELLOW	1ST
6,500	BLUE	ORANGE	1ST
7,000	GREEN	RED	1ST

Effect of relative retardation in birefringent substances



BOOKS

Geomythology (including Velikovsky) and the energetic history of the U.S.

by Philip Morrison

LEGENDS OF THE EARTH: THEIR GEOLOGIC ORIGINS, by Dorothy B. Vitaliano. Indiana University Press ($12.50). Hot-tempered Pele, goddess of volcanoes, dwells to this day with her many relatives in the fire pit called Halemaumau, deep in the caldera of Kilauea in the Hawaii National Park. Their shifting dwellings are the lava mounds built up when a fountain of fire jets high above the surface of the lava lake. For a century before 1924 Pele's lake of flame rose and fell continually: fiery surf for family sport. Since then the climate has grown less favorable. A congealed gray layer conceals the glow; only during specific eruptions do fountains leap and do flung bombs bear witness to the games and the temper of the gods. Not since 1790 has angry Pele hurled a violent explosion from her abode. Then she acted in defense of the ruler of Hawaii, her protégé Kamehameha, who in the end came to hold sway over all the islands. She reduced the rebellious army of the great chief's cousin Keoua by a third, engulfing it with a rare blast of steam, ash and rock just as it passed her crater. An entire day spent in appeasing her had been inadequate to win her neutrality.

Pele has not always lived in Kilauea. When she first came to the islands, it was to Kauai in the northwest, where she left only an eroded core of a volcanic cone, now long extinct. Next she went to Oahu, where a fire pit she left became filled with salt water to make a lake, and then, island by island, she moved southeastward, at last to dwell in fiery peace near the tip of Hawaii. She still leaves Halemaumau from time to time, but only by "a road under the ground from her house in the crater to the shore." It is an apt enough account of flank eruptions, the latest of which began in 1969 and is still in progress.

The dramatic and coherent legend of Pele is a fine example of geomythology, the subject of this unusually interesting book. Pele's story includes both types of earth legend: the euhemerist, which incorporates actual observed geological events and is a kind of folk field report, and the causal, or etiological, which is rather a folk theory conceived to account for features of the environment, from odd minerals such as tektites to landforms, floods, volcanoes and earthquakes. Many photographs (some taken by the geologist author herself), detailed maps and drawings enrich this book, whose clarity and liveliness are no less visible than the informed scholarship manifested in a multilingual list of references citing 277 books and papers.

The folklore (and some fakelore) presented is catholic indeed, spanning the world. Iceland, Japan and Hawaii are evidently particularly rich veins, and they are well worked here. The lore does not stand alone; it is supported by detailed and skeptical appraisal of the real geological situations, both in the relevant locality and as more general analogy. One chapter makes it clear that the slow motion of most geologic process has not come to the attention of the legend makers. "Man adapts to the slow-but-sure changes so naturally that in most cases he forgets them completely."

In almost every part of the world there are traditions of a great flood, such as the deluge Noah alone escaped. Once even geologists believed that the marine fossils of the high mountains bore mute witness to such a worldwide event. It is now clear that the imprints refer to no single time, and even clearer that worldwide rains of the most remarkable volume and simultaneity could at most flood many large rivers but never the sea.

Immanuel Velikovsky is the most important remaining proponent of the folkloric evidence of a universal flood (which, since it cannot be of normal geological origin, he ascribes to "worlds in collision"). Tsunamis in the wide Pacific littoral, meltwater floods from postglacial lakes and disastrous river floods augmented by typhoon conditions are some more likely events behind the tales. For flood stories are not literally worldwide; they are conspicuously lacking in Africa, where the main rivers are seasonal and floods such as the Nile's are often not evil but benign. "Velikovsky gets around the lack of African flood traditions with an ingenuity that must be admired: he invokes a 'collective amnesia.'" The verdict is plain. It is just as implausible to ascribe the widespread deluge stories to an actual worldwide flood as it is to claim that all the stories derive from one diffused account of a great overflow of the Euphrates. "Flood traditions are nearly universal, partly because of the efforts of [Christian] missionaries, but mainly because floods *in the plural* are the most nearly universal of all geologic catastrophes."

Nearly half of the book is given over to a fascinating argument (owing much to the Greek scholars and scientists Spyridon Marinatos and Angelos Galanopoulos), some of it based on direct work in the field by the author (and her husband). The topic is the Bronze Age eruption of the caldera of Santorini, southernmost of the Cyclades Islands in the Aegean. There stands the sea-flooded broken ring of the volcano, 10 kilometers in diameter. Deep-sea cores show a layer of tephra, the ash of the explosion, over many hundreds of miles of sea floor. (Krakatoa's ash shows up very little in comparable cores, and so we can argue that at least one paroxysm of the event of Santorini must have been more powerful than the 100 or 150 megatons of TNT equivalent estimated for Krakatoa, a well-studied eruption of the same type.)

