The development of field theory in the last 50 years

Victor F. Weisskopf

Victor Weisskopf is Institute Emeritus Professor of physics at MIT. From 1961–1965 he was Director General of CERN.

This article is devoted to the development of quantum field theory, a discipline that began with quantum electrodynamics,[1] which was born in 1927 when P. A. M. Dirac published his famous paper "The Quantum Theory of the Emission and Absorption of Radiation." Figure 1 reproduces the first page. Note that it was communicated by Niels Bohr himself. Also note the second and third sentences. The latter is an understatement indeed: *Nothing* had been done up to this time on quantum electrodynamics.

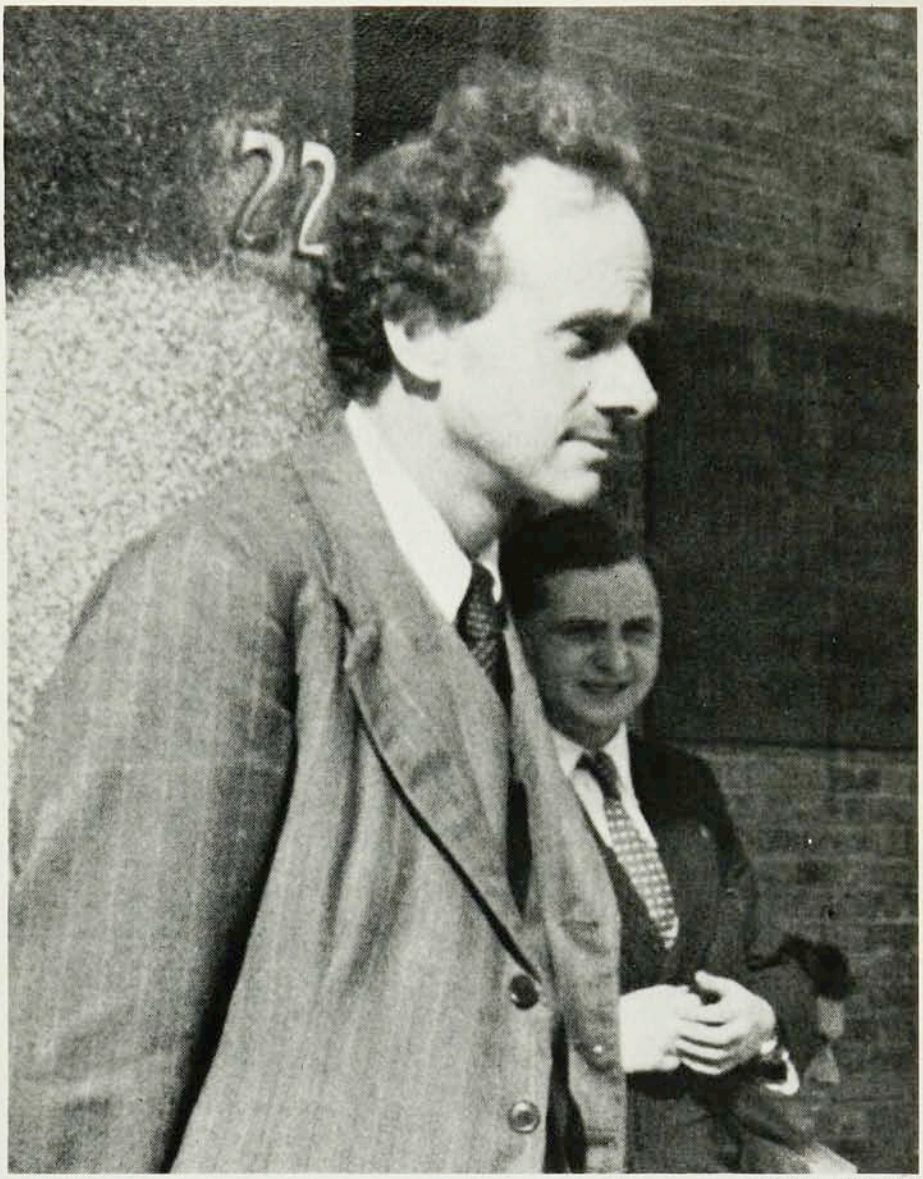
The pre-Dirac time

Classical electrodynamics started in 1862 when James Clerk Maxwell created his equations connecting the electric field **E** and the magnetic field **B** with the charge density ρ and the current density **j**. Together with the expression of the Lorentz force acting on a system carrying charge and current in an electromagnetic field, it led to an understanding of light as an electromagnetic wave, of the radiation emitted by moving charges and of the effects of radiation upon charged bodies. The results were splendidly verified by Heinrich Hertz in 1885 for radiations emitted and absorbed by antennas.

The application to atomic radiation was stymied by two facts: First, ρ and **j** in atoms were unknown to them; second, they faced a fundamental difficulty when the statistical theory of heat was applied to the radiation field. The number of degrees of freedom of a radiation field in a unit volume is infinite, and if each degree is supposed to get an energy $kT/2$ according to the equipartition theorem, the total energy density becomes infinity; empty space would be an infinite sink of radiation energy. Furthermore, apart from this distressing result, the classical theory of light had no explanation of the daily experience that incandescent matter changes its color with rising temperature—from red to yellow and then to white. The physicists must have felt before 1900 much as the neurophysiologists of today feel without any explanation of what memory is.

Then came quantum theory. It developed with increasing speed within a quarter century beginning with Max Planck's insight into the nature of blackbody radiation in 1900, followed by Albert Einstein's revolutionary idea of the existence of a photon in 1905, by Niels Bohr's atomic model in 1913, and by Louis DeBroglie's daring hypothesis of the wave–particle duality of particles in 1924. It reached its peak with the formulation of quantum mechanics by Werner Heisenberg, Erwin Schrödinger, Dirac, Wolfgang Pauli and Bohr in 1925.

0031-9228 / 81 / 110005-17 / $00.50



AIP NIELS BOHR LIBRARY

***Title page of paper** (below) by P. A. M. Dirac (left) on radiation theory (from Proceedings of the Royal Society **114**, 243, 1927).* *Figure 1*

The Quantum Theory of the Emission and Absorption of Radiation.

By P. A. M. Dirac, St. John's College, Cambridge, and Institute for Theoretical Physics, Copenhagen.

(Communicated by N. Bohr, For. Mem. R.S.—Received February 2, 1927.)

§ 1. *Introduction and Summary.*

The new quantum theory, based on the assumption that the dynamical variables do not obey the commutative law of multiplication, has by now been developed sufficiently to form a fairly complete theory of dynamics. One can treat mathematically the problem of any dynamical system composed of a number of particles with instantaneous forces acting between them, provided it is describable by a Hamiltonian function, and one can interpret the mathematics physically by a quite definite general method. On the other hand, hardly anything has been done up to the present on quantum electrodynamics. The questions of the correct treatment of a system in which the forces are propagated with the velocity of light instead of instantaneously, of the production of an electromagnetic field by a moving electron, and of the reaction of this field on the electron have not yet been touched. In addition, there is a serious difficulty in making the theory satisfy all the requirements of the restricted

The difficulties of the classical theory disappeared with one stroke—not without bringing about other difficulties about which much more will be said soon. Of course, the problem of heat radiation was immediately solved and the reasons for the sharp characteristic spectral lines of each atomic species became evident. Atomic stabilities, sizes and excitation energies could be derived from first principles: The chemical forces turned out to be a direct consequence of quantum mechanics; chemistry became part of physics.

