3087
SCIENTIFIC AMERICAN OFFPRINTS

Gauge Theories of the Forces between Elementary Particles

by Gerard 't Hooft

SCIENTIFIC AMERICAN
JUNE 1980
VOL. 242, NO. 6 PP. 104-138

PUBLISHED BY **W. H. FREEMAN AND COMPANY** 660 MARKET STREET, SAN FRANCISCO, CALIFORNIA 94104

Gauge Theories of the Forces between Elementary Particles

All the basic forces of nature are now described by theories of this kind. The properties of the forces are deduced from symmetries or regularities apparent in the laws of physics

by Gerard 't Hooft

An understanding of how the world is put together requires a theory of how the elementary particles of matter interact with one another. Equivalently, it requires a theory of the basic forces of nature. Four such forces have been identified, and until recently a different kind of theory was needed for each of them. Two of the forces, gravitation and electromagnetism, have an unlimited range; largely for this reason they are familiar to everyone. They can be felt directly as agencies that push or pull. The remaining forces, which are called simply the weak force and the strong force, cannot be perceived directly because their influence extends only over a short range, no larger than the radius of an atomic nucleus. The strong force binds together the protons and the neutrons in the nucleus, and in another context it binds together the particles called quarks that are thought to be the constituents of protons and neutrons. The weak force is mainly responsible for the decay of certain particles.

A long-standing ambition of physicists is to construct a single master theory that would incorporate all the known forces. One imagines that such a theory would reveal some deep connection between the various forces while accounting for their apparent diversity. Such a unification has not yet been attained, but in recent years some progress may have been made. The weak force and electromagnetism can now be understood in the context of a single theory. Although the two forces remain distinct, in the theory they become mathematically intertwined. What may ultimately prove more important, all four forces are now described by means of theories that have the same general form. Thus if physicists have yet to find a single key that fits all the known locks, at least all the needed keys can be cut from the same blank. The theories in this single favored class are formally designated non-Abelian gauge theories with local symmetry. What is meant by this forbidding label is the main topic of this article. For now, it will suffice to note that the theories relate the properties of the forces to symmetries of nature.

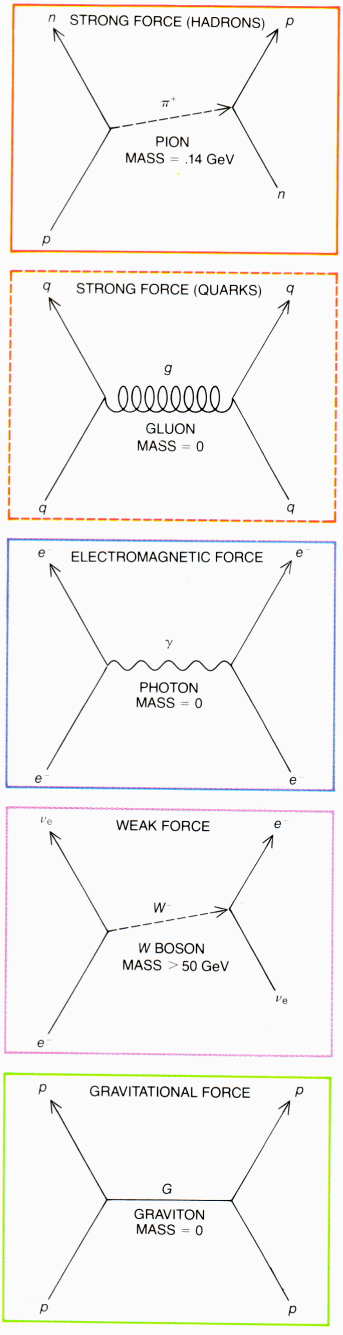
Symmetries and apparent symmetries in the laws of nature have played a part in the construction of physical theories since the time of Galileo and Newton. The most familiar symmetries are spatial or geometric ones. In a snowflake, for example, the presence of a symmetrical pattern can be detected at a glance. The symmetry can be defined as an invariance in the pattern that is observed when some transformation is applied to it. In the case of the snowflake the transformation is a rotation by 60 degrees, or one-sixth of a circle. If the initial position is noted and the snowflake is then turned by 60 degrees (or by any integer multiple of 60 degrees), no change will be perceived. The snowflake is invariant with respect to 60-degree rotations. According to the same principle, a square is invariant with respect to 90-degree rotations and a circle is said to have continuous symmetry because rotation by any angle leaves it unchanged.

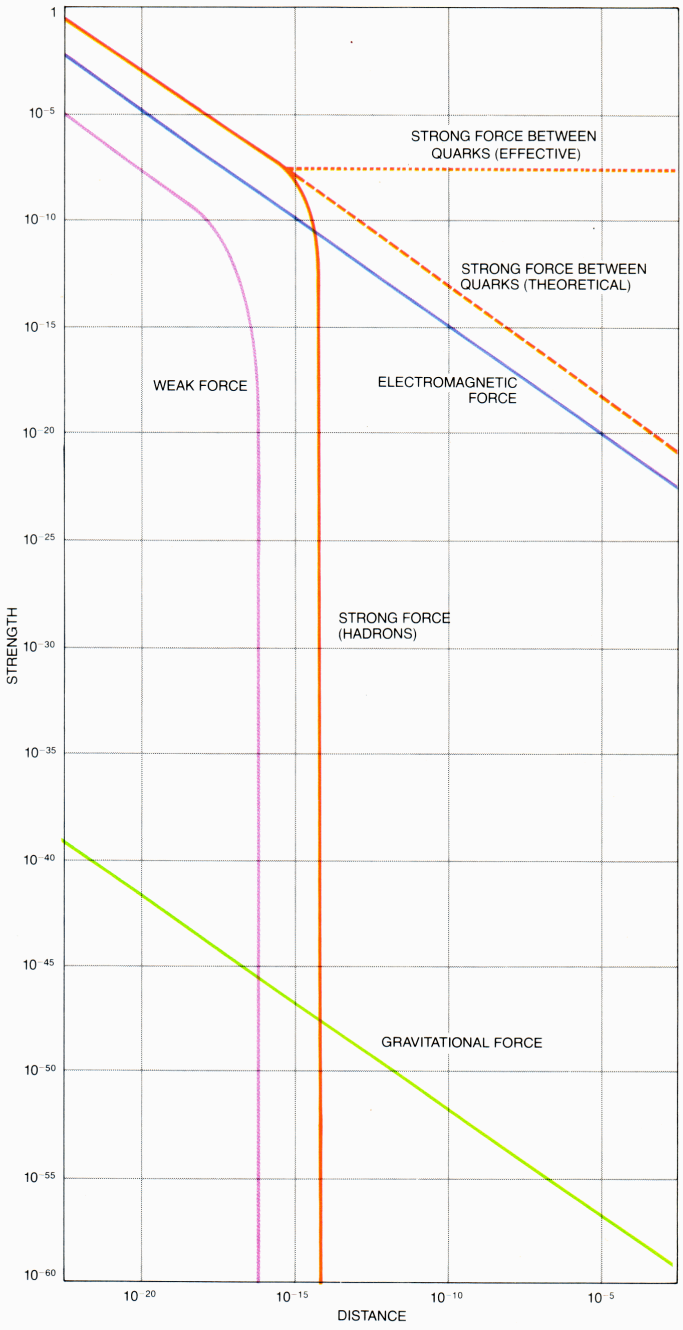
Although the concept of symmetry had its origin in geometry, it is general enough to embrace invariance with respect to transformations of other kinds. An example of a nongeometric symmetry is the charge symmetry of electromagnetism. Suppose a number of electrically charged particles have been set out in some definite configuration and all the forces acting between pairs of particles have been measured. If the polarity of all the charges is then reversed, the forces remain unchanged.

Another symmetry of the nongeometric kind concerns isotopic spin, a property of protons and neutrons and of the many related particles called hadrons, which are the only particles responsive to the strong force. The basis of the symmetry lies in the observation that the proton and the neutron are remarkably similar particles. They differ in mass by only about a tenth of a percent, and except for their electric charge they are identical in all other properties. It therefore seems that all protons and neutrons could be interchanged and the strong interactions would hardly be altered. If the electromagnetic forces (which depend on electric charge) could somehow be turned off, the isotopic-spin symmetry would be exact; in reality it is only approximate.

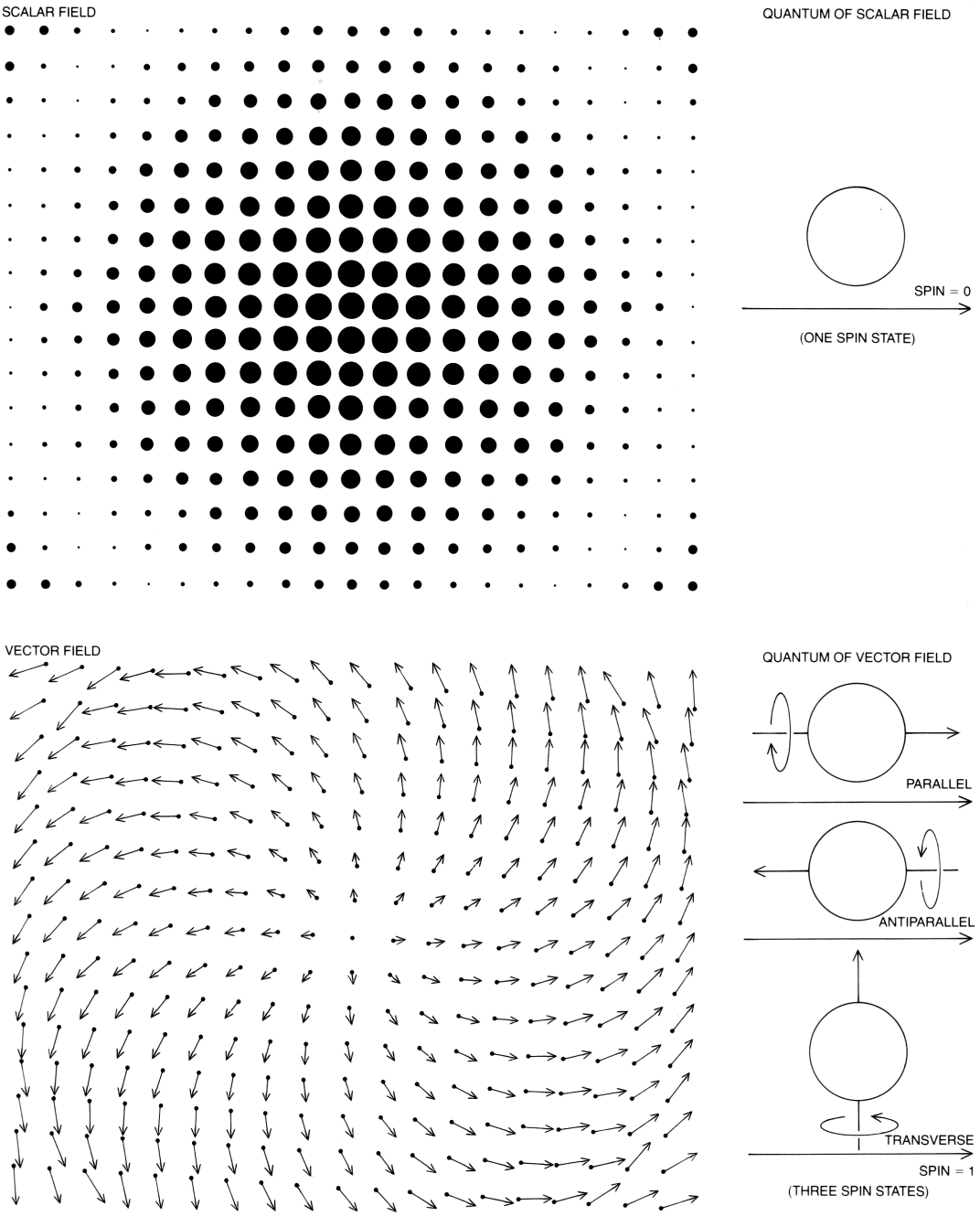
Although the proton and the neutron seem to be distinct particles and it is hard to imagine a state of matter intermediate between them, it turns out that symmetry with respect to isotopic spin is a continuous symmetry, like the symmetry of a sphere rather than like that of a snowflake. I shall give a simplified explanation of why that is so. Imagine that inside each particle are a pair of crossed arrows, one representing the proton component of the particle and the other representing the neutron component. If the proton arrow is pointing up (it makes no difference what direction is defined as up), the particle is a proton; if the neutron arrow is up, the particle is a neutron. Intermediate positions correspond to quantum-mechanical superpositions of the two states, and the particle then looks sometimes like a proton and sometimes like a neutron. The symmetry transformation associated with isotopic spin rotates the internal indicators of all protons and neutrons everywhere in the universe by the same amount and at the same time. If the rotation is by exactly 90 degrees, every proton becomes a neutron and every neutron becomes a proton. Symmetry with respect to isotopic spin, to the extent it is exact, states that no effects of this transformation can be detected.