Santorini erupted around 1500 B.C., the dating coming most surely from the presence of Minoan pottery in "sealed Late Minoan I A" levels in easternmost Crete. (Pottery styles, but not carbon-14 dates, are specific enough to fix the complicated sequence of events.) The event of course destroyed the Minoan colony on Santorini, but not the colonies on Crete or other islands, and yet within a generation the Minoan culture had given way to the mainlanders from Mycenae. Were the mainlanders the beneficiaries

Opel. It doesn't need a lot of gas to be exciting.

The Germans are a demanding people. They expect perfection of most things—themselves, their work, their products, and of course, the automobiles they drive. And more often than not, they get it. Or at least come close.

From the country that gave you the Gullwing and the economy import

It is not surprising, therefore, that some of the fastest, most powerful cars in history have come from Germany. It is also not surprising that some of the most frugal, gas-pinching cars are likewise German in origin.

The optimum compromise: economy *and* spirit

It is equally logical that a proven best seller in such a nation must embody a combination of both. Performance *and* economy. Emotion *and* frugality. Art *and* Science.

The Opel Manta Rallye is such a blend. Its precision 1.9 liter, cam-in-head 4-cylinder engine and 4-speed transmission allow it to cope handily with the rigors of the German Autobahn. And for urban situations, there's an automatic transmission available. But in either case, the powerplant is fundamentally rugged and simple. Hence, maintenance and fuel consumption are held to practical minimums.

Old World practicality

Virtually everywhere you look, Manta Rallye is an exercise in sound, no-nonsense thinking. The hood is painted a businesslike flat-black, for efficiency as well as looks. The instrumentation is as sensuous to the eye as it is plentiful to watch. A large tachometer checks engine speed, and there are warning lights and gauges for amps and oil pressure. The front seats, in best German fashion, recline individually at the twist of a knob. And the steering wheel carries a skin of soft vinyl.

It works

The total effect is that of function. The car looks like it works. And it does work. It works because of things like front and rear stabilizer bars in the suspension and an efficient forced-air ventilation system. It works because of power front disc brakes and dozens of other things, little and big, that help make it a genuine pleasure to drive. It works because, well, it *has* to, to be the best selling car in Germany the past two years.

And the price?

The price of an Opel Manta Rallye in America is $3,584.22,* including road wheels, special instrumentation, and radio.

The standard Manta Coupe is priced at $3,298.50,* the Manta Luxus and Opel Sportwagon at $3,535.50* and $3,521.50,* respectively.

See and drive one at any of more than 2,200 Buick/Opel dealers throughout America.

BUICK
OPEL
GM

*Manufacturer's Suggested Retail Price, including dealer new vehicle preparation charges, other optional equipment, state and local taxes and transportation charges from Port of Entry are additional.

Opel. A proven best seller in Germany.

perhaps of a "knockout blow from Nature... not all at once, but in the form of a one-two or even a triple punch"? The Minoan decline was the aftermath of ashfalls that ruined crops, followed by moderate but repeated tsunamis and finally by a major earthquake that decimated fisheries, towns and royal courts. In the end there was a massive emigration and final surrender to powerful Mycenaean expeditions of conquest.

The author provides this scenario in one interesting chapter of "science fiction." The classical deluge of Deukalion, the Platonic tale of sunken Atlantis, the plagues of Egypt (ashfall and its consequences), the parting of the Red Sea (a tsunami on the Mediterranean coast of Sinai), the flight of Icarus and even more have been ascribed to Santorini. The euhemeristic relationship of all these classical and Exodus events is still to be proved; it remains possible until tested by more reliable absolute dating. "Until then in these cases... we are left swimming in a sea of speculation. But is it not a delightful sport?"

Here is excellent, skeptical, and knowing geology, fused with the sensible folklore insights so much at home at the Indiana University. Travelers, the fortunate dwellers in many interesting places, those who teach earth sciences and their more imaginative students all need this book.

ENERGY CRISES IN PERSPECTIVE, by John C. Fisher. John Wiley & Sons, Inc. ($9.95). In 1850 the U.S. was a rapidly industrializing nation. Our population was already about that of Britain; by 1875 we were to pass her to become the leading industrial producer in the world. Steamboats, locomotives, spinning mills, ironworks were the mechanical wonders of the age; the plow and the reaper were still man-held or horse-drawn. Adding up all the sources of energy—bread, hay, wood and coal alike—when Lincoln was in Congress the average American spent an energy equivalent to that yielded by burning just under six tons of coal per year. In 1970, naturally, we used more energy. Can the reader guess how much more? About twice as much per head: 13 tons of coal equivalent.