However, before the publication of Dirac's 1927 paper, it was not possible to derive the expressions for ρ and **j** within the atoms for the purpose of calculating the emission of light quanta.

Actually, the Schrödinger equation allowed the calculation of transitions under the influence of an external radiation field, that is the absorption of light and the *forced* emission of an additonal photon in the presence of an incident radiation. The field of an incident light wave could be considered as a perturbation on the atom in the initial state; it was possible by means of the Schrödinger equation to calculate the probability of a transition, which turned out to be proportional to the intensity of the incident light wave. However, the emission by a transition from a higher to a lower state in a field-free vacuum could not be

treated. It was assumed at that time the matrix elements $\langle a|\rho|b\rangle$ and $\langle a|\mathbf{j}|b\rangle$ between two stationary states a, b of the atom play the role of charge and current density responsible for the radiation connected with the quantum transition from a to b or vice versa. The atom was considered as an "orchestra of oscillators," and the matrix elements determined the strengths of those oscillators ascribed to each pair of states. To determine the intensity of spontaneous emission, one had to use either the oscillator model and equate the emission with the classical radiation of these oscillators, or one had to use the Einstein relations, from which it follows that the probability of spontaneous emission from b to a is equal to the absorption probability from a to b when the light intensity per frequency interval $d\omega$ is put equal to a certain value I_0:

$$I_0 d\omega = \frac{\hbar\omega^3}{4\pi^2 c^2} d\omega \qquad (1)$$

This happens to be the light intensity when each degree of freedom of the radiation field contained one photon. According to this rule the probability of spontaneous emission is equal to the probability of a *forced* emission by a fictitious radiation field of the intensity 1.

But why? According to the Schrödinger equation, any stationary state should have an infinite lifetime when there is no radiation present.

Quantization of the radiation field

Dirac's fundamental paper in 1927 changed all that. Quantum mechanics must be applied not only to the atom via the Schrödinger equation, but also to the radiation field. Dirac made use of the old idea of Paul Ehrenfest (1906) and Peter Debye (1910), to describe the electromagnetic field in empty space as a system of quantized oscillators. In the presence of atoms or of other systems of charged particles, the coupling between the charged particles and the field is expressed by an interaction energy

$$H^1 = e\int \mathbf{j}\cdot\mathbf{A}\,dx^3 \qquad (2)$$

where $\mathbf{j}$ is the current density of the particles. The value e of the particle charge is inserted here as an explicit factor and $\mathbf{A}$ is the vector potential. Both magnitudes are operators in the quantized system of the atom and the field oscillators. Expression 2 is a direct consequence of Maxwell's equations. The Hamiltonian of the combined system then has the form

$$H = H_0 + H^1 \qquad (3)$$
$$H_0 = H_{\text{field}} + H_{\text{atom}}$$

where H_{field} is the Hamiltonian of the isolated field oscillators and H_{atom} is the Schrödinger Hamiltonian of the atom isolated from the electromagnetic fields.

The Hamiltonian H_0 describes field and atom without interaction. The effects of H^1 are treated as a perturbation upon the system H_0. The stationary states of H_0 are characterized by

$$(\ldots n_i \ldots;\ a) \qquad (4)$$

Here n_i are the occupation numbers of the radiation oscillators (the numbers of photons present in each oscillator i) and a indicates the stationary state of the atom.

The states 4 are no longer stationary when the perturbation energy H^1 is taken into account. The theory yields simply and directly the laws of emission and absorption of light. Indeed, the state $(\ldots 0,0,\ldots;a)$ of an atom in an excited state a without any radiation present is not stationary according to the Hamiltonian 3. A first-order perturbation calculation gives a probability $P_{ab}\,d\Omega$ per unit time for a transition from a to a lower state b, accompanied by the emission of a photon of a frequency $\omega = (\epsilon_a - \epsilon_b)/\hbar$ into the solid angle $d\Omega$ and with a polarization vector $\mathbf{s}$:

$$P_{ab}\,d\Omega = \frac{e^2}{\hbar c}\frac{(2\pi)^2}{\hbar\omega^2} I_0 |\mathbf{s}\mathbf{j}_{ab}|^2 d\Omega \qquad (5)$$

I_0 is given by the expression 1. The matrix element is determined by (for a one-electron system)

$$\mathbf{j}_{ab} = \int \psi_a^* \mathbf{j} \exp(i\,\mathbf{k}_{ab}\mathbf{x})\psi_b\,dx^3$$

where $\mathbf{j}$ is the operator of the current, and $\mathbf{k}_{ab}$ the wave vector of the emitted quantum. The effect of the size of the system compared to the wavelength is taken into account by the exponential; it was neglected in the oscillator picture (dipole approximation). According to equation 5 spontaneous emission appears as a forced emission caused by the zero-point oscillations of the electromagnetic field, which are always present, even in a space without any photons.

This was the start of an interesting development in theoretical physics. After Einstein had put an end to the concept of aether, the field-free and matter-free vacuum was considered as a truly "empty space." The introduction of quantum mechanics changed this situation and the vacuum gradually became "populated." In quantum mechanics an oscillator cannot be exactly at its rest position except at the expense of an infinite momentum, according to Heisenberg's uncertainty relation. The oscillatory nature of the radiation field therefore requires zero-point oscillations of the electromagnetic fields in the vacuum state, which is the state of lowest energy. The spontaneous emission process can be interpreted as a consequence of these oscillations.

Dirac's theory produced all results regarding the absorption and emission of light by atoms that previously were obtained by unreliable arguments. The results followed from the Hamiltonian 3 when the interaction energy 2 was treated as a first-order perturbation. Some other radiation phenomena such as photon scattering processes, resonance fluorescence and nonrelativistic Compton scattering of photons by electrons, appear in the second order of the perturbation treatment. The theory gave excellent account of all radiation phenomena in that order of perturbation in which they first appear. The higher approximations give rise to difficulties, which will be discussed later on.

Coupling to relativistic systems

In 1928 Dirac published two papers on a new relativistic wave equation of the electron. It was his third great contribution to the foundations of physics; the first was the reformulation of quantum mechanics, the "transformation theory,"[2] the second was the theory of radiation. The Dirac equation was supposed to replace Schrödinger's equation for cases where electron energies and momenta are too high for a nonrelativistic treatment. It immediately gave rise to four great triumphs:

▶ The spin $\hbar/2$ of the electron appeared to be a natural consequence of the relativistic wave equation. (It turned out later that there exist relativistic wave equations for particles with different spin. Dirac's equation for a spin $\hbar/2$ is distinguished by the fact that the energy operator appears linearly.)
▶ The g-factor of the electron necessarily has the value $g = 2$. The value of the magnetic moment of the electron followed directly from the equation.
▶ When applied to the hydrogen atom, the equation yields the correct Sommerfeld formula for the fine structure of the hydrogen spectrum.

The coupling of the quantized radiation field with the Dirac equation made it possible to calculate the interaction of light with relativistic electrons. The most important results were the derivation of the Klein–Nishina formula for the scattering of light by electrons, the Møller formula for the scattering of two relativistic electrons, and the emission of photons when electrons are scattered by the Coulomb field of nuclei.