All the symmetries I have discussed so far can be characterized as global symmetries; in this context the word global means "happening everywhere at once." In the description of isotopic-spin symmetry this constraint was made explicit: the internal rotation that transforms





FOUR BASIC FORCES mediate all known interactions among the particles of matter. The forces differ greatly in strength and effective range, but they are all described by theories of the same mathematical form, namely local gauge theories. Electromagnetism and gravitation are said to be of infinite range, although their influence declines as the square of the distance between two interacting particles. The weak force is confined to an exceedingly small range of about 10^{-15} centimeter. The properties of the strong force are somewhat more complicated. As the strong force is observed acting between hadrons, such as the proton and the neutron (*solid colored line*), it has a finite range of some 10^{-13} centimeter. The strong force also binds together the particles called quarks that make up hadrons, and in that context it could be expected to follow an inverse-square law (*broken colored line*). The actual behavior is apparently stranger: the force remains constant regardless of distance (*dotted colored line*). In quantum field theories (*diagrams at left*) the force between two particles is made manifest through the exchange of a third particle, which is called a virtual particle. The range of the force is determined by the mass of the exchanged virtual particle. Massless virtual particles, such as the photon and the graviton, give rise to forces that have infinite range.



CONCEPT OF A FIELD, a quantity defined at each point throughout some region of space and time, is important in the construction of gauge theories. A scalar field has only a magnitude at each point; in this example the magnitude is given by the area of the dots. A vector field has both a magnitude and a direction and can be illustrated by drawing an arrow at each point. A scalar field might represent a quantity such as the temperature or the density of a fluid, whereas a vector field could represent its velocity. In quantum field theories the influence of a field can be embodied in a virtual particle. The number of components in the field is reflected in the number of distinct orientations of the particle, which in turn depends on its spin angular momentum. A scalar field has just one component (its value can be given by a single number) and is represented by a spin-zero particle with one spin state, or orientation. A vector field in three-dimensional space has three components (a magnitude and two angles), and it corresponds to a spin-one particle with three spin states.

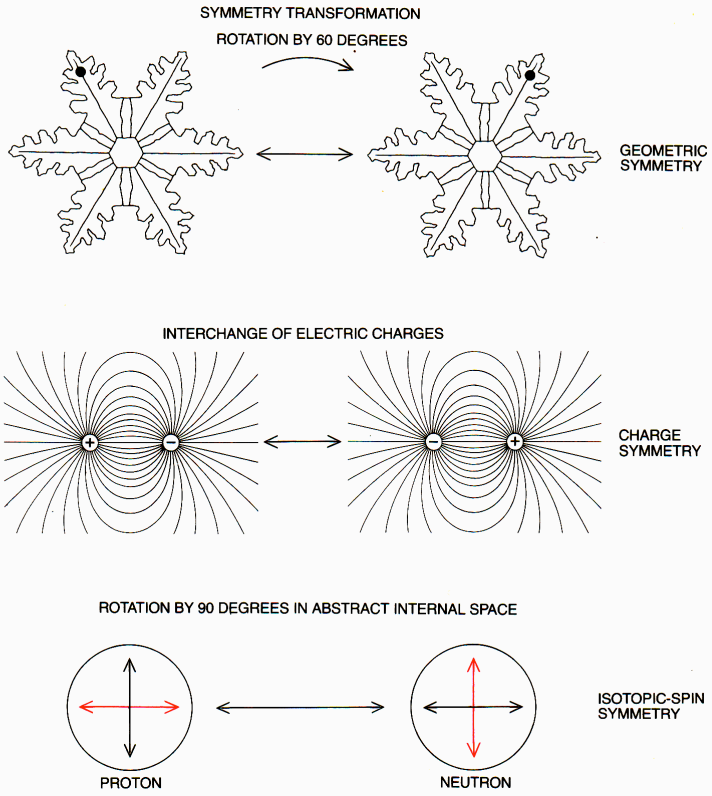
protons into neutrons and neutrons into protons is to be carried out everywhere in the universe at the same time. In addition to global symmetries, which are almost always present in a physical theory, it is possible to have a "local" symmetry, in which the convention can be decided independently at every point in space and at every moment in time. Although "local" may suggest something of more modest scope than a global symmetry, in fact the requirement of local symmetry places a far more stringent constraint on the construction of a theory. A global symmetry states that some law of physics remains invariant when the same transformation is applied everywhere at once. For a local symmetry to be observed the law of physics must retain its validity even when a different transformation takes place at each point in space and time.

Gauge theories can be constructed with either a global or a local symmetry (or both), but it is the theories with local symmetry that hold the greatest interest today. In order to make a theory invariant with respect to a local transformation something new must be added: a force. Before showing how this comes about, however, it will be necessary to discuss in somewhat greater detail how forces are described in modern theories of elementary-particle interactions.

The basic ingredients of particle theory today include not only particles and forces but also fields. A field is simply a quantity defined at every point throughout some region of space and time. For example, the quantity might be temperature and the region might be the surface of a frying pan. The field then consists of temperature values for every point on the surface.

Temperature is called a scalar quantity, because it can be represented by position along a line, or scale. The corresponding temperature field is a scalar field, in which each point has associated with it a single number, or magnitude. There are other kinds of field as well, the most important for present purposes being the vector field, where at each point a vector, or arrow, is drawn. A vector has both a magnitude, which is represented by the length of the arrow, and a direction, which in three-dimensional space can be specified by two angles; hence three numbers are needed in order to specify the value of the vector. An example of a vector field is the velocity field of a fluid; at each point throughout the volume of the fluid an arrow can be drawn to show the speed and direction of flow.

In the physics of electrically charged objects a field is a convenient device for expressing how the force of electromagnetism is conveyed from one place to another. All charged particles are supposed to emanate an electromagnetic field; each particle then interacts with the sum of all the fields rather than directly with the other particles.

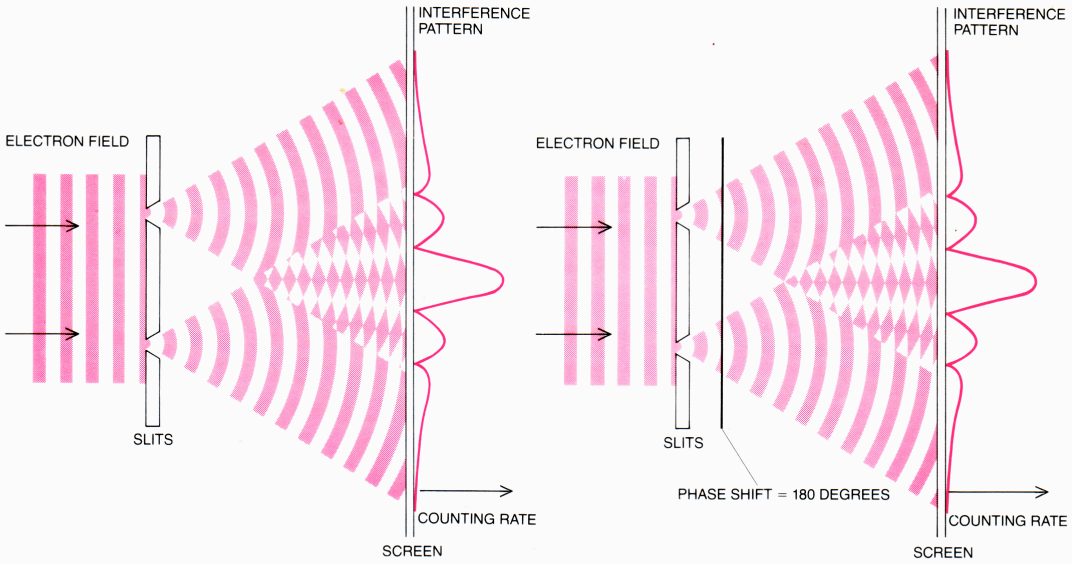


SYMMETRIES OF NATURE determine the properties of forces in gauge theories. The familiar symmetry of a snowflake can be characterized by noting that the pattern is unchanged when it is rotated 60 degrees; the snowflake is said to be invariant with respect to such rotations. In physics nongeometric symmetries are introduced. Charge symmetry, for example, is the invariance of the forces acting among a set of charged particles when the polarities of all the charges are reversed. Isotopic-spin symmetry is based on the observation that little would be changed in the strong interactions of matter if the identities of all protons and neutrons were interchanged. Hence proton and neutron become merely the alternative states of a single particle, the nucleon, and transitions between the states can be made (or imagined) by adjusting the orientation of an indicator in an internal space. It is symmetries of this kind, where the transformation is an internal rotation or a phase shift, that are referred to as gauge symmetries.

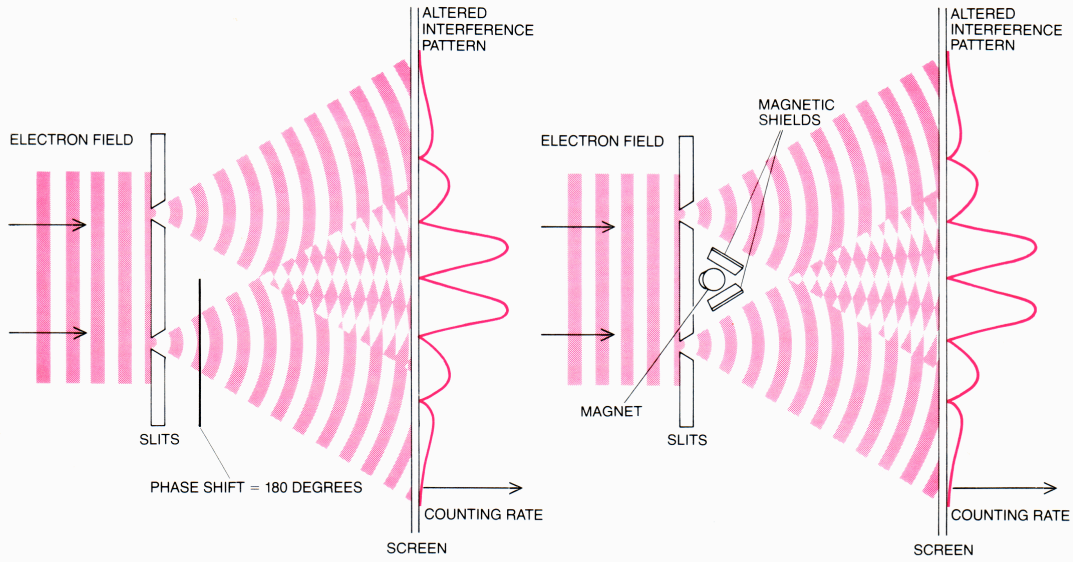
In quantum mechanics the particles themselves can be represented as fields. An electron, for example, can be considered a packet of waves with some finite extension in space. Conversely, it is often convenient to represent a quantum-mechanical field as if it were a particle. The interaction of two particles through their interpenetrating fields can then be summed up by saying the two particles exchange a third particle, which is called the quantum of the field. For example, when two electrons, each surrounded by an electromagnetic field, approach each other and bounce apart, they are said to exchange a photon, the quantum of the electromagnetic field.

The exchanged quantum has only an ephemeral existence. Once it has been emitted it must be reabsorbed, either by the same particle or by another one, within a finite period. It cannot keep going indefinitely, and it cannot be detected in an experiment. Entities of this kind are called virtual particles. The larger their energy, the briefer their existence. In effect a virtual particle borrows or embezzles a quantity of energy, but it must repay the debt before the shortage can be noticed.

The range of an interaction is related to the mass of the exchanged quantum. If the field quantum has a large mass, more energy must be borrowed in order to support its existence, and the debt must be repaid sooner lest the discrepancy be discovered. The distance the particle can travel before it must be re-



GAUGE SYMMETRY OF ELECTROMAGNETISM is an invariance with respect to shifts in the phase of the matter field that represents an electron. The phase itself cannot be measured, but it has an influence on such observable quantities as the interference pattern formed when the waves of an electron field pass through a pair of slits. The peaks in this pattern are found wherever the waves are in phase, and the nodes are found where the waves are out of phase. A shift in phase greatly alters the configuration of the field, but it leaves the observable interference pattern unchanged. The symmetry is an exact one, so that the phase shift cannot be detected. It is therefore only a matter of convention what phase is chosen in any theoretical description of the field. In the absence of forces acting between the electrons, however, the symmetry is a global one: the observed pattern is invariant only if the same phase shift is applied everywhere.