Of course, we also count up some nine times more American heads, but our per capita use has only somewhat more than doubled, notwithstanding the traffic jams, jet planes, aluminum smelters, dishwashers and television of our fuming era. In Britain or Germany even today they use a little less energy per capita than we were burning already in 1850! Only Canada among nations expends appreciably more per head than our 1850 consumption. Americans have been profligate fire makers, it is plain, from the days when we were first clearing our forests. Every farmhouse in a new clearing could afford the winter luxury of an inefficient, roaring wood fire no prince in the Old World could then command. We North Americans started out as energy spendthrifts in a generous land, and that we remain.

Per capita energy utilization remained steady in the U.S. until around 1900. By the 1920's it had risen about 40 percent to reach an uneven plateau marked by the peaks and troughs of depression and war. It began rising again steeply in the late 1950's and is still upward bound. The first rise represented the growth of urban employment at the expense of farm employment: the farmer has one heated and lighted place, his house, but the factory or office worker's family makes basic energy demands twice, at home and at work. The second rise follows the recent increase in the number of employed persons in every household, largely the movement of married women from home life to outside employment. In both events family income rises as well, and so other demands are added to the basic ones.

This unexpected result opens a concise book by a physicist and planner at the General Electric Company headquarters; he presents no official corporate view, of course, but rather a crisp, candid, reasoned and cool personal statement. No book in the flood of books touching this issue makes its case so sharply; the argument and its documentation, in tables and graphs, are indispensable to the well-informed reader, even if he can manage here and there to remain skeptical of some of Fisher's conclusions.

History is here used to project the future, but it is used with respect. Many writers have felt it was enough to extend straight line segments along semilog paper. Fisher works at a level deeper than that. For example, he breaks the recent energy-use growth rate into five measurable portions: the increase in working population off the farm, the energy content per unit of goods and services, the total population, the increasing residential use ("the thermal comforts of home") and "new" energy uses: those unknown long ago, such as lighting, electronics, air conditioning, refrigeration and clothes drying. He projects these individually to saturation; allowing for increased affluence and greater participation in the work force, he expects the final plateau to stand about 50 percent higher per capita. The increased energy costs per unit of the complex production mix (fighting against lower-quality material inputs and meeting new demands for pollution control and for safety) he trades off at par against the steady cost reduction from better technology.

Can we find 20 tons of coal per person per year? Indeed we can, Fisher says, for a big population, and even in the U.S. and even at present real prices. We went from fuelwood to coal to meet the first big rise; petroleum and gas took over from coal to nourish the second. Basing his projections on U.S. Geological Survey estimates of resources in the ground (in our own ample territories and offshore), he foresees enough crude oil for decades, enough coal for centuries and nuclear resources for a million years. We ought to see the price of domestic crude oil, which has shown a fluctuating but persistent fall over the entire history of the industry, continue its decline. Only now should it slowly begin to reflect the inevitable depletion of the physical resources, which in 1970 were perhaps depleted only by a few percent.

The trend he documents is a 5 percent price decline, based on economies of technology and scale, with each doubling of "experience": the total cumulative production. The rise in depletion is counterbalanced by this decline, putting it in a long flat trough of price. Our wells ought to yield oil at less than $4 (1970 dollars) per barrel for a couple of doublings yet, or into the next century. Meanwhile synthetic crude oil from coal or shale should become economically ready to take over, a supply fit for a longer term still. These "learning curve" projections are economically plausible, if not certain; they are far more persuasive than straight-line fits against time. The externalities Fisher postulates are not unrealistic or extreme: there would be no burning of unrefined fuels and all land disturbed by mining would be reclaimed. These changes he credits to environmental concern.

His reliance on continuity makes one political assumption: it is that national policy will continue to protect marginal U.S. producers by the use of quotas and tariffs against irreversible competition from low-cost but uncertain energy imports. Here the Organization of Petroleum Exporting Countries (OPEC) and its customers share some interest in keeping prices from declining sharply to the level of the real cost of oil from those giant, low-cost, rationally developed fields of the Middle East. One can estimate the incremental supply cost (in 1970 dollars) of a barrel of oil as 10 cents at dockside in the Persian Gulf, "including a 20 percent return on all necessary investments." The smaller oilfields of the U.S. and the crowded, competitive U.S. wells cannot match that even in order of magnitude.