In spite of these amazing successes a number of serious difficulties turned up immediately and it took a long time to solve them. The difficulties came from the existence of states of negative kinetic energy or negative mass. There was no way to get rid of them. If one tried to exclude them from the Hilbert space of the electron, the space becomes incomplete; furthermore, the Klein–Nishina formula could not be derived without them. Taken at face value, the existence of those states would imply that the hydrogen atom is not stable because of radiative transitions from the ordinary states to the states of negative energy. The properties of those impossible states were constantly in the center of discussion during those years. George Gamow referred to electrons in these states as "donkey electrons" because they tend to move in the opposite direction to the applied force.

Triumph and curse of the filled vacuum

It was again Dirac who proposed a way out of the difficulty in 1929. As it happens with ideas of great men, it was not only "a way out of a difficulty" but it was a seminal idea that led to the recognition of the existence of antimatter and ultimately to the development of field theory with all its concomitant insights into the nature of matter. He made use of the Pauli principle and assumed that, in the vacuum, all states of negative kinetic energy are occupied. This was the second step in the development of "populating" the vacuum. Later on this step was somewhat mitigated by eliminating the notion of an actual presence of those electrons, but the fluctuations of matter density in the vacuum remained as an additional property of the vacuum besides the electromagnetic vacuum fluctuations.

Dirac's daring assumption had most disturbing consequences, such as an infinite charge density and infinite (negative) energy density of the vacuum. Some of these impossible consequences were circumvented later, as is reported in the next section. However, the assumption not only solved most of the problems of the negative energy states but led to an impressive and unexpected broadening of our views about matter.

First of all, the transitions from positive to negative energy states were excluded, and the stability of the atoms was assured. Furthermore, Dirac's assumption required the existence of processes in which one particle from the "sea" of filled negative states is lifted to a state of positive energy, if the necessary energy is supplied by absorption of photons or by other means. A hole in the sea and a normal particle would be created. The hole would have all the properties of a particle of opposite charge. Moreover, a particle could fall back into a hole with the emission of photons of the right amount of energy and momentum. This, of course, would be a process of particle–antiparticle annihilation. Thus Dirac's assumption led to the recognition of the existence of antiparticles and of the existence of two new fundamental processes: pair creation and annihilation.

In the beginning these ideas seemed incredible and unnatural to everybody. No positive electron was ever seen at that time; the asymmetry of charges, positive for the heavy nuclei, negative for the light electrons, seemed to be a basic property of matter. Even Dirac shrank away from the concept of antimatter and tried to interpret the positive "holes" in the sea of the vacuum electrons as being protons. It was soon recognized, however, by Hermann Weyl, Robert Oppenheimer and by Dirac himself, that this interpretation would again lead to an unstable hydrogen atom and that the holes must have the same mass as the particles. Antimatter ought to exist. Indeed the positron was found by Carl Anderson in 1932; the antiproton was discovered 25 years later because its production needed energy concentrations several thousand times higher than were available before the invention of the synchrocyclotron. (The possibility of antiparticles was already mentioned by Pauli[3] and Einstein.[4] More about this can be found in a review by A. Pais.[5])

Once the idea of the filled vacuum took hold, it was relatively easy to calculate the cross section for the annihilation of an electron and a positron into two photons and the cross section for pair creation by photons in the Coulomb field of atomic nuclei. It is astonishing that it took more than three years after the identification of the holes with positrons, before the pair creation in a Coulomb field was calculated, although it was a very simple determination of a transition probability. It illustrates the wonder and incredulity that those ideas encountered during the first years.

Today it is hard to realize the excitement, the skepticism and the enthusiasm aroused in the early years by the development of all the new insights that emerged from the Dirac equation. A great deal more was hidden in the Dirac equation than the author had expected when he wrote it down in 1928. Dirac himself remarked in one of his talks that his equation was more intelligent than its author. But it was Dirac who found most of the additional insights himself.

The formulas derived for the creation of pairs and for radiative scattering (Bremsstrahlung) also gave an excellent account of the development of cosmic-ray cascade showers in matter, once the incoming energy is transformed into electrons and photons. It is interesting to observe how this success was interpreted. First it was considered as proof that radi-

ation theory and pair creation are valid even at very high energy. Then, when it turned out that a part of the cosmic rays do not form showers (the part consisting of the then-unknown muons), doubts were expressed as to the validity of radiation theory at high energies. But it was shown by Enrico Fermi[6] and then by C. F. Von Weizsäcker[7] and E. J. Williams[8] that the effect of a Coulomb field on a fast-moving electron can be expressed as the effect of light quanta whose energy is only a few mc^2, when a suitable system of reference was used (the system in which the electron is at rest). This analysis of the production of cascade showers showed clearly that only energies and momenta of the order mc^2 and mc are exchanged in the relevant processes. Hence the shower production does not test the theory at high energies, nor could any deviation from the expected showers be explained by a breakdown of the theory at high energies.

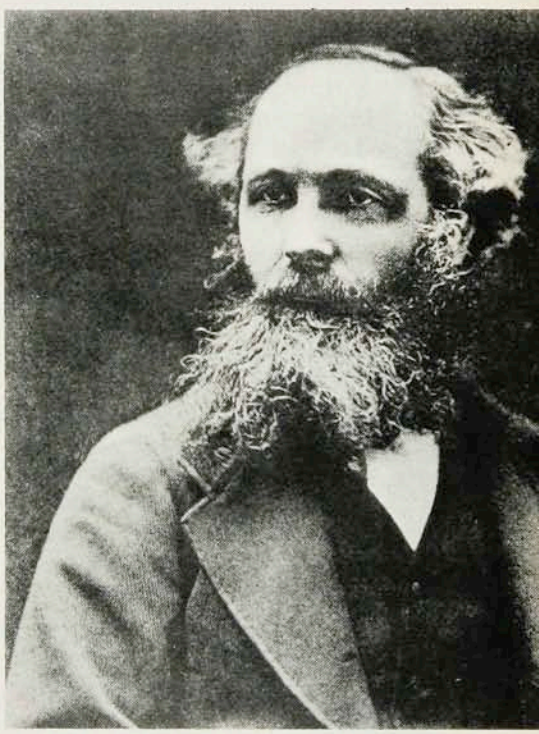
Indeed, electron accelerators of many GeV were needed to test the theory at large energies. Recent measurements with electron–positron colliders have shown radiation theory to be valid at least up to energy exchanges of 100 GeV.

How unreasonable the idea of antimatter seemed at that time may be illustrated by the fact that many of us did not believe in the existence of an antiparticle to the proton because of its anomalous magnetic moment. The latter was measured by Otto Stern in 1933 and could be interpreted as an indication that the proton does not obey the Dirac equation. The fundamental character of the matter–antimatter symmetry and its independence of the special wave equations was recognized only very slowly by most physicists.

The following conclusions must be drawn from the new interpretation of the negative-energy states in the Dirac equation. There are no real one-particle systems in Nature, not even few-particle systems. Only in nonrelativistic quantum mechanics are we justified to consider the hydrogen atom as a two-particle system; not so in the relativistic case, because we must include the presence of an infinite number of vacuum electrons. Even if we consider the filled vacuum as a clumsy description of reality, the existence of virtual pairs and of pair fluctuations shows that the days of fixed particle numbers are over.