LOCAL GAUGE SYMMETRY of the electron matter field is restored when magnetic fields are taken into account. Shifting the phase of one diffracted electron beam but not the other clearly alters the observed interference pattern (*diagram at left*). The same effect can be obtained, however, by introducing a small magnetic field perpendicular to the electron beam and between the slits (*diagram at right*). Remarkably, the magnetic field induces the phase shift even when shields are arranged so that the field cannot penetrate the region where the electron waves propagate and interfere. An experimenter examining the interference patterns could not distinguish between the effects of a phase shift imposed arbitrarily on one electron beam and the effects of a magnetic field introduced between the slits. Any local shift in the phase of the electron matter field could therefore be reproduced by electric and magnetic fields, and so the phase of the electron field is arbitrary. The theory that combines electron matter fields with electric and magnetic fields is quantum electrodynamics.

absorbed is thereby reduced and so the corresponding force has a short range. In the special case where the exchanged quantum is massless the range is infinite.

The number of components in a field corresponds to the number of quantum-mechanical states of the field quantum. The number of possible states is in turn related to the intrinsic spin angular momentum of the particle. The spin angular momentum can take on only discrete values; when the magnitude of the spin is measured in fundamental units, it is always an integer or a half integer. Moreover, it is not only the magnitude of the spin that is quantized but also its direction or orientation. (To be more precise, the spin can be defined by a vector parallel to the spin axis, and the projections, or components, of this vector along any direction in space must have values that are integers or half integers.) The number of possible orientations, or spin states, is equal to twice the magnitude of the spin, plus one. Thus a particle with a spin of one-half, such as the electron, has two spin states: the spin can point parallel to the particle's direction of motion or antiparallel to it. A spin-one particle has three orientations, namely parallel, antiparallel and transverse. A spin-zero particle has no spin axis; since all orientations are equivalent, it is said to have just one spin state.

A scalar field, which has just one component (a magnitude), must be represented by a field quantum that also has one component, or in other words by a spin-zero particle. Such particles are therefore called scalar particles. Similarly, a three-component vector field requires a spin-one field quantum with three spin states: a vector particle. The electromagnetic field is a vector field, and the photon, in conformity with these specifications, has a spin of one unit. The gravitational field is a more complicated structure called a tensor and has 10 components; not all of them are independent, however, and the quantum of the field, the graviton, has a spin of two units, which ordinarily corresponds to five spin states.

In the cases of electromagnetism and gravitation one further complication must be taken into account. Since the photon and the graviton are massless, they must always move with the speed of light. Because of their velocity they have a property not shared by particles with a finite mass: the transverse spin states do not exist. Although in some formal sense the photon has three spin states and the graviton has five, in practice only two of the spin states can be detected.

The first gauge theory with local symmetry was the theory of electric and magnetic fields introduced in 1868 by James Clerk Maxwell. The foundation of Maxwell's theory is the proposition that an electric charge is surrounded by an electric field stretching to infinity, and that the movement of an electric charge gives rise to a magnetic field also of infinite extent. Both fields are vector quantities, being defined at each point in space by a magnitude and a direction.

In Maxwell's theory the value of the electric field at any point is determined ultimately by the distribution of charges around the point. It is often convenient, however, to define a potential, or voltage, that is also determined by the charge distribution: the greater the density of charges in a region, the higher its potential. The electric field between two points is then given by the voltage difference between them.

The character of the symmetry that makes Maxwell's theory a gauge theory can be illustrated by considering an imaginary experiment. Suppose a system of electric charges is set up in a laboratory and the electromagnetic field generated by the charges is measured and its properties are recorded. If the charges are stationary, there can be no magnetic field (since the magnetic field arises from movement of an electric charge); hence the field is purely an electric one. In this experimental situation a global symmetry is readily perceived. The symmetry transformation consists in raising the entire laboratory to a high voltage, or in other words to a high electric potential. If the measurements are then repeated, no change in the electric field will be observed. The reason is that the field, as Maxwell defined it, is determined only by differences in electric potential, not by the absolute value of the potential. It is for the same reason that a squirrel can walk without injury on an uninsulated power line.

This property of Maxwell's theory amounts to a symmetry: the electric field is invariant with respect to the addition or subtraction of an arbitrary overall potential. As noted above, however, the symmetry is a global one, because the result of the experiment remains constant only if the potential is changed everywhere at once. If the potential were raised in one region and not in another, any experiment that crossed the boundary would be affected by the potential difference, just as a squirrel is affected if it touches both a power line and a grounded conductor.

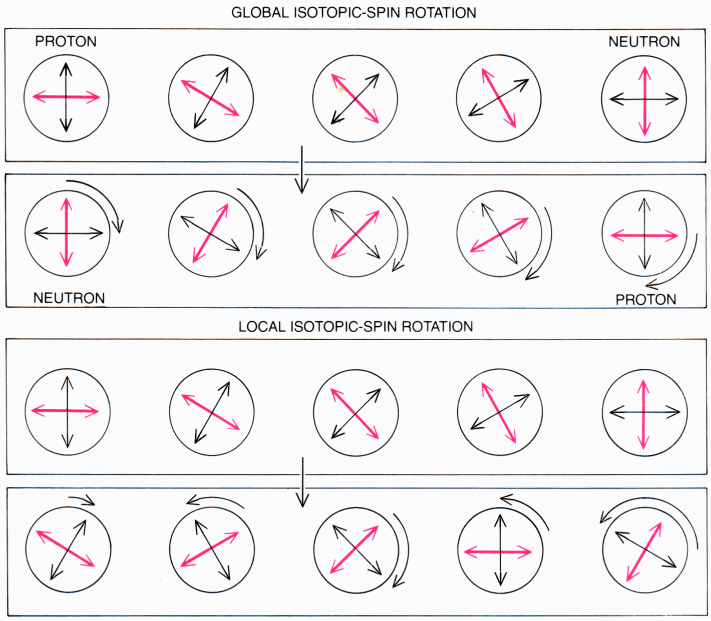
A complete theory of electromagnetic fields must embrace not only static arrays of charges but also moving charges. In order to do that the global symmetry of the theory must be converted into a local symmetry. If the electric field were the only one acting between charged particles, it would not have a local symmetry. Actually when the charges are in motion (but only then), the electric field is not the only one present: the movement itself gives rise to a second field, namely the magnetic field. It is the effects of the magnetic field that restore the local symmetry.

Just as the electric field depends ultimately on the distribution of charges but can conveniently be derived from an electric potential, so the magnetic field is generated by the motion of the charges but is more easily described as resulting from a magnetic potential. It is in this system of potential fields that local transformations can be carried out leaving all the original electric and magnetic fields unaltered. The system of dual, interconnected fields has an exact local symmetry even though the electric field alone does not. Any local change in the electric potential can be combined with a compensating change in the magnetic potential in such a way that the electric and magnetic fields are invariant.

Maxwell's theory of electromagnetism is a classical or non-quantum-mechanical one, but a related symmetry can be demonstrated in the quantum theory of electromagnetic interactions. It is necessary in that theory to describe the electron as a wave or a field, a convention that in quantum mechanics can be adopted for any material particle. It turns out that in the quantum theory of electrons a change in the electric potential entails a change in the phase of the electron wave.

The electron has a spin of one-half unit and so has two spin states (parallel and antiparallel). It follows that the associated field must have two components. Each of the components must be represented by a complex number, that is, a number that has both a real, or ordinary, part and an imaginary part, which includes as a factor the square root of -1. The electron field is a moving packet of waves, which are oscillations in the amplitudes of the real and the imaginary components of the field. It is important to emphasize that this field is not the electric field of the electron but instead is a matter field. It would exist even if the electron had no electric charge. What the field defines is the probability of finding an electron in a specified spin state at a given point and at a given moment. The probability is given by the sum of the squares of the real and the imaginary parts of the field.

In the absence of electromagnetic fields the frequency of the oscillations in the electron field is proportional to the energy of the electron, and the wavelength of the oscillations is proportional to the momentum. In order to define the oscillations completely one additional quantity must be known: the phase. The phase measures the displacement of the wave from some arbitrary reference point and is usually expressed as an angle. If at some point the real part of the oscillation, say, has its maximum positive amplitude, the phase at that point might be assigned the value zero de-



ISOTOPIC-SPIN SYMMETRY serves as the basis of another gauge theory, first discussed in 1954 by C. N. Yang and Robert L. Mills. If isotopic-spin symmetry is valid, the choice of which position of the internal arrow indicates a proton and which a neutron is entirely a matter of convention. Global symmetry (*upper diagram*) requires the same convention to be adopted everywhere, and any rotation of the arrow must be made in the same way at every point. In the Yang-Mills theory isotopic spin is made a local symmetry (*lower diagram*), so that the orientation of the arrow is allowed to vary from place to place. In order to preserve the invariance of all observable quantities with respect to such local isotopic-spin transformations it is necessary to introduce at least six fields, corresponding to three massless vector particles, or vector bosons. One of these particles can be identified as the photon; the other two carry electric charge. The theory has been influential, but in its original form it was unrealistic. It makes protons and neutrons indistinguishable and predicts massless charged particles that do not exist.

grees. Where the real part next falls to zero the phase is 90 degrees and where it reaches its negative maximum the phase is 180 degrees. In general the imaginary part of the amplitude is 90 degrees out of phase with the real part, so that whenever one part has a maximal value the other part is zero.

It is apparent that the only way to determine the phase of an electron field is to disentangle the contributions of the real and the imaginary parts of the amplitude. That turns out to be impossible, even in principle. The sum of the squares of the real and the imaginary parts can be known, but there is no way of telling at any given point or at any moment how much of the total derives from the real part and how much from the imaginary part. Indeed, an exact symmetry of the theory implies that the two contributions are indistinguishable. Differences in the phase of the field at two points or at two moments can be measured, but not the absolute phase.

The finding that the phase of an electron wave is inaccessible to measurement has a corollary: the phase cannot have an influence on the outcome of any possible experiment. If it did, that experiment could be used to determine the phase. Hence the electron field exhibits a symmetry with respect to arbitrary changes of phase. Any phase angle can be added to or subtracted from the electron field and the results of all experiments will remain invariant.

This principle can be made clearer by considering an example: the two-slit diffraction experiment with electrons, which is the best-known demonstration of the wavelike nature of matter. In the experiment a beam of electrons passes through two narrow slits in a screen and the number of electrons reaching a second screen is counted. The distribution of electrons across the surface of the second screen forms a diffraction pattern of alternating peaks and valleys.

The quantum-mechanical interpretation of this experiment is that the electron wave splits into two segments on striking the first screen and the two diffracted waves then interfere with each other. Where the waves are in phase the interference is constructive and many electrons are counted at the second screen; where the waves are out of phase destructive interference reduces the count. Clearly it is only the difference in phase that determines the pattern formed. If the phases of both waves were shifted by the same amount, the phase difference at each point would be unaffected and the same pattern of constructive and destructive interference would be observed.

It is symmetries of this kind, where the phase of a quantum field can be adjusted at will, that are called gauge symmetries. Although the absolute value of the phase is irrelevant to the outcome of experiments, in constructing a theory of electrons it is still necessary to specify the phase. The choice of a particular value is called a gauge convention.

Gauge symmetry is not a very descriptive term for such an invariance, but the term has a long history and cannot now be dislodged. It was introduced in about 1920 by Hermann Weyl, who was then attempting to formulate a theory that would combine electromagnetism and the general theory of relativity. Weyl was led to propose a theory that remained invariant with respect to arbitrary dilatations or contractions of space. In the theory a separate standard of length and time had to be adopted at every point in space-time. He compared the choice of a scale convention to a choice of gage blocks, the polished steel blocks employed by machinists as a standard of length. The theory was nearly correct, the necessary emendation being to replace "length scales" by "phase angles." Writing in German, Weyl had referred to "Eich Invarianz," which was initially translated as "calibration invariance," but the alternative translation "gauge" has since become standard.

The symmetry of the electron matter field described above is a global symmetry: the phase of the field must be shifted in the same way everywhere at once. It can easily be demonstrated that a theory of electron fields alone, with no other forms of matter or radiation, is not invariant with respect to a corresponding local gauge transformation. Consider again the two-slit diffraction experiment with electrons. An initial experiment is carried out as before and the electron-diffraction pattern is recorded. Then the experiment is repeated, but one slit is fitted with the electron-optical equivalent of a half-wave plate, a device that shifts the phase of a wave by 180 degrees. When the waves emanating from the two slits now interfere, the phase difference between them will be altered by 180 degrees. As a result wherever the interference was constructive in the first experiment it will now be destructive, and vice versa. The observed diffraction pattern will not be unchanged; on the contrary, the positions of all the peaks and depressions will be interchanged.