The last part of Fisher's little volume concentrates on electrical energy. His projections of future electric-power generation spread over a factor of three by the year 2000. He does not devote much space to strictly technological questions; his emphasis is more on commonsense economic analysis than on engineering decisions, although he lists and comments briefly on a wide range of possible technological novelties. He sees "no chance at all" for fusion at competitive prices in this century. His bets for the next decade are gas turbines for peak-load generation, many nuclear base-load plants and heat pumps for economical electric heating and cooling. Hay-burners (not four-footed ones but large-scale steam boilers fueled by hay produced from special strains bred for combustion and not for feed, with low nitrogen and high cellulose) are his guess as providing the most promising new path to solar-energy utilization for electric power! (He does not consider decentralized solar space heating and water heating.) Hay fields collect sunlight on clear and cloudy days alike, whereas lenses do not. Solar steam can compete on even terms with fossil-fuel steam once hay costs 60 cents per million B.t.u. The going price is now $2 per million B.t.u., but that is for a hay to be oxidized by a distinct and fastidious combustion process: animal metabolism. "I believe it is possible, I do not expect that it will happen." All these optional competitive technologies will ensure, he expects, the growth of electrification at a declining real cost.

This book, founded on the demonstrated consequences of a less certain continuity, shares no part of the ambience of alarm that fills our days. Perhaps it is too cool; perhaps real discontinuities are about to appear in the technological, economic and social parameters that abstractly measure our history. Important, if transient, crises are yet to come, like Suez in 1967 and the oil embargo of 1973, or worse. The argument stands; under conditions sensibly assumed, "the United States energy economy, with its moderate fuel prices and high technology content, might be expected to continue its historic evolution." These points are well supported. Still, the plight of the poorer world, from Newark to Calcutta, and the growing energy problem not of feeding boilers but of properly feeding human beings remain outside this work, yet within the mind of every thoughtful reader.

PATTERNS IN NATURE, by Peter S. Stevens. Atlantic Monthly Press Book. Little, Brown and Company ($10). There is a certain body of knowledge and lore, less than a science yet more than a hobby, that has claimed devotees for a long time. It is the study of perceived form in nature, on every scale, into every material, with an eye for the unity of the disparate: the milk poured into a draining slate sink looks like a spiral galaxy. More than one inarticulate artist, more than one all too articulate philosophical systematist, have produced volumes either to celebrate or to generalize grandly on such resemblances. Consequential studies exist as well, those in which the actual theory of the form is sought and resemblances are explained by the physical principles that underlie the formations.

By now there is a classical corpus on this subject, dear to its adepts and widely recognized among scientific readers. Its theory is a pungent mix of topology and surface energies, of fluid mechanics and crystal growth, of osteology, stress analysis and the calculus of variations. A sample catalogue of specimens strikes a familiar note: the occasional pentagon in a hexagonal network, as in a radiolarian or a geodesic dome; the open microstructure of human bone and the tensile structures of Frei Otto; the infolded outline of an "absorbing" region, whether the intestinal lining or the piers of a busy port; the shell of a box turtle and a symmetrical cluster of 13 soap bubbles; the space-filling truncated octahedrons of Lord Kelvin and the froth on a beer; the crackle glaze of a Chinese bowl and the shrinking of a blob of poster paint as it dries; snail and ram's horn; Hortonspheroids (the patented oblate storage tanks that are built by the Chicago Bridge and Iron Company) and a drop of mercury....

Those who love forms might be dubbed with the nonce word philomorph; indeed, at Harvard there has met for some years a group of enthusiasts that goes by that name: architect-designers, historians of science and of art, a physicist or two, a geographer and so on. Among these the author, himself an architect, painter, photographer and acoustician, found the stimulus for this rich book. Just about every one of the items listed above is shown here in a handsome photograph along with some account of the abstract structures behind the physical form. The well-made volume is a long pleasure to look at; there are 180 numbered figures, most of them organized photomontages or multiple drawings.

Stevens has a flair for pedagogy, and his text is aimed at those who know little or no mathematics. A few diagrams and tables, a little geometry and a handful of equations at about the level of difficulty of the theorem of Pythagoras are all the equipment demanded of a reader; Stevens seeks to explain a great deal of the pattern in nature with no more tools than that. His quarry is not petty. The chapter topics make clear how he seeks to net the world. We begin with the effect of scale and the simplest geometry of surfaces and volumes in three-space. Then we examine a little topology, the lengths of networks and the relations of points and lines. We consider fluid flow, both laminar and turbulent, with examples from electrostatics to bone and galaxy. A little is done with spirals and their kin; group theory and space-filling are only alluded to. The most original chapters deal with branching and models of the process as we find it in rivers, sparks, arterial nets and trees. Here Stevens has assembled a good deal of powerful recent theory, not known to D'Arcy Thompson and the older philomorphs. He goes on to growth patterns, to very well done sections on bubbles and froth and on cracks and packing, from dried peas to continents.

His is a grand design; his own photographs and others he has collected from the diverse literature present it attractively. Some topics, such as the relation of bubble curvature and internal pressure, he explains more clearly than anyone has done for the nonmathematical. The book is a success, and it deserves a wide and contented readership.