Furthermore, relativity requires that time and space be treated equivalently. In nonrelativistic quantum mechanics, time is a parameter, whereas the space coordinates of the particles are considered as operators. In relativistic quantum mechanics the

James Clerk Maxwell



AIP NIELS BOHR LIBRARY

AIP NIELS BOHR LIBRARY

Discovery of the positron. *Cloud chamber photo by Carl Anderson in 1931 showing the first recorded positron track.*

particles appear as quanta of a field, just as the photons are quanta of the electromagnetic one. The fields assume the role of operators and the coordinates are parameters indicating the space- or time-dependence of the field operators. The theory of the interaction of charged particles with the radiation field becomes a field theory in which two (or more) quantized fields interact: the matter field and the radiation field.

The field amplitudes are expressed as linear combinations of creation and destruction operators that increase or decrease the number of particles in the quantum states of the system. It is a direct generalization of the quantization of the electromagnetic field as decomposed into oscillator amplitudes. The operator of an oscillator amplitude contains matrix elements only between states that differ by one unit of excitation. The corresponding operator either adds (creates) or subtracts (destroys) a quantum of the oscillator.

There are essential differences between a field of particles with spin ½ and the radiation field. The former describes the behavior of fermions, whereas the latter is an example of a boson field. In the classical limit, the boson fields are classical fields whose field strength is a well-defined function of space and time (radio wave). The fermion fields cannot have a classical limit because no more than one fermion can be put into one wave; its classical limit is a particle with a well-defined momentum and position. So far, the constituents of matter have all been shown to be fermions interacting by means of boson fields.

Furthermore, the interaction between fermion and boson fields in its simplest form necessarily is bilinear in the fermion fields and linear in the boson fields. This is indicated by the fact that the current density is a bilinear expression of the particle wave functions. One cannot construct a Lorentz-invariant expression that is linear or cubic in the spinor wave functions. Boson field (vector or scalar), however, may appear linearly in the interaction.

When the fields are expressed in terms of creation and annihilation operators, the form of the interaction can be interpreted in the following way: The fundamental interaction between fermions and bosons consists of the product of

two fermion creation or destruction operators $b^\dagger$ and b, and one boson operator a or $a^\dagger$: $b^\dagger ba$ or $b^\dagger ba^\dagger$. It is interpreted as a change of state of a fermion "destroyed" in one state and "created" in another) accompanied with either an emission or an absorption of a boson.

The fight against infinities: elimination of the vacuum electrons

In spite of all successes of the hole theory of the positron, the infinite charge density and the infinite negative energy density of the vacuum made it very difficult to accept the theory at its face value. A war against infinities started at that time. It was waged with increasing fervor by the developers of quantum electrodynamics when more intricate infinities appeared besides those mentioned before, as will be described in the subsequent sections.

There is a rather primitive way to take care of the infinite charge density, by a slight change in the definition of charge and current. It amounts to the following argument: Because the theory is completely symmetric in regard to electrons and positrons, it would be equally valid to construct a theory in which the positrons are the particles and the electrons are the holes in a sea of positrons that occupy negative energy states. The actual theory then could be considered as a superposition of these two theories, one with an infinite negative charge density and the other with infinite positive one. This combination also serves to emphasize the symmetry between matter and antimatter. The vacuum charge densities cancel; the corresponding expressions for charge and current indeed give a more satisfactory description of the phenomena.

It was recognized in 1934 by Heisenberg[9] and by Oppenheimer and Wendell Furry[10] that the creation and destruction operators are most suitable for turning the liability of the negative energy states into an asset, by interchanging the role of creation and destruction of those operators that act upon the negative states. This interchange can be done in a consistent way without any fundamental change of the equations. The consequences are identical to those of the filled-vacuum assumption, but it is not necessary to introduce that disagreeable assumption explicitly. Particles and antiparticles enter symmetrically into the formalism, and the infinite charge density of the vacuum disappears. One even can get rid of the infinite negative-energy density by a suitable rearrangement of the bilinear terms of the creation and destruction operators in the Hamiltonian. After all, in a relativistic theory the vacuum must have vanishing energy and momentum. There remains, however, the unpleasant fact of the existence of vacuum fluctuations without any energy.

The fundamental interaction between charged fermions and photons now contains three basic processes: the scattering of a fermion with the emission or absorption of a photon, the creation and the annihilation of a fermion–antifermion pair with the emission or absorption of a photon. All electrodynamic interaction processes are combinations of these fundamental steps.

Surprisingly enough, it took many years before the physicists realized the great advantages of this new formalism. One still reads about the "hole theory" of positrons in papers written in the late 1940s, when renormalization was the topic of the day.

An interesting episode in the fight for the elimination of vacuum electrons was the quantization of the Klein–Gordon relativistic wave equation for scalar particles. It seemed to be a rather academic activity because no scalar particle was known at that time. In that theory, the charge density $(\phi^*\phi - \phi\phi^*)$ and the wave intensity $|\phi|^2$ are not identical. Therefore, it seemed posssible that, under the influence of external electromagnetic fields, the total intensity $\int|\phi|^2 dx^3$ may change in time, although the total charge remains conserved. It smelled of a creation or annihilation process of oppositely charged particles. The problem attracted the attention of Pauli and myself[11] because we saw that the quantized Klein–Gordon equation gives rise to particles and antiparticles and to pair creation and annihilation processes without introducing a vacuum full of particles. Note that at the time the method of exchanging the creation and destruction operators (for negative energy states) was not yet in fashion; the hole theory of the filled vacuum was still the accepted way of dealing with positrons. Pauli called our work the "anti-Dirac paper;" he considered it as a weapon in the fight against the filled vacuum, which he never liked. We thought that this theory only served the purpose of an unrealistic example of a theory that contained all the advantages of the hole theory without the necessity of filling the vacuum. We had no idea that the world of particles would abound with spin-zero entities a quarter of a century later. This was the reason why we published it in the venerable but not widely read *Helvetica Physica Acta*.

Our work on the quantization of the Klein–Gordon equation led Pauli to formulate the famous relation between spin and statistics. Pauli demonstrated in 1936 the impossibility of quantizing equations of scalar or vector fields that obey anticommutation rules. He showed that such relations would have the consequence that physical operators do not commute at two points that differ by a spacelike interval. This lack of commutativity would contradict causality because it would require that mea-

Hideki Yukawa

surements interfere with each other when no signal can pass from one to the other. Thus Pauli concluded that particles with integer spin cannot obey Fermi statistics. They must be bosons. During the days of the hole theory it was obvious that particles with spin ½ cannot obey Bose statistics because it would be impossible to "fill" the vacuum. Four years later Pauli proved the necessity of Fermi statistics for half-integer spins, also on the basis of the same causality arguments.

The fight against infinities: infinite self mass

The infinities of the filled vacuum and of the zero-point energy of the vacuum turned out to be relatively harmless compared to other infinities that appeared in quantum electrodynamics when the coupling between the charged particles and the radiation field was considered in detail. No difficulties appeared as long as only the first terms of the perturbation treatment were taken into account, that is those terms in which the phenomena under consideration appear in the lowest order. It soon turned out that the higher terms always contain infinities, as Oppenheimer[12] had pointed out for the first time.