Suppose one wanted to make the theory consistent with a local gauge symmetry. Perhaps it could be fixed in some way; in particular, perhaps another field could be added that would compensate for the changes in electron phase. The new field would of course have to do more than mend the defects in this one experiment. It would have to preserve the invariance of all observable quantities when the phase of the electron field was altered in any way from place to place and from moment to moment. Mathematically the phase shift must be allowed to vary as an arbitrary function of position and time.

Although it may seem improbable, a field can be constructed that meets these specifications. It turns out that the required field is a vector one, corresponding to a field quantum with a spin of one unit. Moreover, the field must have infinite range, since there is no limit to the distance over which the phases of the electron fields might have to be reconciled. The need for infinite range implies that the field quantum must be massless. These are the properties of a field that is already familiar: the electromagnetic field, whose quantum is the photon.

How does the electromagnetic field ensure the gauge invariance of the electron field? It should be remembered that the effect of the electromagnetic field is to transmit forces between charged particles. These forces can alter the state of motion of the particles; what is most important in this context, they can alter the phase. When an electron absorbs or emits a photon, the phase of the electron field is shifted. It was shown above that the electromagnetic field itself exhibits an exact local symmetry; by describing the two fields together the local symmetry can be extended to both of them.

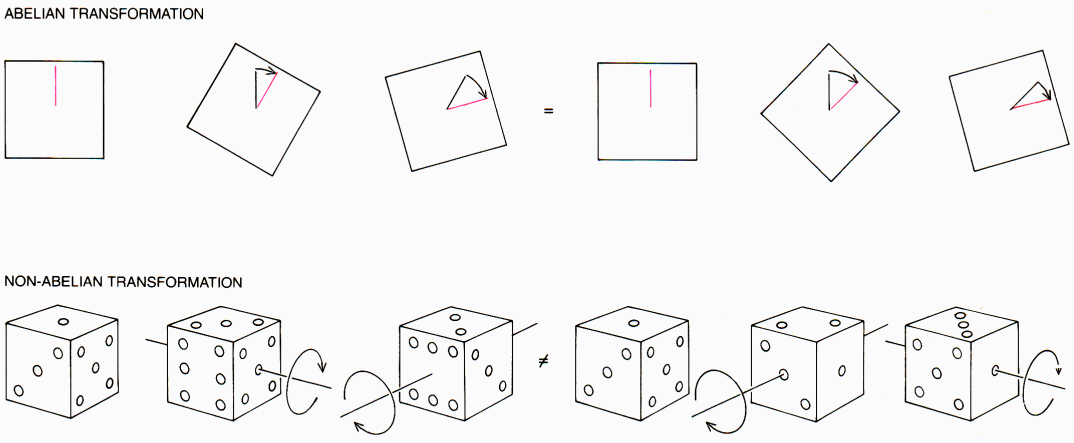
The connection between the two fields lies in the interaction of the electron's charge with the electromagnetic field. Because of this interaction the propagation of an electron matter wave in an electric field can be described properly only if the electric potential is specified. Similarly, to describe an electron in a magnetic field the magnetic vector potential must be specified. Once these two potentials are assigned definite values the phase of the electron wave is fixed everywhere. The local symmetry of electromagnetism, however, allows the electric potential to be given any arbitrary value, which can be chosen independently at every point and at every moment. For this reason the phase of the electron matter field can also take on any value at any point, but the phase will always be consistent with the convention adopted for the electric and the magnetic potentials.

What this means in the two-slit diffraction experiment is that the effects of an arbitrary shift in the phase of the electron wave can be mimicked by applying an electromagnetic field. For example, the change in the observed interference pattern caused by interposing a half-wave plate in front of one slit could be caused instead by placing the slits between the poles of a magnet. From the resulting pattern it would be impossible to tell which procedure had been followed. Since the gauge conventions for the electric and the magnetic potentials can be chosen locally, so can the phase of the electron field.

The theory that results from combining electron matter fields with electromagnetic fields is called quantum electrodynamics. Formulating the theory and proving its consistency was a labor of some 20 years, begun in the 1920's by P. A. M. Dirac and essentially completed in about 1948 by Richard P. Feynman, Julian Schwinger, Sin-itiro Tomonaga and others.

The symmetry properties of quantum electrodynamics are unquestionably appealing, but the theory can be invested with physical significance only if it agrees with the results of experiments. Indeed, before sensible experimental predictions can even be made the theory must pass certain tests of internal consistency. For example, quantum-mechanical theories predict the probabilities of events: the probabilities must not be negative, and all the probabilities taken together must add up to 1. In addition energies must be assigned positive values but should not be infinite.

It was not immediately apparent that quantum electrodynamics could qualify as a physically acceptable theory. One problem arose repeatedly in any attempt to calculate the result of even the simplest electromagnetic interactions, such as the interaction between two electrons. The likeliest sequence of events in such an encounter is that one electron emits a single virtual photon and the other electron absorbs it. Many



EFFECTS OF REPEATED TRANSFORMATIONS distinguish quantum electrodynamics, which is an Abelian theory, from the Yang-Mills theory, which is non-Abelian. An Abelian transformation is commutative: if two transformations are applied in succession, the outcome is the same no matter which sequence is chosen. An example is rotation in two dimensions. Non-Abelian transformations are not commutative, so that two transformations will generally yield different results if their sequence is reversed. Rotations in three dimensions exhibit this dependence on sequence. Quantum electrodynamics is Abelian in that successive phase shifts can be applied to an electron field without regard to the sequence. The Yang-Mills theory is non-Abelian because the net effect of two isotopic-spin rotations is generally different if the sequence of rotations is reversed. One sequence might yield a proton and the opposite sequence a neutron.

more complicated exchanges are also possible, however; indeed, their number is infinite. For example, the electrons could interact by exchanging two photons, or three, and so on. The total probability of the interaction is determined by the sum of the contributions of all the possible events.

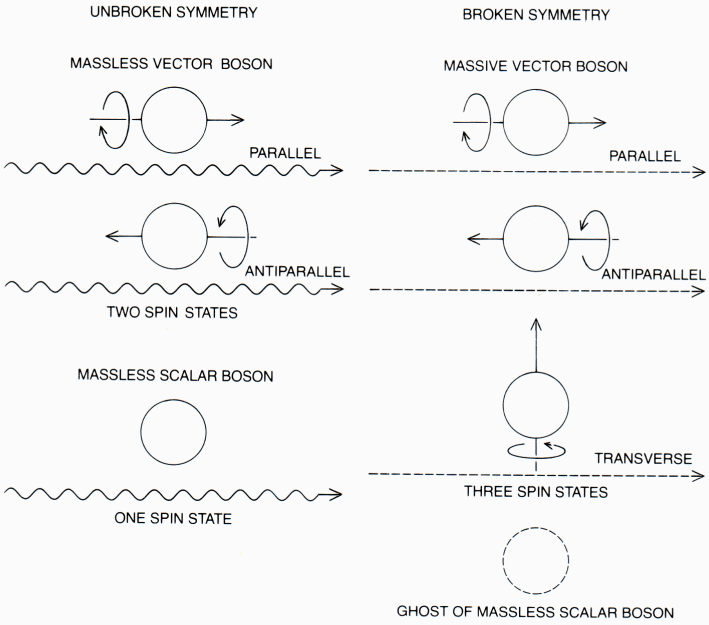
Feynman introduced a systematic procedure for tabulating these contributions by drawing diagrams of the events in one spatial dimension and one time dimension. A notably troublesome class of diagrams are those that include "loops," such as the loop in space-time that is formed when a virtual photon is emitted and later reabsorbed by the same electron. As was shown above, the maximum energy of a virtual particle is limited only by the time needed for it to reach its destination. When a virtual photon is emitted and reabsorbed by the same particle, the distance covered and the time required can be reduced to zero, and so the maximum energy can be infinite. For this reason some diagrams with loops make an infinite contribution to the strength of the interaction.

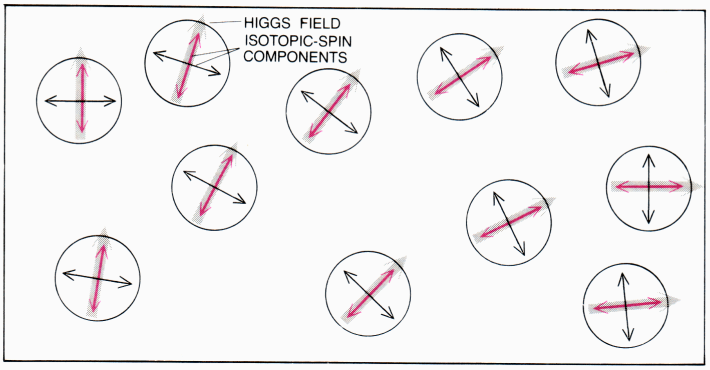
The infinities encountered in quantum electrodynamics led initially to predictions that have no reasonable interpretation as physical quantities. Every interaction of electrons and photons was assigned an infinite probability. The infinities spoiled even the description of an isolated electron: because the electron can emit and reabsorb virtual particles it has infinite mass and infinite charge.

The cure for this plague of infinities is the procedure called renormalization. Roughly speaking, it works by finding one negative infinity for each positive infinity, so that in the sum of all the possible contributions the infinities cancel. The achievement of Schwinger and of the other physicists who worked on the problem was to show that a finite residue could be obtained by this method. The finite residue is the theory's prediction. It is uniquely determined by the requirement that all interaction probabilities come out finite and positive.

The rationale of this procedure can be explained as follows. When a measurement is made on an electron, what is actually measured is not the mass or the charge of the pointlike particle with which the theory begins but the properties of the electron together with its enveloping cloud of virtual particles. Only the net mass and charge, the measurable quantities, are required to be finite at all stages of the calculation. The properties of the pointlike object, which are called the "bare" mass and the "bare" charge, are not well defined.

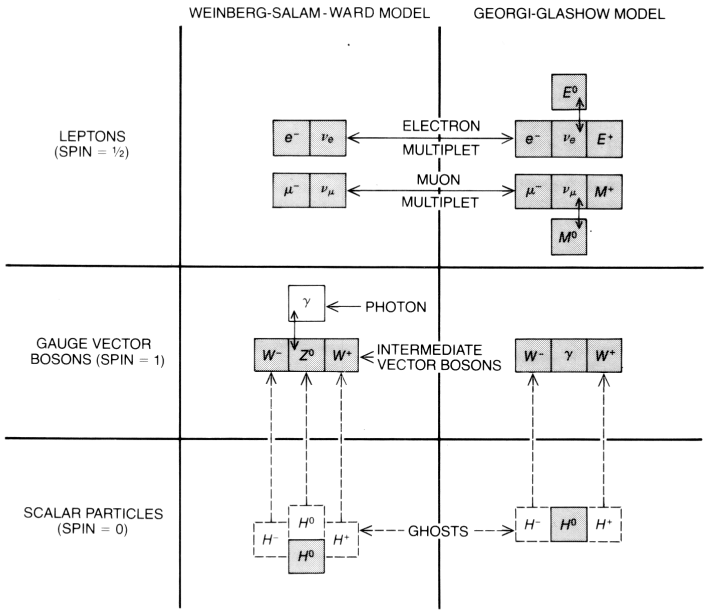
Initially it appeared that the bare mass would have to be assigned a value of negative infinity, an absurdity that made many physicists suspicious of the renormalized theory. A more careful analysis, however, has shown that if the bare mass is to have any definite value, it tends to zero. In any case all quantities with implausible values are unobservable, even in principle. Another objection to the theory is more profound: mathematically quantum electrodynamics is not perfect. Because of the methods that must be used for making predictions in the theory the predictions are limited to a finite accuracy of some hundreds of decimal places.