Yet there is a heavy caveat. Stevens' captions and his conclusions are too elliptical. He is very often in waters so deep he cannot stand. He provides many explanations he has himself seen into only quite murkily, like that of the scale-invariant Reynolds number, or the vague account of the averaging unity behind the Laplacean operator. He seems not explicitly to have heard of the foundations of thermodynamics and its profound relation between energy minimums and order, although he senses it. In short, he cannot guide the beginner as far as he claims. Every philomorph will want the book; its display of examples and its other good points are too many to neglect. But those who come for the first time to these wonders will find themselves puzzled nearly as often as they are enlightened. Stevens has constructed a most handsome edifice, but it has many blind corridors.

THE PARTICLE ATLAS, EDITION TWO: AN ENCYCLOPEDIA OF TECHNIQUES FOR SMALL PARTICLE IDENTIFICATION, by Walter C. McCrone and John Gustav Delly. Ann Arbor Science Publishers, Inc. ($240). In any ordinarily clean urban room a rain of dust, about 1,000 particles larger than five microns across, is "settling every hour on each square centimeter of bench surface." Each particle has its history written in the myriad of atoms that make up even so small a sample. About seven years ago the first edition of this encyclopedia appeared, then in one volume instead of four, outlining the remarkable detective skill of the senior author and his colleagues at deducing particle biographies in their pioneering Chicago laboratory. In those days the central tool was the polarizing optical microscope, with its requisite file collection of identified particles. The atlas presented that file in some 500 color photographs, an attractive field guide to the dust. The power of analysis shown with these 10^{-11}-gram samples, using elegant refractive index and birefringence measurements to augment visible color, form and texture, was all but incredible. It was routine to spot the pink eraser particles in office dust and to part the fly ash of a chain-grate stoker from that of a spreader. Plenty of tables and graphs supplemented the pictorial guide for dust detectives.

The art has matured since 1967, stimulated by the profound concern of the past years, a concern less with the "clean room" of the laboratory than with dirty living. Artful employment of the polarizing optical microscope and the file collection remains the center of the work. The second volume of the new edition once more presents those color photographs on a gray ground, now about 700 of them, including moon dust and red rain. (The London red rain was colored by Sahara dust, the Chicago example by iron oxide from the steel mills south of the city.) It remains true that "there are not too many particles not identified by John, his Zeiss Universal and his reference preps." But the laboratory has now gone well beyond the eye.

The electron beam, the ion beam, the X-ray beam augment the visible. A second field guide fills the third volume of the new atlas. Its pages contain some 600 samples, each seen in three black-and-white views, two usually the textured grays of the scanning electron microscope and one the oscilloscope record of an elemental analysis of the material in the field of view, performed by the scanning electron beam as it induces X-ray emission in the sample. The X rays are sorted in energy by a solid-state detector attachment and displayed by a multichannel analyzer. The steel mill dust shows a strong peak at iron, but so does Sahara dust; both were red. It is the simpler chemistry of the mill red that distinguishes it from the complex geochemistry of the North African desert.

Once a first-class dust-analyzing laboratory could be expected to cost as much as a fine automobile. Now the cost is closer to that of a small jet aircraft. The three authors of the first edition have become a party of a couple of dozen. A computer patiently runs the long repetitive scans looking for tiny needles in microhaystacks, and the problems John does not solve by eye, hand and mind are sent on to the photoelectron spectrometer, the ion microprobe (which carries out mass spectroscopy in its delicate beam), the scanning electron microscope, the transmission electron microscope, the X-ray-diffraction powder camera or the electron microprobe (a versatile beam-analysis device combining many modes of analysis, all of them responsive to the computer, "which rules the roost").

The first volume expands on the wealth of techniques and the fourth is a rich handbook of tables and charts for the particle analyst. It is delightful to see that these experts prefer to draw on the sample block with a sharp tungsten needle the orienting outline of a toy horse a millimeter high rather than use a bland grid of squares. "It is easier and faster to find... the eye of a horse than a particle 'near the right center edge of a square.'" These detective-scientists remain elegant craftsmen; they still handle their particles by hand, working with fine needles and a steady touch (the book tells you how to brace your hand) in preference to any micromanipulator.

These volumes are undoubtedly meant for the specialist, who probably cannot do without them. Where else, however, will you find micrographs of the hair of the goat, the rat, the dog, the cat, the bat, the sheep, the horse, the rabbit, the skunk and man (several varieties), plus spider web, by far the finest and smoothest of all the 70 fibers shown?