In 1934 Pauli asked me to calculate the self energy of an electron according to the positron theory. It was a modern repetition of an old problem of electrodynamics. In classical theory the energy contained in the field of an electron of radius a (neglecting the inside) is $4\pi e^2/a$ and would diverge linearly if the radius goes to zero. The corresponding calculation in the positron theory is much more complicated. One had to calculate the difference between two infinite amounts: the energy of the vacuum and the energy of the vacuum plus one electron. The result was equivalent to the statement that the electric field inside one Compton wave length $\lambda_c = h/mc$ from the electron is not e/r^2 but $(e/r^2)(r/\lambda_c)^{1/2}$. When r goes to zero it increases only as $r^{-3/2}$. The self energy then becomes[12]

$$E = m_0c^2 + (3/2\pi)m_0c^2 \times (e^2/\hbar c)\log(\lambda_c/a) \qquad (6)$$

where m_0 is the intrinsic or "mechanical" mass of the electron, which appears in the Hamiltonian of the electron when it is decoupled from the electromagnetic field. It diverges only logarithmically.

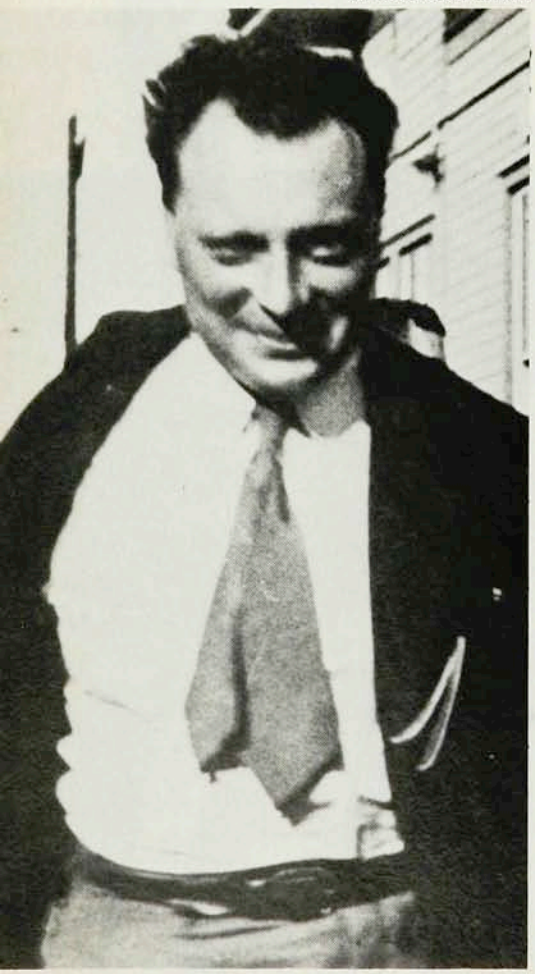
(This brings back one of the dark moments of my professional career. I made a mistake in the first publication that resulted in a quadratic divergence of the self-energy. Then I received a letter from Furry, who kindly pointed out my rather silly mistake and the fact that actually the divergence is logarithmic. Instead of publishing the result himself, he allowed me to publish a correction quoting his intervention. Since then the discovery of the logarithmic divergence of the electron self-energy is wrongly ascribed to me instead of to Furry.)

A consistent relativistic theory requires a point electron, that is $a \to 0$. It is worth noting, however, that the value of a for which the second term of 6 becomes half of the first is as small as 10^{-72} cm! Even the Schwarzschild radius of the electron is only 10^{-55} cm. This value means that the deformation of the space around the electron is strong enough to prevent the electron from interacting with photons of that wave length, thus providing a natural cut-off long before the electromagnetic self-energy becomes important. Unfortunately, no consistent calculation of this effect has ever succeeded.

Another somewhat more benign type of infinities appeared in quantum electrodynamics when emissions of photons of very low frequencies were considered. Such emissions take place, for example, when electron beams are scattered by static electric fields. Classical theory predicts that the emitted energy does not vanish in the limit of zero frequencies. The quantum result ought to be identical with the classical one at that limit; it would indicate that the number of emitted quanta goes to infinity. This trouble, called "infrared catastrophe," can be avoided by describing this limit with the help of classical fields, as Bloch and Arnold Nordsieck[14] have shown in their important paper of 1937. It put an end to any worries about this kind of infinity.

***Wolfgang Pauli** in 1931*

AIP NIELS BOHR LIBRARY



The fight against infinities: infinite vacuum polarization

The virtual pairs endow the vacuum with properties similar to a dielectric medium. We may ascribe a dielectric coefficient ϵ to the vacuum. A direct calculation of this dielectric effect leads to a dielectric coefficient that consists of a constant part ϵ_0 and an additional part that depends upon the electromagnetic fields and their derivatives in time and space.

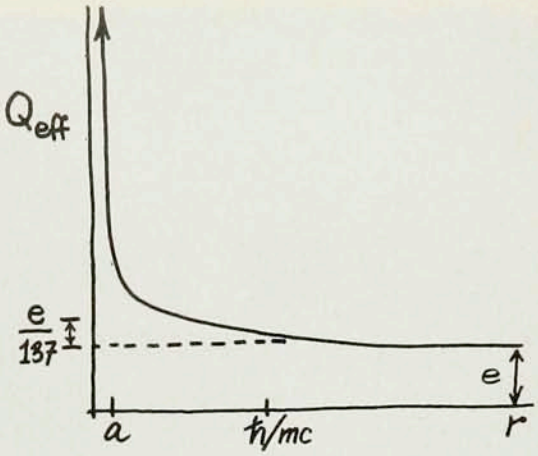
$$\epsilon = \epsilon_0 + \epsilon(\text{field}) \qquad (7)$$

The constant part ϵ_0 cannot have any physical significance because it serves only to redefine the unit of charge. Any charge Q_0 would appear as $Q = Q_0/\epsilon$. The actual value of ϵ_0 turns out to be logarithmically divergent (it goes as $\log(\Lambda/m)$ where Λ is the highest momentum considered in the cal-

culation). The additional field-dependent term, however, turns out to be finite and therefore should have physical significance.

Let us now consider what happens to a charge Q_0 when placed in a vacuum with a dielectric coefficient of the form 7. At large distances r the effective charge will be Q_0/ϵ_0. When r becomes of the order $\lambda_c = \hbar/(mc)$ or less the second term of 7 becomes important. Calculations of this term for a Coulomb field were carried out by Robert Serber[15] and E. Uehling.[16] They found that $\epsilon(r)$ decreases with r when r becomes smaller than the Compton wave length λ_c. This is so because, for smaller r, only those virtual pairs contribute whose energy is larger than $\hbar c/r$. This decrease is finite and calculable. The infinite value of ϵ_0 was interpreted as an indication that the intrinsic "true" charge Q_0 is infinite so that the observed charge becomes finite and equal to $e = Q_0/\epsilon_0$ for $r \to \infty$. The decrease of ϵ with decreasing r when $r < \lambda_c$ would then amount to an increase of the effective charge Q_{eff} at those small distances.

This increase of Q_{eff} for $r < \lambda_c$ over the value e at large distances is rather small; it is of the order of $e/137$. A strong increase occurs only at very small distances $r \sim \lambda_c \exp(-\hbar c/e^2)$; these are the same distances as the ones we discussed in connection with the self-energy, at which the theory most likely is inapplicable. We then get a dependence of Q_{eff} on the distance as shown in figure 2. It is the first example of a "running coupling constant," which plays an important role in quantum chromodynamics.