HIGGS MECHANISM can lend mass to the photonlike vector bosons of the Yang-Mills theory, thereby making the theory more realistic. The massless bosons have three possible spin orientations (parallel, antiparallel and transverse to the direction of motion), but only two of these are observable; the transverse state does not exist, a peculiarity of all massless particles, which move with the speed of light. If the Yang-Mills particles were to acquire a mass, the transverse state would become observable, and this added mode of motion must have some source. In the Higgs mechanism the source is an extra scalar field, corresponding to a massless spin-zero boson. The Yang-Mills particle is said to "eat" the Higgs boson, which thereupon becomes an unobservable "ghost." The Higgs field also provides a frame of reference (*gray arrows*) in which protons can be distinguished from neutrons. The arrow of the Higgs field rotates along with the other arrows in a gauge transformation, and so there is no absolute orientation, but the relative orientation of the isotopic-spin arrows can be measured with respect to the Higgs arrow. The symmetry of the theory, which without the Higgs mechanism would have abolished all differences between the proton and the neutron, has not been lost but only hidden.

Clearly the logic and the internal consistency of the renormalization method leave something to be desired. Perhaps the best defense of the theory is simply that it works very well. It has yielded results that are in agreement with experiments to an accuracy of about one part in a billion, which makes quantum electrodynamics the most accurate physical theory ever devised. It is the model for theories of the other fundamental forces and the standard by which such theories are judged.



WEINBERG-SALAM-WARD MODEL incorporates electromagnetism and the weak force in a local gauge theory. The model applies to the interactions of the particles called leptons, which include the electron (e^-), the muon (μ^-) and two kinds of neutrino (ν_e and ν_μ). A requirement that the interactions of these particles remain invariant with respect to local transformations of a leptonic equivalent of isotopic spin gives rise to four massless fields. Three of these fields are then given a mass through the Higgs mechanism; they become the intermediate vector bosons W^+, W^- and Z^0. The fourth vector boson is the photon. Three of the Higgs particles are eaten by the vector bosons and become ghosts, but a fourth is left over and should be observable. The theory does not truly unify the electromagnetic forces and the weak forces because the photon is still in a family of its own. A theory proposed by Howard Georgi and Sheldon Lee Glashow suggested a more profound unification, where the photon and the massive vector bosons were in the same family, but that theory is now contradicted by experiment.

At the time quantum electrodynamics was completed another theory based on a local gauge symmetry had already been known for some 30 years. It is Einstein's general theory of relativity. The symmetry in question pertains not to a field distributed through space and time but to the structure of space-time itself.

Every point in space-time can be labeled by four numbers, which give its position in the three spatial dimensions and its sequence in the one time dimension. These numbers are the coordinates of the event, and the procedure for assigning such numbers to each point in space-time is a coordinate system. On the earth, for example, the three spatial coordinates are commonly given as longitude, latitude and altitude; the time coordinate can be given in hours past noon. The origin in this coordinate system, the point where all four coordinates have values of zero, lies at noon at sea level where the prime meridian crosses the Equator.

The choice of such a coordinate system is clearly a matter of convention. Ships at sea could navigate just as successfully if the origin of the coordinate system were shifted to Utrecht in the Netherlands. Every point on the earth and every event in its history would have to be assigned new coordinates, but calculations made with those coordinates would invariably give the same results as calculations made in the old system. In particular any calculation of the distance between two points would give the same answer.

The freedom to move the origin of a coordinate system constitutes a symmetry of nature. Actually there are three related symmetries: all the laws of nature remain invariant when the coordinate system is transformed by translation, by rotation or by mirror reflection. It is vital to note, however, that the symmetries are only global ones. Each symmetry transformation can be defined as a formula for finding the new coordinates of a point from the old coordinates. Those formulas must be applied simultaneously in the same way to all the points.

The general theory of relativity stems from the fundamental observation that the structure of space-time is not necessarily consistent with a coordinate system made up entirely of straight lines meeting at right angles; instead a curvilinear coordinate system may be needed. The lines of longitude and latitude employed on the earth constitute such a system, since they follow the curvature of the earth.

In such a system a local coordinate transformation can readily be imagined. Suppose height is defined as vertical distance from the ground rather than from mean sea level. The digging of a pit would then alter the coordinate system, but only at those points directly over the pit. The digging itself represents the local coordinate transformation. It would appear that the laws of physics (or the rules of navigation) do not remain invariant after such a transformation, and in a universe without gravitational forces that would be the case. An airplane set to fly at a constant height would dip suddenly when it flew over the excavation, and the accelerations needed to follow the new profile of the terrain could readily be detected.

As in electrodynamics, local symmetry can be restored only by adding a new field to the theory; in general relativity the field is of course that of gravitation. The presence of this field offers an alternative explanation of the accelerations detected in the airplane: they could result not from a local change in the coordinate grid but from an anomaly in the gravitational field. The source of the anomaly is of no concern: it could be a concentration of mass in the earth or a distant object in space. The point is that any local transformation of the coordinate system could be reproduced by an appropriate set of gravitational fields. The pilot of the airplane could not distinguish one effect from the other.

Both Maxwell's theory of electromagnetism and Einstein's theory of gravitation owe much of their beauty to a local gauge symmetry; their success has long been an inspiration to theoretical physicists. Until recently theoretical accounts of the other two forces in nature have been less satisfactory. A theory of the weak force formulated in the 1930's by Enrico Fermi accounted for some basic features of the weak interaction, but the theory lacked local symmetry. The strong interactions seemed to be a jungle of mysterious fields and resonating particles. It is now clear why it took so long to make sense of these forces: the necessary local gauge theories were not understood.

The first step was taken in 1954 in a theory devised by C. N. Yang and Robert L. Mills, who were then at the Brookhaven National Laboratory. A similar idea was proposed independently at about the same time by R. Shaw of the University of Cambridge. Inspired by the success of the other gauge theories, these theories begin with an established global symmetry and ask what the consequences would be if it were made a local symmetry.

The symmetry at issue in the Yang-Mills theory is isotopic-spin symmetry, the rule stating that the strong interactions of matter remain invariant (or nearly so) when the identities of protons and neutrons are interchanged. In the global symmetry any rotation of the internal arrows that indicate the isotopic-spin state must be made simultaneously everywhere. Postulating a local symmetry allows the orientation of the arrows to vary independently from place to place and from moment to moment. Rotations of the arrows can depend on any arbitrary function of position and time. This freedom to choose different conventions for the identity of a nuclear particle in different places constitutes a local gauge symmetry.

As in other instances where a global symmetry is converted into a local one, the invariance can be maintained only if something more is added to the theory. Because the Yang-Mills theory is more complicated than earlier gauge theories it turns out that quite a lot more must be added. When isotopic-spin rotations are made arbitrarily from place to place, the laws of physics remain invariant only if six new fields are introduced. They are all vector fields, and they all have infinite range.

The Yang-Mills fields are constructed on the model of electromagnetism, and indeed two of them can be identified with the ordinary electric and magnetic fields. In other words, they describe the field of the photon. The remaining Yang-Mills fields can also be taken in pairs and interpreted as electric and magnetic fields, but the photons they describe differ in a crucial respect from the known properties of the photon: they are still massless spin-one particles, but they carry an electric charge. One photon is negative and one is positive.

The imposition of an electric charge on a photon has remarkable consequences. The photon is defined as the field quantum that conveys electromagnetic forces from one charged particle to another. If the photon itself has a charge, there can be direct electromagnetic interactions among the photons. To cite just one example, two photons with opposite charges might bind together to form an "atom" of light. The familiar neutral photon never interacts with itself in this way.

The surprising effects of charged photons become most apparent when a local symmetry transformation is applied more than once to the same particle. In quantum electrodynamics, as was pointed out above, the symmetry operation is a local change in the phase of the electron field, each such phase shift being accompanied by an interaction with the electromagnetic field. It is easy to imagine an electron undergoing two phase shifts in succession, say by emitting a photon and later absorbing one. Intuition suggests that if the sequence of the phase shifts were reversed, so that first a photon was absorbed and later one was emitted, the end result would be the same. This is indeed the case. An unlimited series of phase shifts can be made, and the final result will be simply the algebraic sum of all the shifts no matter what their sequence.

In the Yang-Mills theory, where the symmetry operation is a local rotation of the isotopic-spin arrow, the result of multiple transformations can be quite different. Suppose a hadron is subjected to a gauge transformation, A, followed soon after by a second transformation, B; at the end of this sequence the isotopic-spin arrow is found in the orientation that corresponds to a proton. Now suppose the same transformations were applied to the same hadron but in the reverse sequence: B followed by A. In general the final state will not be the same; the particle may be a neutron instead of a proton. The net effect of the two transformations depends explicitly on the sequence in which they are applied.

Because of this distinction quantum electrodynamics is called an Abelian theory and the Yang-Mills theory is called a non-Abelian one. The terms are borrowed from the mathematical theory of groups and honor Niels Henrik Abel, a Norwegian mathematician who lived in the early years of the 19th century. Abelian groups are made up of transformations that, when they are applied one after another, have the commutative property; non-Abelian groups are not commutative.

Commutation is familiar in arithmetic as a property of addition and multiplication, where for any numbers A and B it can be stated that $A + B = B + A$ and $A \times B = B \times A$. How the principle can be applied to a group of transformations can be illustrated with a familiar example: the group of rotations. All possible rotations of a two-dimensional object are commutative, and so the group of such rotations is Abelian. For instance, rotations of $+60$ degrees and -90 degrees yield a net rotation of -30 degrees no matter which is applied first. For a three-dimensional object free to rotate about three axes the commutative law does not hold, and the group of three-dimensional rotations is non-Abelian. As an example, consider an airplane heading due north in level flight. A 90-degree yaw to the left followed by a 90-degree roll to the left leaves the airplane heading west with its left wing tip pointing straight down. Reversing the sequence of transformations, so that a 90-degree roll to the left is followed by a 90-degree left yaw, puts the airplane in a nose dive with the wings aligned on the north-south axis.

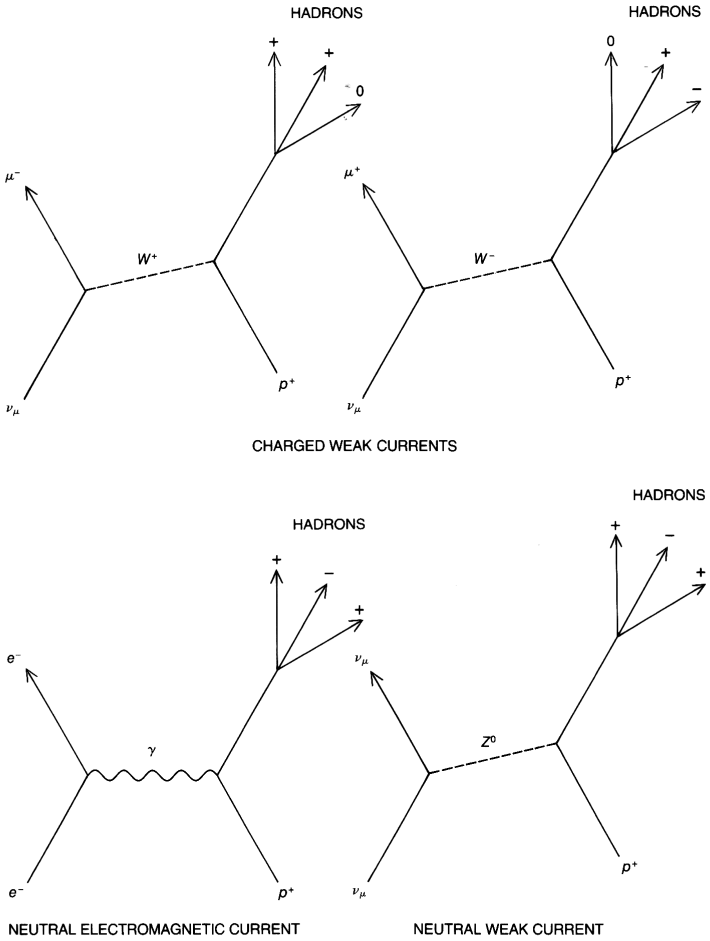
Like the Yang-Mills theory, the general theory of relativity is non-Abelian: in making two successive coordinate transformations, the order in which they are made usually has an effect on the outcome. In the past 10 years or so several more non-Abelian theories have been devised, and even the electromagnetic interactions have been incorporated into a larger theory that is non-Abelian. For now, at least, it seems all the forces of nature are governed by non-Abelian gauge theories.

The Yang-Mills theory has proved to be of monumental importance, but as it was originally formulated it was totally unfit to describe the real world. A first objection to it is that isotopic-spin symmetry becomes exact, with the result that protons and neutrons are indistinguishable; this situation is obviously contrary to fact. Even more troubling is the prediction of electrically charged photons. The photon is necessarily massless because it must have an infinite range. The existence of any electrically charged particle lighter than the electron would alter the world beyond recognition. Of course, no such particle has been observed. In spite of these difficulties the theory has great beauty and philosophical appeal. One strategy adopted in an attempt to fix its defects was to artificially endow the charged field quanta with a mass greater than zero.