THE GRAMMAR OF ORNAMENT, by Owen Jones. Van Nostrand Reinhold Company ($40). FACSIMILE-ATLAS TO THE EARLY HISTORY OF CARTOGRAPHY, by A. E. Nordenskiöld. Translated from the Swedish original by Johan Adolf Ekelöf and Clements R. Markham. New introduction by J. B. Post. Dover Publications, Inc. ($6.95). TEXTILES OF ANCIENT PERU AND THEIR TECHNIQUES, by Raoul D'Harcourt. Edited by Grace G. Denny and Carolyn M. Osborne. Translated by Sadie Brown. University of Washington Press ($8.95). The three volumes here noticed are new in only one way: they are all freshly published editions of classical works, prized for many years and now available again at modest prices. All three center on representations of a variety of artifacts that are displayed to the eye; the texts serve to explain and annotate what the drawings, lithographs and photographs exhibit.

The big book by Owen Jones was a product—the first edition was even bigger, full folio size—of the London Exhibition of 1851. Like the Victoria and Albert Museum, to which it is in some way kin, it is a kind of legacy of that celebration, for which the architect Jones was superintendent of works. The volume we have here reproduces the many large plates, mostly painted in rich polychrome, that in 1856 were drawn "upon stone" by Francis Bedford and his assistants, who were able to "render this work as perfect as the advanced stage of chromolithography demanded." No one can turn the big pages without pleasure as this lavish cornucopia of decorated surfaces spills out Greek frets from vases and pavements, Egyptian lotus pillars, "a splendid copy of the Koran in the Mosque El Barkookeyeh," the tile mosaics of the Alhambra, a "beautiful New Zealand paddle," a page of "Typographic Embellishments of the Sixteenth Century in Italy and France" and nearly 100 more. (The 10 plates in monochrome, which present the leaves and flowers of the holly, the ivy and the vine "full size, traced from Natural Leaves," are something of a letdown.) The aesthetic theory of the text is not so much out of fashion as it is marred by the prejudice and complacency of the period, affecting even a man whose taste was as catholic as this fine display demonstrates. The group-theory examples of the Alhambra mosaics are one of the prize exhibits here, but it would be provincial to single out so few forms in a volume that brings treasure from Nineveh to Trinity College. For the size and wealth of color given, the volume is not costly.

The facsimile atlas goes back to the year 1889, when it was issued by the Swedish scholar, mineralogist and Arctic explorer Nordenskiöld nine years after his successful voyage around all Asia. (The Netherlands government awarded him the prize it had offered first in the year 1611 for the Northeast Passage!) It is with this book that the modern history of cartography began. It is a bargain of modern (and 19th-century) photolithography; the paperbound pages are 10½ inches by 16, and the maps are mostly the size of Nordenskiöld's version, although some are reduced by a few percent. The work is of course in black and white; the old maps shown are nearly all printed maps, with very few manuscript examples; the originals were woodcuts or copperplate. There are 51 big plate pages and more than 100 figures in the text. Most striking is the complete world atlas of Ptolemy, from the plates of a Rome edition of 1490. "The principles of geography may be said still to be published with Ptolemy's alphabet," Nordenskiöld wrote. Here is New England when it was Norumbega, a generation before the *Mayflower;* here is the world as mapped by Hakluyt in 1599, or "so much of the world as hath beene hitherto discovered," as he says cautiously; here is a map of Mexico City about 1550, little Aztec figures at work in the lakes. This is a map fancier's bargain.

The textile book appeared first in French in 1934 and then in English translation in 1962; both editions have been long out of print. This new paperbound version is a good buy, although it lacks the color plates that made more real the marvels of the ancient coastal Peruvian artists who, in dyed and natural cotton or in wool of the llama and its kin, achieved some of the most intricate and beautiful examples of the textile art in all history. D'Harcourt's work has the distinction of being valued no less as a technical guide and stimulus to the modern handweaver than as a contribution to the archaeology of pre-Inca America, because he teased out of the structure of the fabrics just how they were made.

This is applied topology, and readers without experience or a flair for that topic should be wary. The cloths are not all true woven textiles; they include complex knotted networks, braids and plaits "with thirty-eight strands of varying thickness in four colors . . . producing a square cross section." The yarn was mainly hand-twisted, perhaps without the use of a spindle whorl. The finest two-ply yarns in wool yielded some 340 kilometers to the kilogram, "which is quite remarkable." The embroidered figures in an extensive range of colors, forming entire friezes of tiny personages and animals in complex branching forms, are the most striking objects, to the uninitiate at least. The adept weaver may be found wonderstruck by what appears to be a much simpler cloth.

INDEX OF ADVERTISERS

JULY 1974

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

FIRE AND FIRE PROTECTION

HEAT TRANSFER IN FIRE. H. W. Emmons in *Journal of Heat Transfer,* Vol. 95, No. 2, pages 145–151; May, 1973.

AMERICA BURNING: THE REPORT OF THE NATIONAL COMMISSION ON FIRE PREVENTION AND CONTROL. U.S. Printing Office, Washington, D.C., May 4, 1973.

AN OVERVIEW OF THE FIRE BEHAVIOR OF POLYMERS. C. J. Hidalgo in *Fire Technology,* August, 1973.