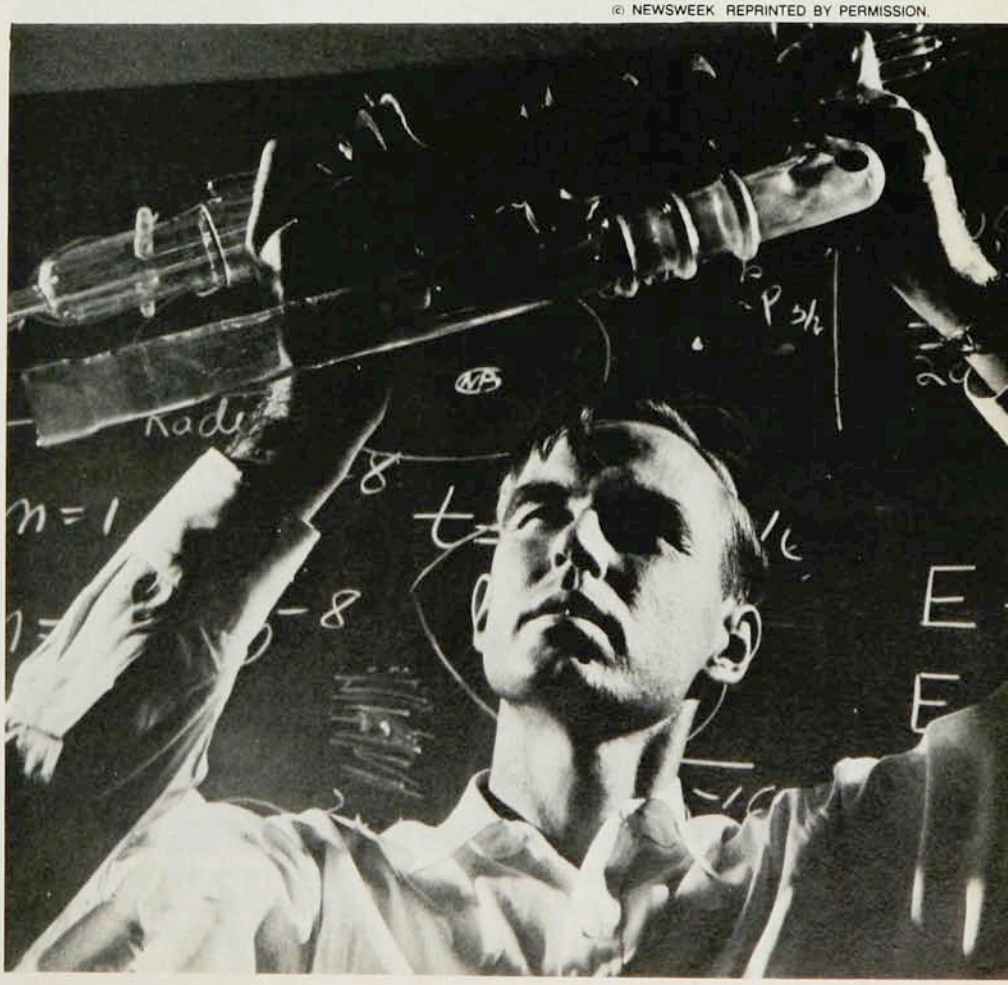


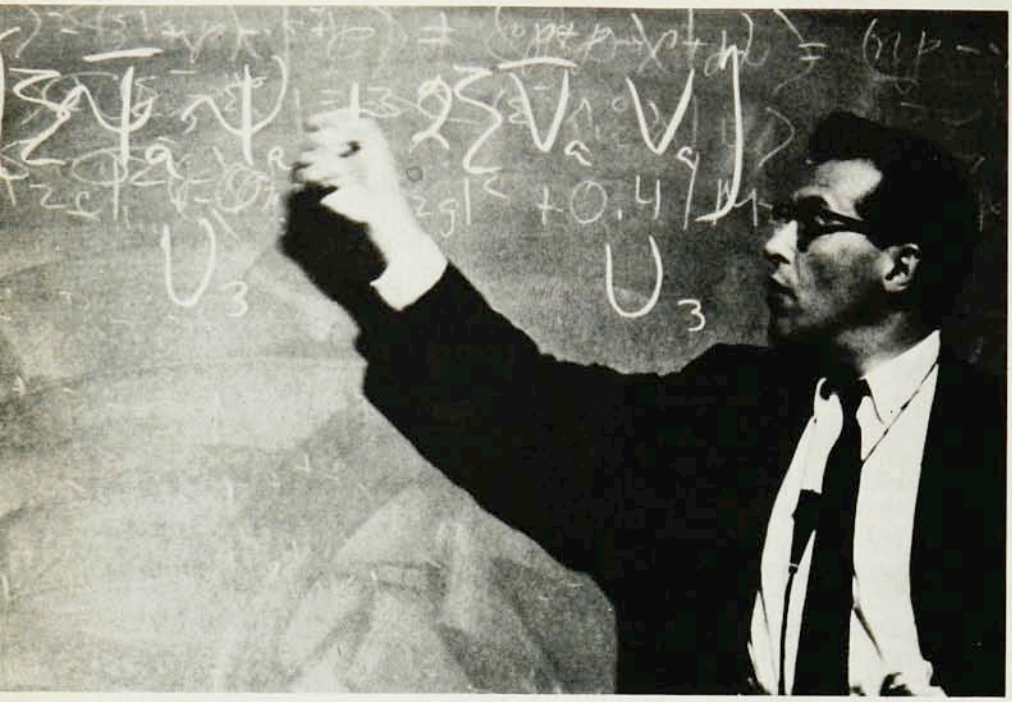
Running coupling constant *in QED. The effective charge* Q_{eff} *as a function of the distance* **r**. *The distance* **a**, *the distance at which* Q_{eff} *is about 137* **e**, *is very much smaller than indicated in this drawing.* *Figure 2*