Imposing a mass on the quanta of the charged fields does not make the fields disappear, but it does confine them to a finite range. If the mass is large enough, the range can be made as small as is wished. As the long-range effects are removed the existence of the fields can be reconciled with experimental observations. Moreover, the selection of the neutral Yang-Mills field as the only real long-range one automatically distinguishes protons from neutrons. Since this field is simply the electromagnetic field, the proton and the neutron can be distinguished by their differing interactions with it, or in other words by their differing electric charges.

With this modification the local symmetry of the Yang-Mills theory would no longer be exact but approximate, since rotation of the isotopic-spin arrow would now have observable consequences. That is not a fundamental objection: approximate symmetries are quite commonplace in nature. (The bilateral symmetry of the human body is only approximate.) Moreover, at distance scales much smaller than the range of the massive components of the Yang-Mills field, the local symmetry becomes better and better. Thus in a sense the microscopic structure of the theory could remain locally symmetric, but not its predictions of macroscopic, observable events.

The modified Yang-Mills theory was easier to understand, but the theory still had to be given a quantum-mechanical



NEUTRAL WEAK CURRENTS provide the decisive test of the Weinberg-Salam-Ward model. It once appeared that all weak interactions involved a transfer of electric charge between the interacting particles; these events could be explained by just two intermediate vector bosons, the W^+ and the W^-. Events in which no charge was transferred were characteristic of electromagnetic interactions, where the exchanged virtual particle is a photon. The Weinberg-Salam-Ward model predicts that weak interactions can also proceed without charge transfer; these neutral weak currents are mediated by the neutral boson Z^0, which is identical with the photon except that it has a very large mass. Neutral weak currents were first observed in 1973.

interpretation. The problem of infinities turned out to be severer than it had been in quantum electrodynamics, and the standard recipe for renormalization would not solve it. New techniques had to be devised.

An important idea was introduced in 1963 by Feynman: it is the notion of a "ghost" particle, a particle added to a theory in the course of a calculation that vanishes when the calculation is finished. It is known from the outset that the ghost particle is fictitious, but its use can be justified if it never appears in the final state. This can be ensured by making certain the total probability of producing a ghost particle is always zero.

Among theoretical groups that continued work on the Yang-Mills theory the ghost-particle method was taken seriously only at the University of Utrecht, where I was then a student. Martin J. G. Veltman, my thesis adviser, together with John S. Bell of the European Organization for Nuclear Research (CERN) in Geneva, was led to the conclusion that the weak interactions might be described by some form of the Yang-Mills theory. He undertook a systematic analysis of the renormalization problem in the modified Yang-Mills model (with massive charged fields), examining each class of Feynman diagrams in turn. The diagrams having no closed loops were readily shown to make only finite contributions to the total interaction probability. The diagrams with one loop do include infinite terms, but by exploiting the properties of the ghost particles it was possible to make the positive infinities and the negative ones cancel exactly.

As the number of loops increases, the number of diagrams rises steeply; moreover, the calculations required for each diagram become more intricate. To assist in the enormous task of checking all the two-loop diagrams a computer program was written to handle the algebraic manipulation of the probabilities. The output of the program is a list of the coefficients of the infinite quantities remaining after the contributions of all the diagrams have been summed. If the infinities are to be expunged from the theory, the coefficients must without exception be zero. By 1970 the results were known and the possibility of error had been excluded; some infinities remained.

The failure of the modified Yang-Mills theory was to be blamed not on any defect in the Yang-Mills formulation itself but rather on the modifications. The masses of the charged fields had to be put in "by hand" and as a result the invariance with respect to local isotopic-spin rotations was not quite perfect. It was suggested at the time by the Russian investigators L. D. Faddeev, V. N. Popov, E. S. Fradkin and I. V. Tyutin that the pure Yang-Mills theory, with only massless fields, could indeed be renormalized. The trouble with this theory is that it not only is unrealistic but also has long-range fields that are difficult to work with.

In the meantime another new ingredient for the formulation of gauge theories had been introduced by F. Englert and Robert H. Brout of the University of Brussels and by Peter Higgs of the University of Edinburgh. They found a way to endow some of the Yang-Mills fields with mass while retaining exact gauge symmetry. The technique is now called the Higgs mechanism.

The fundamental idea of the Higgs mechanism is to include in the theory an extra field, one having the peculiar property that it does not vanish in the vacuum. One usually thinks of a vacuum as a space with nothing in it, but in physics the vacuum is defined more precisely as the state in which all fields have their lowest possible energy. For most fields the energy is minimized when the value of the field is zero everywhere, or in other words when the field is "turned off." An electron field, for example, has its minimum energy when there are no electrons. The Higgs field is unusual in this respect. Reducing it to zero costs energy; the energy of the field is smallest when the field has some uniform value greater than zero.

The effect of the Higgs field is to provide a frame of reference in which the

orientation of the isotopic-spin arrow can be determined. The Higgs field can be represented as an arrow superposed on the other isotopic-spin indicators in the imaginary internal space of a hadron. What distinguishes the arrow of the Higgs field is that it has a fixed length, established by the vacuum value of the field. The orientation of the other isotopic-spin arrows can then be measured with respect to the axis defined by the Higgs field. In this way a proton can be distinguished from a neutron.

It might seem that the introduction of the Higgs field would spoil the gauge symmetry of the theory and thereby lead again to insoluble infinities. In actuality, however, the gauge symmetry is not destroyed but merely concealed. The symmetry specifies that all the laws of physics must remain invariant when the isotopic-spin arrow is rotated in an arbitrary way from place to place. This implies that the absolute orientation of the arrow cannot be determined, since any experiment for measuring the orientation would have to detect some variation in a physical quantity when the arrow was rotated. With the inclusion of the Higgs field the absolute orientation of the arrow still cannot be determined because the arrow representing the Higgs field also rotates during a gauge transformation. All that can be measured is the angle between the arrow of the Higgs field and the other isotopic-spin arrows, or in other words their relative orientations.

The Higgs mechanism is an example of the process called spontaneous symmetry breaking, which was already well established in other areas of physics. The concept was first put forward by Werner Heisenberg in his description of ferromagnetic materials. Heisenberg pointed out that the theory describing a ferromagnet has perfect geometric symmetry in that it gives no special distinction to any one direction in space. When the material becomes magnetized, however, there is one axis—the direction of magnetization—that can be distinguished from all other axes. The theory is symmetrical but the object it describes is not. Similarly, the Yang-Mills theory retains its gauge symmetry with respect to rotations of the isotopic-spin arrow, but the objects described—protons and neutrons—do not express the symmetry.

How does the Higgs mechanism lend mass to the quanta of the Yang-Mills field? The process can be explained as follows. The Higgs field is a scalar quantity, having only a magnitude, and so the quantum of the field must have a spin of zero. The Yang-Mills fields are vectors, like the electromagnetic field, and are represented by spin-one quanta. Ordinarily a particle with a spin of one unit has three spin states (oriented parallel, antiparallel and transverse to its direction of motion), but because the Yang-Mills particles are massless and move with the speed of light they are a special case; their transverse states are missing. If the particles were to acquire a mass, they would lose this special status and all three spin states would have to be observable. In quantum mechanics the accounting of spin states is strict and the extra state must come from somewhere; it comes from the Higgs field. Each Yang-Mills quantum coalesces with one Higgs particle; as a result the Yang-Mills particle gains mass and a spin state, whereas the Higgs particle disappears. A picturesque description of this process has been suggested by Abdus Salam of the International Center for Theoretical Physics in Trieste: the massless Yang-Mills particles "eat" the Higgs particles in order to gain weight, and the swallowed Higgs particles become ghosts.

In 1971, Veltman suggested that I investigate the renormalization of the pure Yang-Mills theory. The rules for constructing the needed Feynman diagrams had already been formulated by Faddeev, Popov, Fradkin and Tyutin, and independently by Bryce S. DeWitt of the University of Texas at Austin and Stanley Mandelstam of the University of California at Berkeley. I could adapt to the task the powerful methods for renormalization studies that had been developed by Veltman.

Formally the results were encouraging, but if the theory was to be a realistic one, some means had to be found to confine the Yang-Mills fields to a finite range. I had just learned at a summer school how Kurt Symanzik of the German Electron Synchrotron and Benjamin W. Lee of the Fermi National Accelerator Laboratory had successfully handled the renormalization of a theoretical model in which a global symmetry is spontaneously broken. It therefore seemed natural to try the Higgs mechanism in the Yang-Mills theory, where the broken symmetry is a local one.

A few simple models gave encouraging results: in these selected instances all infinities canceled no matter how many gauge particles were exchanged and no matter how many loops were included in the Feynman diagrams. The decisive test would come when the theory was checked by the computer program for infinities in all possible diagrams with two loops. The results of that test were available by July, 1971; the output of the program was an uninterrupted string of zeros. Every infinity canceled exactly. Subsequent checks showed that infinities were also absent even in extremely complicated Feynman diagrams. My results were soon confirmed by others, notably by Lee and by Jean Zinn-Justin of the Saclay Nuclear Research Center near Paris.

The Yang-Mills theory had begun as a model of the strong interactions, but by the time it had been renormalized interest in it centered on applications to the weak interactions. In 1967 Steven Weinberg of Harvard University and independently (but later) Salam and John C. Ward of Johns Hopkins University had proposed a model of the weak interactions based on a version of the Yang-Mills theory in which the gauge quanta take on mass through the Higgs mechanism. They speculated that it might be possible to renormalize the theory, but they did not demonstrate it. Their ideas therefore joined many other untested conjectures until some four years later, when my own results showed it was just that subclass of Yang-Mills theories incorporating the Higgs mechanism that can be renormalized.

The most conspicuous trait of the weak force is its short range: it has a significant influence only to a distance of 10^{-15} centimeter, or roughly a hundredth the radius of a proton. The force is weak largely because its range is so short: particles are unlikely to approach each other closely enough to interact. The short range implies that the virtual particles exchanged in weak interactions must be very massive. Present estimates run to between 80 and 100 times the mass of the proton.

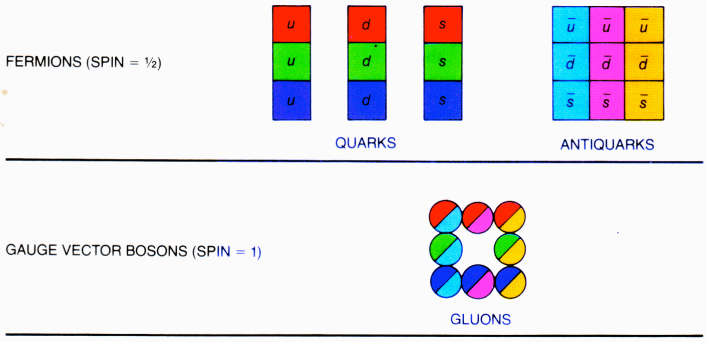
The Weinberg-Salam-Ward model actually embraces both the weak force and electromagnetism. The conjecture on which the model is ultimately founded is a postulate of local invariance with respect to isotopic spin; in order to preserve that invariance four photonlike fields are introduced, rather than the three of the original Yang-Mills theory. The fourth photon could be identified with some primordial form of electromagnetism. It corresponds to a separate force, which had to be added to the theory without explanation. For this reason the model should not be called a unified field theory. The forces remain distinct; it is their intertwining that makes the model so peculiar.

At the outset all four of the fields in the Weinberg-Salam-Ward model are of infinite range and therefore must be conveyed by massless quanta; one field carries a negative electric charge, one carries a positive charge and the other two fields are neutral. The spontaneous symmetry breaking introduces four Higgs fields, each field represented by a scalar particle. Three of the Higgs fields are swallowed by Yang-Mills particles, so that both of the charged Yang-Mills particles and one of the neutral ones take on a large mass. These particles are collectively named massive intermediate vector bosons, and they are designated W^+, W^- and Z^0. The fourth Yang-Mills particle, which is a neutral one, remains massless: it is the photon of electromagnetism. Of the Higgs parti-

cles, the three that lend mass to the Yang-Mills particles become ghosts and are therefore unobservable, but the last Higgs particle is not absorbed, and it should be seen if enough energy is available to produce it.

The most intriguing prediction of the model was the existence of the Z^0, a particle identical with the photon in all respects except mass, which had not been included in any of the earlier, provisional accounts of the weak force. Without the Z^0 any weak interaction would necessarily entail an exchange of electric charge. Events of this kind are called charged-weak-current events. The Z^0 introduced a new kind of weak interaction, a neutral-weak-current event. By exchanging a Z^0, particles would interact without any transfer of charge and could retain their original identities. Neutral weak currents were first observed in 1973 at CERN.