THE SEX-ATTRACTANT RECEPTOR OF MOTHS

CHEMICALS CONTROLLING INSECT BEHAVIOR: SYMPOSIUM OF THE AMERICAN CHEMICAL SOCIETY. Edited by M. Beroza. Academic Press, 1970.

OLFACTORY RECEPTORS FOR THE SEXUAL ATTRACTANT (BOMBYKOL) OF THE SILK MOTH. D. Schneider in *The Neurosciences: Second Study Program,* edited by F. O. Schmitt. Rockefeller University Press, 1970.

INSECT OLFACTION. Karl-Ernst Kaissling in *Handbook of Sensory Physiology: Vol. IV/1,* edited by L. M. Beidler. Springer-Verlag, 1971.

INSECT SEX PHEROMONES. M. Jacobson. Academic Press, 1972.

GYPSY MOTH CONTROL WITH THE SEX ATTRACTANT PHEROMONE. M. Beroza and E. F. Knipling in *Science,* Vol. 177, No. 4043, pages 19–27; July 7, 1972.

HYBRID CELLS AND HUMAN GENES

HYBRIDIZATION OF SOMATIC CELLS. B. Ephrussi. Princeton University Press, 1972.

LINKAGE ANALYSIS IN MAN BY SOMATIC CELL GENETICS. Frank H. Ruddle in *Nature,* Vol. 242, No. 5394, pages 165–167; March 16, 1973.

ASSIGNMENT OF THREE GENE LOCI (PGK, HGPRT, G6PD) TO THE LONG ARM OF THE HUMAN X CHROMOSOME BY SOMATIC CELL GENETICS. Florence C. Ricciuti and Frank H. Ruddle in *Genetics,* Vol. 74, No. 4, pages 661–678; August, 1973.

LOCALIZATION AND INDUCTION OF THE HUMAN THYMIDINE KINASE GENE BY ADENOVIRUS 12. Raju Kucherlapati, James K. McDougall and Frank H. Ruddle in *Nature New Biology,* Vol. 245, No. 145, pages 170–173; October 10, 1973.

UNIFIED THEORIES OF ELEMENTARY-PARTICLE INTERACTION

RELATIVISTIC QUANTUM FIELDS. J. D. Bjorken and S. D. Drell. McGraw-Hill Book Co., Inc., 1965.

A MODEL OF LEPTONS. S. Weinberg in *Physical Review Letters,* Vol. 19, No. 21, pages 1264–1266; November 20, 1967.

THEORY OF WEAK INTERACTIONS IN PARTICLE PHYSICS. R. E. Marshak, Riazuddin and C. P. Ryan. Wiley-Interscience, 1969.

GAUGE THEORIES. E. S. Abers and B. W. Lee in *Physics Reports,* Vol. 9C, No. 1. North Holland Publishing Company, 1973.

SPONTANEOUS SYMMETRY BREAKING, GAUGE THEORIES, THE HIGGS MECHANISM AND ALL THAT. Jeremy Bernstein in *Reviews of Modern Physics,* Vol. 46, pages 7–48; January, 1974.

RECENT PROGRESS IN UNIFIED GAUGE THEORIES OF THE WEAK, ELECTROMAGNETIC AND STRONG INTERACTIONS. Steven Weinberg in *Reviews of Modern Physics,* in press.

THE GLORY

A STUDY OF THE BACKSCATTER OF LIGHT FROM A SINGLE WATER DROPLET (dissertation). T. S. Fahlen. The University of New Mexico, 1967.

NUCLEAR GLORY SCATTERING. H. C. Bryant and N. Jarmie in *Annals of Physics,* Vol. 47, pages 127–140; 1968.

RAINBOWS AND GLORIES IN MOLECULAR SCATTERING. E. A. Mason, R. J. Munn and F. J. Smith in *Endeavor,* Vol. 30, page 91; May, 1971.

A FAMILY OF PROTEIN-CUTTING PROTEINS

STRUCTURE OF THE COMPLEX FORMED BY BOVINE TRYPSIN AND BOVINE PANCREATIC TRYPSIN INHIBITOR. A. Rühlmann, B. Kukla, T. Schwager, K. Bartels and R. Huber in *Journal of Molecular Biology,* Vol. 77, No. 3, pages 417–436; July 5, 1973.

CARBON NUCLEAR MAGNETIC RESONANCE STUDIES OF THE HISTIDINE RESIDUE IN α-LYTIC PROTEASE: IMPLICATIONS FOR THE CATALYTIC MECHANISM OF SERINE PROTEASES. M. Hunkapiller, S. Smallcombe, D. Whitaker and J. Richards in *Biochemistry,* Vol. 12, No. 23, pages 4732–4743; November 6, 1973.