The fight against infinities: renormalization

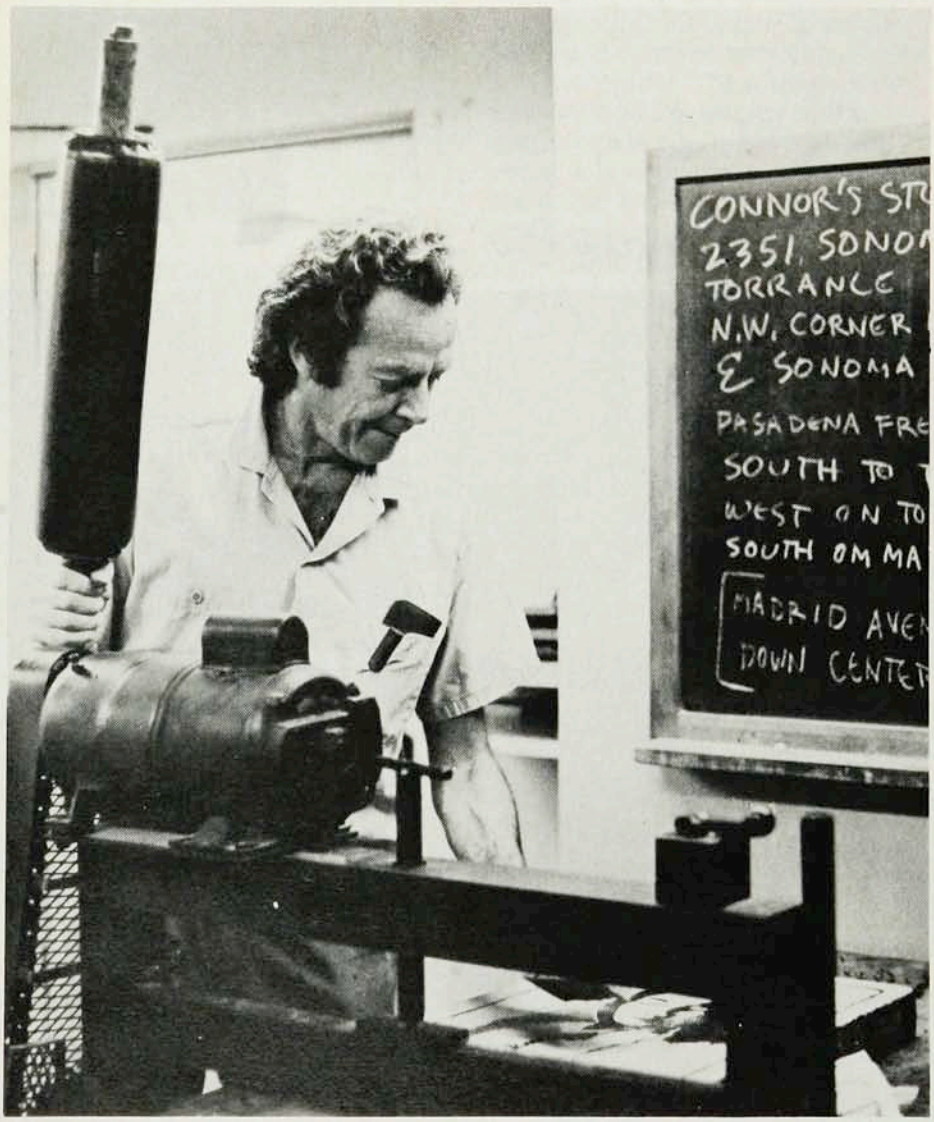
The appearance of infinite magnitudes in quantum electrodynamics was noticed in 1930. Because they only occurred when a certain phenomenon was calculated to a higher order of perturbation theory than the lowest one in which it appeared, it was possible to ignore the infinities and stick to the lowest-order results that were good enough for the experimental accuracy at that period. However, the infinities at higher order indicated that the formalism contained unrealistic contributions from the inter-

Willis Lamb *in 1947*





Julian Schwinger



Richard Feynman *(photo by Sylvia Posner, courtesy of the CalTech Archives)*

action with high-momentum photons.

Already in 1936 the conjecture was expressed[17,18] that the infinite contributions of the high-momentum photons are all connected with the infinite self mass, the infinite intrinsic charge Q_0 and with nonmeasurable vacuum quantities such as a constant dielectric coefficient of the vacuum. Thus it seemed that a systematic theory could be developed in which these infinities are circumvented. At that time nobody attempted to formulate such a theory, although it would have been possible then to develop what is now known as the method of renormalization.

There was one tragic exception and that was E. C. G. Stueckelberg.[19,20] He wrote several important papers in 1934–38, putting forward a manifestly invariant formulation of field theory. This could have been a basis of developing the ideas of renormalization. Later on (in 1947) he actually formulated the complete renormalization procedure quite independently of the efforts of other authors. Unfortunately, his writings and his talks were rather obscure and it was very difficult to understand them or to make use of his methods. Had the theorists been capable of grasping his ideas they may well have calculated the Lamb shift and the correction to the magnetic moment of the electron at a much earlier time.

A new impetus to such attempts came from an experimental result. Willis Lamb and R. C. Retherford[21] were able to measure reliably the difference in energy between the $2S_{1/2}$ and $2P_{1/2}$ state of hydrogen (Lamb shift). The two states should have been exactly degenerate according to the Dirac equation applied to the hydrogen problem. Already in the 1930s the degeneracy of these two levels was in doubt from spectroscopic measurements, but Lamb and Retherford, using newly developed microwave methods, definitely established the splitting and measured it with great precision.

It had been conjectured long ago that such a splitting should be caused by the coupling of the radiation field with the atom, but early attempts to calculate it ran into difficulties because the infinite mass and vacuum polarization appeared in the same approximation. It was

H. A. Kramers who pointed out[22] that one ought to be able to calculate the effect by carefully subtracting the infinite energy of the bound electron from that of the free one and thereby separating the parts that contribute to the mass and charge from those of real significance. Infinities are always difficult to subtract in an unambiguous way. After the Lamb shift had been measured, Bethe had made an attempt to estimate the effect of the radiation coupling, simply by omitting the coupling with photons of an energy larger than mc^2. This attempt was successful because most of the effect comes from the coupling with photons of lower energy, which can be treated nonrelativistically.

An exact calculation to the lowest order in $(e^2/\hbar c)$ was then performed by Norman M. Kroll and Lamb[23] and by J. B. French and myself[24] (1949) and resulted in good agreement with the experiment. However, the methods used by those authors of subtracting two infinities were clumsy and unreliable. Subsequently, a formidable group of physicists, including Julian Schwinger, Richard Feynman, Freeman Dyson and Sin-Itiro Tomonaga, developed a reliable way to deal with the infinities. They introduced a method of renormalization in which the initial parameters were eliminated in favor of those with immediate physical significance. In any computation of an electrodynamical result, the effects of the mass and charge redefinitions had to be incorporated. Infinite "counterterms" are introduced into the Hamiltonian in such a manner that they compensate for the infinite mass and charge. In order to make this procedure unambiguous it was necessary to keep the expressions in a manifestly relativistic and gauge-invariant form throughout the calculations.

The results were most encouraging. Schwinger found that the magnetic moment of the electron should indeed be larger by the factor $1 + \alpha/(2\pi)$ than the Bohr magneton, a result that was observed shortly before by I. I Rabi and his disciples and then more accurately by Henry Foley and Polykarp Kusch. The Lamb-shift results were recalculated in a much simpler way, radiative corrections of higher order in $e^2/\hbar c$ to scattering processes were unambiguously determined, and the vacuum polarization effects were worked out in detail; the latter found an impressive experimental confirmation in the measurements of the spectrum of muonic atoms (the electron replaced by a muon); the muon moves in the region $r \sim (\hbar/m_e c)$ where the vacuum polarization is a one-percent effect. Another remarkable test of the new methods was the agreement between the predicted and observed properties of positronium—the atom consisting of an electron and a positron, discovered and investigated for the first time by Martin Deutsch.

The war against infinities was ended. There was no reason any more to fear the higher-order terms. The renormalization took care of all infinities and provides an unambiguous way to calculate with any desired accuracy any phenomenon resulting from the coupling of electrons with the electromagnetic field. It was not a complete victory, because infinite counter-terms had to be introduced to remove the infinities. Furthermore, the procedure of eliminating infinities could be carried out only by renormalizing successively at each step of the perturbation expansion in powers of the coupling parameter. It still is not clear whether this method leads to a convergent series. It is like Hercules's fight against Hydra, the many-headed sea monster, which grows a new head for every one cut off. But Hercules won his fight and so did the physicists. Sidney Drell characterized the situation most aptly as "a peaceful coexistence with the infinities."