The elaboration of a successful gauge theory of the strong interactions, which are unique to hadrons, could not be undertaken until a fundamental fact about the hadrons was understood: they are not elementary particles. A model of hadrons as composite objects was proposed in 1963 by Murray Gell-Mann of the California Institute of Technology; a similar idea was introduced independently and at about the same time by Yuval Ne'eman of Tel Aviv University and George Zweig of Cal Tech. In this model hadrons are made up of the smaller particles Gell-Mann named quarks. A hadron can be built out of quarks according to either of two blueprints. Combining three quarks gives rise to a baryon, a class of hadrons that includes the proton and the neutron. Binding together one quark and one antiquark makes a meson, a class typified by the pions. Every known hadron can be accounted for as one of these allowed combinations of quarks.



QUARK MODEL describes all hadrons, including the proton and the neutron, as being composite particles made up of the smaller entities called quarks. In the original form of the model the quarks were assumed to come in three "flavors," labeled *u*, *d* and *s*, each of which is now said to have three possible "colors," red, green and blue. There are also antiquarks with the corresponding anticolors cyan, magenta and yellow. The interactions of the quarks are now described by means of a gauge theory based on invariance with respect to local transformations of color. Sixteen fields are needed to hold this invariance. They are taken in pairs to make up eight massless vector bosons, called gluons, each bearing a combination of color and anticolor.

In the original model there were just three kinds of quark, designated "up," "down" and "strange." James D. Bjorken of the Stanford Linear Accelerator Center and Sheldon Lee Glashow of Harvard soon proposed adding a fourth quark bearing a property called charm. In 1971 a beautiful argument by Glashow, John Iliopoulos of Paris and Luciano Maiani of the University of Rome showed that a quark with charm is needed to cure a discrepancy in the gauge theory of weak interactions. Charmed quarks, it was concluded, must exist if both the gauge theory and the quark theory are correct. The discovery in 1974 of the *J* or psi particle, which consists of a charmed quark and a charmed antiquark, supported the Weinberg-Salam-Ward model and persuaded many physicists that the quark model as a whole should be taken seriously. It now appears that at least two more "flavors," or kinds, of quark are needed; they have been labeled "top" and "bottom."

The primary task of any theory of the strong interactions is to explain the peculiar rules for building hadrons out of quarks. The structure of a meson is not too difficult to account for: since the meson consists of a quark and an antiquark, it is merely necessary to assume that the quarks carry some property analogous to electric charge. The binding of a quark and an antiquark would then be explained on the principle that opposite charges attract, just as they do in the hydrogen atom. The structure of the baryons, however, is a deeper enigma. To explain how three quarks can form a bound state one must assume that three like charges attract.

The theory that has evolved to explain the strong force prescribes exactly these interactions. The analogue of electric charge is a property called color (although it can have nothing to do with the colors of the visible spectrum). The term color was chosen because the rules for forming hadrons can be expressed succinctly by requiring all allowed combinations of quarks to be "white," or colorless. The quarks are assigned the primary colors red, green and blue; the antiquarks have the complementary "anticolors" cyan, magenta and yellow. Each of the quark flavors comes in all three colors, so that the introduction of the color charge triples the number of distinct quarks.

From the available quark pigments there are two ways to create white: by mixing all three primary colors or by mixing one primary color with its complementary anticolor. The baryons are made according to the first scheme: the three quarks in a baryon are required to have different colors, so that the three primary hues are necessarily represented. In a meson a color is always accompanied by its complementary anticolor.

The theory devised to account for these baffling interactions is modeled directly on quantum electrodynamics and is called quantum chromodynamics. It is a non-Abelian gauge theory. The gauge symmetry is an invariance with respect to local transformations of quark color.

It is easy to imagine a global color symmetry. The quark colors, like the isotopic-spin states of hadrons, might be indicated by the orientation of an arrow in some imaginary internal space. Successive rotations of a third of a turn would change a quark from red to green to blue and back to red again. In a baryon, then, there would be three arrows, with one arrow set to each of the three colors. A global symmetry transformation, by definition, must affect all three arrows in the same way and at the same time. For example, all three arrows might rotate clockwise a third of a turn. As a result of such a transformation all three quarks would change color, but all observable properties of the hadron would remain as before. In particular there would still be one quark of each color, and so the baryon would remain colorless.

Quantum chromodynamics requires that this invariance be retained even when the symmetry transformation is a local one. In the absence of forces or interactions the invariance is obviously lost. Then a local transformation can change the color of one quark but leave the other quarks unaltered, which would give the hadron a net color. As in other gauge theories, the way to restore the

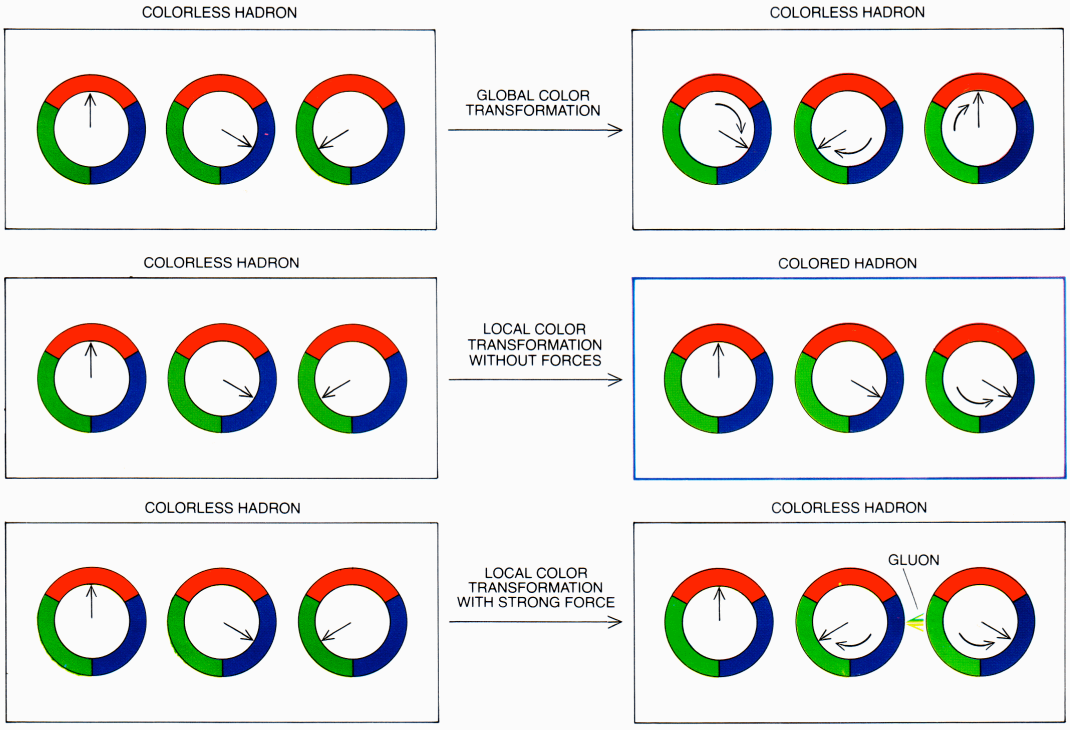
invariance with respect to local symmetry operations is to introduce new fields. In quantum chromodynamics the fields needed are analogous to the electromagnetic field but are much more complicated; they have eight times as many components as the electromagnetic field has. It is these fields that give rise to the strong force.

The quanta of the color fields are called gluons (because they glue the quarks together). There are eight of them, and they are all massless and have a spin angular momentum of one unit. In other words, they are massless vector bosons like the photon. Also like the photon the gluons are electrically neutral, but they are not color-neutral. Each gluon carries one color and one anticolor. There are nine possible combinations of a color and an anticolor, but one of them is equivalent to white and is excluded, leaving eight distinct gluon fields.

The gluons preserve local color symmetry in the following way. A quark is free to change its color, and it can do so independently of all other quarks, but every color transformation must be accompanied by the emission of a gluon, just as an electron can shift its phase only by emitting a photon. The gluon, propagating at the speed of light, is then absorbed by another quark, which will have its color shifted in exactly the way needed to compensate for the original change. Suppose, for example, a red quark changes its color to green and in the process emits a gluon that bears the colors red and antigreen. The gluon is then absorbed by a green quark, and in the ensuing reaction the green of the quark and the antigreen of the gluon annihilate each other, leaving the second quark with a net color of red. Hence in the final state as in the initial state there is one red quark and one green quark. Because of the continual arbitration of the gluons there can be no net change in the color of a hadron, even though the quark colors vary freely from point to point. All hadrons remain white, and the strong force is nothing more than the system of interactions needed to maintain that condition.

In spite of the complexity of the gluon fields, quantum electrodynamics and quantum chromodynamics are remarkably similar in form. Most notably the photon and the gluon are identical in their spin and in their lack of mass and electric charge. It is curious, then, that the interactions of quarks are very different from those of electrons.

Both electrons and quarks form bound states, namely atoms for the electrons and hadrons for the quarks. Electrons, however, are also observed as independent particles; a small quantity of energy suffices to isolate an electron by ionizing an atom. An isolated quark has never been detected. It seems to be impossible to ionize a hadron, no matter how much energy is supplied. The quarks are evidently bound so tightly that they cannot be pried apart; paradoxically, however, probes of the in-



COLOR SYMMETRY requires that every hadron remain white, or colorless, even when the colors of its constituent quarks have been altered. The color of a quark can be indicated by the position of an arrow in an imaginary internal space. Global symmetry is easily achieved. If a hadron initially consists of three quarks, one in each of the three colors, then any synchronized rotation of all three of the arrows must leave the overall balance of the colors unchanged. In the absence of forces between the quarks, however, the global symmetry cannot be converted into a local symmetry. Changing the position of one color arrow while leaving the other two arrows fixed gives the hadron a net color. In order to preserve the local color symmetry, forces must be introduced. In particular when the color of one quark is changed, a virtual particle must be emitted that will readjust the colors of the other quarks so that the hadron as a whole will remain colorless. The fields that are required to ensure the colorlessness of all the hadrons are the eight gluon fields of quantum chromodynamics.

ternal structure of hadrons show the quarks moving freely, as if they were not bound at all.

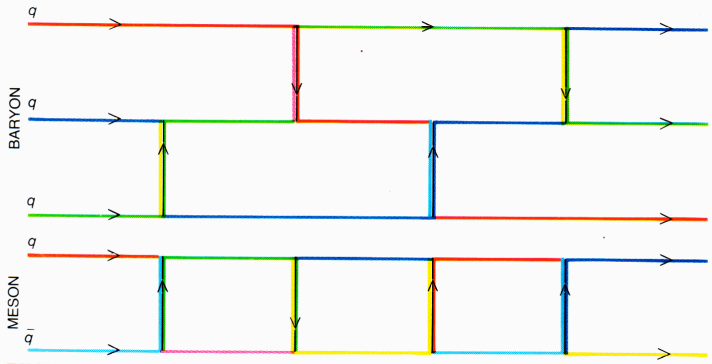
Gluons too have not been seen directly in experiments. Their very presence in the theory provokes objections like those raised against the pure, massless Yang-Mills theory. If massless particles that so closely resemble the photon existed, they would be easy to detect and they would have been known long ago. Of course, it might be possible to give the gluons a mass through the Higgs mechanism. With eight gluons to be concealed in this way, however, the project becomes rather cumbersome. Moreover, the mass would have to be large or the gluons would have been produced by now in experiments with high-energy accelerators; if the mass is large, however, the range of the quark-binding force becomes too small.

A tentative resolution of this quandary has been discovered not by modifying the color fields but by examining their properties in greater detail. In discussing the renormalization of quantum electrodynamics I pointed out that even an isolated electron is surrounded by a cloud of virtual particles, which it constantly emits and reabsorbs. The virtual particles include not only neutral ones, such as the photon, but also pairs of oppositely charged particles, such as electrons and their antiparticles, the positrons. It is the charged virtual particles in this cloud that under ordinary circumstances conceal the "infinite" negative bare charge of the electron. In the vicinity of the bare charge the electron-positron pairs become slightly polarized: the virtual positrons, under the attractive influence of the bare charge, stay closer to it on the average than the virtual electrons, which are repelled. As a result the bare charge is partially neutralized; what is seen at long range is the difference between the bare charge and the screening charge of the virtual positrons. Only when a probe approaches to within less than about 10^{-10} centimeter do the unscreened effects of the bare charge become significant.