STRUCTURE AND SPECIFIC BINDING OF TRYPSIN: COMPARISON OF INHIBITED DERIVATIVE AND A MODEL FOR SUBSTRATE BINDING. M. Krieger, L. M. Kay and R. M. Stroud in *Journal of Molecular Biology,* Vol. 83, No. 2, pages 209–230; February 25, 1974.

SOURCES OF AMBIGUITY IN THE PRINTS OF MAURITS C. ESCHER

SOME FACTORS DETERMINING FIGURE-GROUND ARTICULATION. M. R. Harrower in *British Journal of Psychology,* Vol. 26, No. 4, pages 407–424; 1936.

PERCEPTION. Hans-Lukas Teuber in *Handbook of Physiology: Section 1—Neurophysiology, Vol. 3.* Edited by H. W. Magoun. Williams & Wilkins, 1961.

THE GRAPHIC WORK OF M. C. ESCHER. M. C. Escher. Ballantine Books, Inc., 1971.

THE WORLD OF M. C. ESCHER, edited by J. L. Locher. Harry N. Abrams, Inc., 1971.

NEW ASPECTS OF PAUL KLEE'S BAUHAUS STYLE. M. L. Teuber in *Paul Klee, Paintings and Watercolors from the Bauhaus Years, 1921–1931.* Des Moines Art Center, 1973.

THE CASTS OF FOSSIL HOMINID BRAINS

THE EVOLUTION OF THE PRIMATE BRAIN: SOME ASPECTS OF QUANTITATIVE RELATIONSHIPS. R. L. Holloway in *Brain Research,* Vol. 7, pages 121–172; 1968.

AUSTRALOPITHECINE ENDOCASTS, BRAIN EVOLUTION IN THE HOMINOIDEA, AND A MODEL OF HOMINID EVOLUTION. R. L. Holloway in *The Functional and Evolutionary Biology of Primates,* edited by R. Tuttle. Aldine Press, 1972.

EVOLUTION OF PRIMATE BRAINS: A COMPARATIVE ANATOMICAL INVESTIGATION. H. Stephan in *The Functional and Evolutionary Biology of Primates,* edited by R. Tuttle. Aldine Press, 1972.

THE AMATEUR SCIENTIST

FUNDAMENTALS OF OPTICS. Francis A. Jenkins and Harvey E. White. McGraw-Hill Book Company, Inc., 1950.

Your next can of beer may have been a Coors can that was a Falstaff can that was a Carling can.

The aluminum beer cans chilling in your cooler may have led other lives. Possibly as soft drink cans. Possibly as iced tea cans. Possibly as other beer cans. All because of aluminum's recyclability.

It's no secret that aluminum cans are popular with people. They like them because they're lightweight, chill quickly, open with a snap and are friendly to food. Unfortunately, some of these cans end up in places we'd prefer they didn't. Like streets, roads, woods and beaches. So, Alcoa's doing something.

Alcoa is supporting "Yes We Can" recycling programs in areas where there are enough aluminum cans to economically justify collection efforts. Cooperating recycling centers pay as much as 15 cents a pound for all-aluminum cans. People are picking them up. By the millions.

Through combined recycling center collections of the aluminum and beverage industries, almost 3¾ billion cans have been recovered since 1970 and over 15 million dollars have been paid to collectors.

Although aluminum is the most abundant metallic element in the earth's crust, recycling aluminum cans is a sensible way to help conserve this natural resource. It requires only 5 percent as much energy to recycle the metal in an aluminum can as it takes to make the metal in the first place.

If you would like more information on aluminum can recycling and how one community established a reclamation program, please write for our free brochures. Aluminum Company of America, 352-G Alcoa Building, Pittsburgh, PA 15219.

The reasons for using aluminum are found in aluminum itself.

ALCOA

The Dodge Colt.

For a little car, it's a lot of car.

Sure, all Dodge Colts are economical to drive. Our 4-cylinder engine sees to that.

In fact, in a mileage test conducted and sanctioned by the United States Auto Club (USAC) in January, Colt got substantially better mileage* than either Vega or Mustang II.

If that's all there were to Colt, it would seem like plenty. But Colt gives you much more.

For example, Dodge Colts have standard features such as front disc brakes, four-on-the-floor, and an adjustable steering column.

The '74 Dodge Colt. As you can see, for a little car, it is a lot of car.

Here're just a few of the features standard on Dodge Colts.

- **Thrifty four-cylinder engine.**
- **Front disc brakes.**
- **Four-speed transmission.**
- **Adjustable steering column.**
- **Flow-through ventilation.**
- **Reclining bucket seats (except coupe).**
- **Hidden radio antenna (except wagons).**

Dodge

Dodge Trucks

*The mileage you get depends on many factors including how and where you drive and the condition of your vehicle.

CHRYSLER
MOTORS CORPORATION