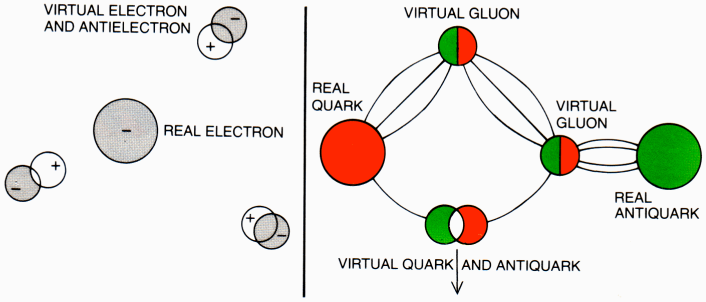
It is reasonable to suppose the same process would operate among color charges, and indeed it does. A red quark is enveloped by pairs of quarks and antiquarks, and the antired charges in this cloud are attracted to the central quark and tend to screen its charge. In quantum chromodynamics, however, there is a competing effect that is not present in quantum electrodynamics. Whereas the photon carries no electric charge and therefore has no direct influence on the screening of electrons, gluons do bear a color charge. (This distinction expresses the fact that quantum electrodynamics is an Abelian theory and quantum chromodynamics is a non-Abelian one.) Virtual gluon pairs also form a cloud around a colored quark, but it turns out that the gluons tend to enhance the color charge rather than attenuate it. It is as if the red component of a gluon were attracted to a red quark and therefore added its charge to the total effective charge. If there are no more than 16 flavors of quark (and at present only six are known), the "antiscreening" by gluons is the dominant influence.

This curious behavior of the gluons follows from rather involved calculations, and the interpretation of the results depends on how the calculation was done. When I calculate it, I find that the force responsible is the color analogue of the gluon's magnetic field. It is also significant, however, that virtual gluons can be emitted singly, whereas virtual quarks always appear as a quark and an antiquark. A single gluon, bearing a net color charge, enhances the force acting between two other color charges.

As a result of this "antiscreening" the effective color charge of a quark grows larger at long range than it is close by. A distant quark reacts to the combined fields of the central quark and the reinforcing gluon charges; at close range, once the gluon cloud has been penetrated, only the smaller bare charge is effective. The quarks in a hadron therefore act somewhat as if they were connected by rubber bands: at very close range, where the bands are slack, the quarks



EXCHANGE OF GLUONS maintains a baryon (made up of three quarks) or a meson (made up of a quark and an antiquark) colorless. In this process the total color of the particles is conserved. For example, a red quark can be converted into a green quark only by emitting a gluon that bears the color red and the anticolor magenta; the magenta can be interpreted as antigreen. Hence the red of the quark is carried off by the red of the gluon, and green and antigreen are created in equal quantities. If the gluon is absorbed by a green quark, the green of the quark and the antigreen of the gluon annihilate each other, leaving the second quark with the color red.



POLARIZATION OF THE VACUUM explains to some extent the peculiar force law that seems to allow quarks complete freedom of movement within a hadron but forbids the isolation of quarks or gluons. In quantum electrodynamics (*left*) pairs of virtual electrons and antielectrons surround any isolated charge, such as an electron. Because of electrostatic forces the positively charged antielectrons tend to remain nearer the negative electron charge and thereby cancel part of it. The observed electron charge is the difference between the "bare" charge and the screening charge of virtual antielectrons. Similarly, pairs of virtual quarks diminish the strength of the force between a real quark and a real antiquark. In quantum chromodynamics, however, there is a competing effect not found in quantum electrodynamics. Because the gluon also has a color charge (whereas the photon has no electric charge), virtual gluons also have an influence on the magnitude of the color force between quarks. The gluons do not screen the quark charge but enhance it. As a result the color charge is weak and the quarks move freely as long as they are close. At long range infinite energy may be needed to separate two quarks.

move almost independently, but at a greater distance, where the bands are stretched taut, the quarks are tightly bound.

The polarization of virtual gluons leads to a reasonably precise account of the close-range behavior of quarks. Where the binding is weak, the expected motion of the particles can be calculated successfully. The long-range interactions, and most notably the failure of quarks and gluons to appear as free particles, can probably be attributed to the same mechanism of gluon antiscreening. It seems likely that as two color charges are pulled apart the force between them grows stronger indefinitely, so that infinite energy would be needed to create a macroscopic separation. This phenomenon of permanent quark confinement may be linked to certain special mathematical properties of the gauge theory. It is encouraging that permanent confinement has indeed been found in some highly simplified models of the theory. In the full-scale theory all methods of calculation fail when the forces become very large, but the principle seems sound. Quarks and gluons may therefore be permanently confined in hadrons.

If the prevailing version of quantum chromodynamics turns out to be correct, color symmetry is an exact symmetry and the colors of particles are completely indistinguishable. The theory is a pure gauge theory of the kind first proposed by Yang and Mills. The gauge fields are inherently long-range and formally are much like the photon field. The quantum-mechanical constraints on those fields are so strong, however, that the observed interactions are quite unlike those of electromagnetism and even lead to the imprisonment of an entire class of particles.

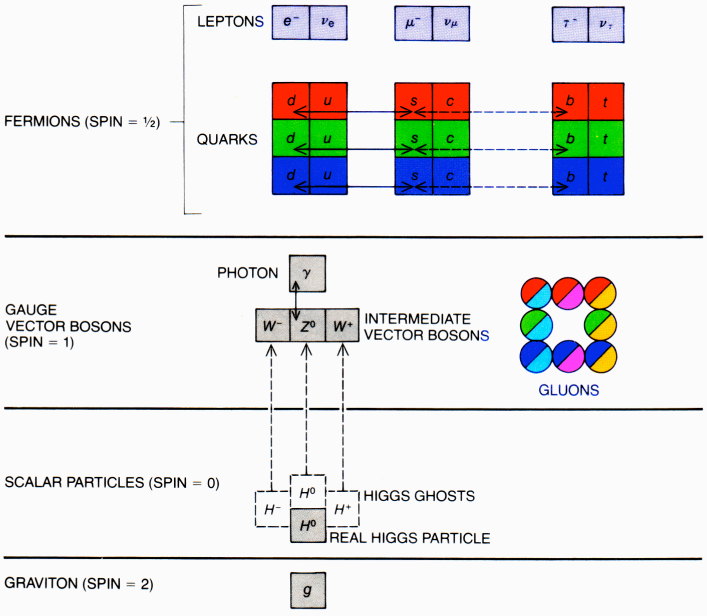
Even where the gauge theories are right they are not always useful. The calculations that must be done to predict the result of an experiment are tedious, and except in quantum electrodynamics high accuracy can rarely be attained. It is mainly for practical or technical reasons such as these that the problem of quark confinement has not been solved. The equations that describe a proton in terms of quarks and gluons are about as complicated as the equations that describe a nucleus of medium size in terms of protons and neutrons. Neither set of equations can be solved rigorously.

In spite of these limitations the gauge theories have made an enormous contribution to the understanding of elementary particles and their interactions. What is most significant is not the philosophical appeal of the principle of local symmetry, or even the success of the individual theories. Rather it is the growing conviction that the class of theories now under consideration includes all possible theories for any system of particles whose mutual interactions are not too strong. Experiment shows that if particles remain closer together than about 10^{-14} centimeter, their total interaction, including the effects of all forces whether known or not, is indeed small. (The quarks are a special case: although the interactions between them are not small, those interactions can be attributed to the effects of virtual particles, and the interactions of the virtual particles are only moderate.) Hence it seems reasonable to attempt a systematic fitting of the existing gauge theories to experimental data.

The mathematics of the gauge theories is rigid, but it does leave some freedom for adjustment. That is, the predicted magnitude of an interaction between particles depends not only on the structure of the theory but also on the values assigned to certain free parameters, which must be considered constants of nature. The theory remains consistent no matter what choice is made for these constants, but the experimental predictions depend strongly on what values are assigned to the constants. Although the constants can be measured by doing experiments, they can never be derived from the theory. Examples of such constants of nature are the charge of the electron and the masses of elementary particles such as the electron and the quarks.

The strength of the gauge theories is that they require comparatively few such free parameters: about 18 constants of nature must be supplied to account for all the known forces. The tangled phenomena of the strongly interacting particles, which seemed incomprehensible 15 years ago, can now be unraveled by means of a theory that includes only a handful of free parameters. Among these all but three are small enough to be safely ignored.

Even if the free parameters have been reduced to a manageable number, they remain an essential part of the theory. No explanation can be offered of why they assume the values they do. The fundamental questions that remain unanswered by the gauge theories center on these apparent constants of nature. Why do the quarks and the other elementary particles have the masses they do? What determines the mass of the Higgs particle? What determines the fundamental unit of electric charge or the strength of the color force? The answers to such questions cannot come from the existing



STANDARD MODEL of elementary-particle interactions describes the four forces of nature by means of three non-Abelian gauge theories. The fundamental particles of matter are six leptons and six flavors of quark, each of the flavors being present in three colors. Electromagnetism and the weak force are mediated by the gauge particles of the Weinberg-Salam-Ward model, namely the massless photon and a triplet of very massive vector bosons, the W^+, W^- and Z^0. The strong force is attributed to the eight massless gluons of quantum chromodynamics. Gravitation results from the exchange of a massless spin-two particle, the graviton, which is described by another local gauge theory: the general theory of relativity. In addition there is one surviving Higgs particle, which is massive and electrically neutral. In the coming years the search for the massive vector bosons and the Higgs particle will serve as tests of this synthesis.

gauge theories but only from a more comprehensive theory.

In the search for a larger theory it is natural to apply once more a recipe that has already proved successful. Hence the obvious program is to search for global symmetries and explore the consequences of making them local symmetries. This principle is not a necessary one, but it is worth trying. Just as Maxwell's theory combined electricity and magnetism and the Weinberg-Salam-Ward model linked electromagnetism and the weak force, so perhaps some larger theory could be found to embrace both the Weinberg-Salam-Ward model and quantum chromodynamics. Such a theory might in principle be constructed on the model of the existing gauge theories. A more sweeping symmetry of nature must be found; making this symmetry a local one would then give rise to the strong force, the weak force and electromagnetism. In the bargain yet more forces, exceedingly weak and so far unobserved, may be introduced.

Work on such theories is proceeding, and it has lately concentrated on symmetries that allow transformations between quarks and leptons, the class of particles that includes the electron. It is my belief the schemes proposed so far are not convincing. The grand symmetry they presuppose must be broken in order to account for the observed disparities among the forces, and that requires several Higgs fields. The resulting theory has as many arbitrary constants of nature as the less comprehensive theories it replaces.

A quite different and more ambitious approach to unification has recently been introduced under the terms "supersymmetry" and "supergravity." It gathers into a single category particles with various quantities of angular momentum; up to now particles with different spins were always assigned to separate categories. The utility of the supersymmetric theories has yet to be demonstrated, but they hold much promise. They offer a highly restrictive description of some hundreds of particles, including the graviton, in terms of only a few adjustable parameters. So far the results do not much resemble the known physical world, but that was also true of the first Yang-Mills theory in 1954.

The form of unification that has been sought longest and most ardently is a reconciliation of the various quantum field theories with the general theory of relativity. The gravitational field seems to lead inevitably to quantized theories that cannot be renormalized. At extremely small scales of distance (10^{-33} centimeter) and time (10^{-44} second) quantum fluctuations of space-time itself become important, and they call into question the very meaning of a space-time continuum. Here lie the present limits not merely of gauge theories but of all known physical theories.

The Author

GERARD 'T HOOFT is professor of physics at the University of Utrecht, from which he received his bachelor's degree in 1969 and his Ph.D. in 1972 and to which he returned as a member of the faculty after two years at the European Organization for Nuclear Research (CERN). "My interest in physics dates back to the age of six," he writes. "I thought I was going to choose between painting, music and physics. I chose physics, inspired by other physicists in my family, but I still like to make a painting or play the piano every now and then."

Bibliography

CONSERVATION OF ISOTOPIC SPIN AND ISOTOPIC GAUGE INVARIANCE. C. N. Yang and R. L. Mills in *Physical Review,* Vol. 96, No. 1, pages 191–195; October 1, 1954.

UNIFIED THEORIES OF ELEMENTARY-PARTICLE INTERACTION. Steven Weinberg in *Scientific American,* Vol. 231, No. 1, pages 50–59; July, 1974.

QUARKS WITH COLOR AND FLAVOR. Sheldon Lee Glashow in *Scientific American,* Vol. 233, No. 4, pages 38–50; October, 1975.