Sin-Itiro Tomonaga

Here are the signs of victory in the war against infinities:

▶ **Lamb shift** (about 10% is due to vacuum polarization; most of the rest is the interaction with the zero-point oscillations of the electromagnetic field):

$$\Delta\nu(2S_{1/2} - 2P_{1/2}) = \begin{matrix} 1057.862\ (20)\ \text{MHz (exp.)} \\ 1057.864\ (14)\ \text{MHz (theor.)} \end{matrix}$$

▶ ***g*-factor of the electron** $(a = \frac{1}{2}(g-2)) \times 10^3$

$$a = \begin{matrix} 1.15965241\ (20)\ \text{(exp.)} \\ 1.159652379\ (261)\ \text{(theor.)} \end{matrix}$$

▶ **Vacuum polarization.** 90% of the Lamb shift in muonic helium (α particle + muon) is caused by vacuum polarization:

$$\Delta E(2S_{1/2} - 2P_{3/2}) = \begin{matrix} 1.5274\ (0.9)\ \text{eV (exp.)} \\ 1.5251\ (9)\ \text{eV (theor.)} \end{matrix}$$

In spite of these victories there remain nagging problems in quantum electrodynamics. There are definite indications that we understand only a partial aspect of what is going on. As was mentioned before, the elimination of infinities is possible only in a perturbation approach; it is contingent upon the smallness of e^2/hc. But the effective coupling constant at very small (indeed incredibly small) distances becomes larger than unity. Will there be a theory that avoids renormalization by using nonperturbative methods? Or will a future unification of electrody-

E. B. BOATNER

Steven Weinberg

Abdus Salam

E. B. BOATNER

Sheldon L. Glashow

namics and general relativity heal the disease of divergencies because of the fact that the dangerous distances are smaller than the Schwarzschild radius of the electron?

Moreover, there is no way to understand and derive the mass of the electron within today's electrodynamics. This problem has become even more acute since heavier electrons such as the muon and the τ-electron have been discovered. There is not the slightest indication why electrons with different masses should exist. In present-day field theories the masses are arbitrary parameters that may assume any values.

Quantum electro-weak dynamics

The tremendous quantitative success of renormalized quantum electrodynamics (QED) has elevated this theory as an (almost) spotless example of a physical theory dealing with the interactions of electrically charged particles with fields. No wonder that the physicists tried to apply similar methods whenever interactions between fermions and bosons occurred. The first well-known use of QED as an example was the attempt of Hideki Yukawa (1935) to describe the nuclear force between protons and neutrons as an emission and subsequent absorption of a virtual boson. He had to ascribe a mass to that boson, because the nuclear force has a short range r_0 of the order of 10^{-13} cm. Any field theory modelled after QED would give an exponential force between fermions of the form $r^{-1}\,e^{-rMc/\hbar}$, with M the mass of the boson. The observed range of nuclear forces leads to a mass of about 200 MeV. No such bosons were known at that time, but he predicted the existence of them. His prediction was confirmed ten years later—an impressive success of a simple idea. Actually the nuclear force turned out to be the effect of somewhat more complicated processes; it does not detract from the beauty of his prediction.

The second early attempt to use QED as an example is a little known contribution by Oskar Klein.[25] He suggested a model for the weak interactions in which massive charged vector bosons mediated processes such as β decay. He even called them by the currently used letter W. He was the first to propose that the neutron decay: $n \rightarrow p + e + \bar{\nu}$ be split into two consecutive steps:

$$n \rightarrow p + W^-, \quad W^- \rightarrow e + \bar{\nu} \qquad (8)$$

He even went as far as to assume that the coupling constant for such processes is $e^2/\hbar c$, the same as for electromagnetic events. He attributed the smallness and the short range of the weak interactions to a large mass of the W, as it is done today, and he arrives at a W mass of about 100 GeV. This was 20 years before Schwinger independently took up this idea again. Schwinger initiated a development that brought forward the present unified quantum electro-weak dynamics, referred to as QEWD, a development in which a large number of theorists took part, including Martinus Veltman, Gerard 't Hooft, P. W. Higgs, R. Brout, Sheldon Glashow, Steven Weinberg, Benjamin W. Lee and Abdus Salam. An excellent historical survey has been written by Sidney Coleman.[26]

Before entering the discussion of those new ideas it is necessary to modernize the relations 8. We assume today that the proton and the neutron are not elementary but are made up of three quarks, the proton being the combination uud, the neutron ddu. Here u and d stand for the two most important quark types; u carries the charge $\frac{2}{3}\,e$ and d carries $-\frac{1}{3}\,e$. They represent an isotopic doublet. Thus the transitions 8 and their inverse are pictured today as transitions between the two doublet states:

$$\begin{aligned} d &\rightarrow u + W^- \\ W^- &\rightarrow \begin{cases} e + \bar{\nu}_e \\ \mu + \bar{\nu}_\mu \\ \tau + \bar{\nu}_\tau \end{cases} \\ W^+ &\rightarrow \begin{cases} \bar{e} + \nu_e \\ \bar{\mu} + \nu_\mu \\ \bar{\tau} + \nu_\tau \end{cases} \end{aligned} \qquad (9)$$

The bar denotes the antiparticle. (There is a refinement that we will not treat in any detail. In the fundamental weak interaction process d is replaced by a linear combination $d' = ad + bs$, where s is the so-called strange quark. This refinement allows a weak transition in which the strangeness changes. These effects are smaller than 9 because $b < a$. Similar mixtures between quark types in weak interactions appear

between the higher quark types.)

C. N. Yang and R. L. Mills[27] provided the key idea that was necessary in order to apply field theory to weak and later to strong interactions. It is a generalization of the field concept that underlies QED. In the latter the source of the field is a scalar magnitude, the charge of the particles. The field does not carry any charge; the charge always stays with the particles. Such theories are called "abelian" theories. Nonabelian field theories, as the ones introduced by Yang and Mills, contain two new features:

▶ The source of the field is not a *scalar* charge, but an internal quantum number of the source particle, for example a *spinor* charge, such as the isotopic-spin quantum number (called "up" or "down" in the case of proton and neutron).

▶ The source particle can exchange its "charge" (the isospin) with the field in the interaction process.

In such theories the field itself carries charge and, therefore, acts as a source of fields; there is a direct interaction process between field quanta. Whereas the fundamental diagram of QED is the coupling of the charged particle with the field (see figure 3a) the nonabelian theories also contain another fundamental diagram denoting the coupling between field quanta. The mathematical formulation of nonabelian field theories is based upon a generalization of gauge invariance; we will not enter here into these formal, though essential, arguments, except by noting that they require the field quanta to be massless vector bosons.

To come closer to an understanding of the present view regarding electro-weak dynamics, we start by discussing the theory at very high energies, much higher than the mass of the W, that is much higher than 100 GeV. In that region the weak interactions and the electric interactions are neatly separated. Let us first discuss the former ones. We introduce the so-called weak isodoublets, consisting of the u-d quark pair (actually u − d′; see parenthetical remark on page 80), and the three neutrino–electron pairs:

Doublet (left-handed)	u	ν_e	ν_μ	ν_τ
	d	e	μ	τ
Hypercharge	η'	η	η	η

Only the left-handed particles form these isodoublets. The right-handed ones have no weak interactions. These doublets emit or absorb three types of bosons according to the scheme:

$$\begin{aligned} a &\rightleftarrows b + W^+ \\ b &\rightleftarrows a + W^- \\ a &\rightleftarrows a + W^0 \\ b &\rightleftarrows b + W^0 \end{aligned} \qquad (10)$$

Here a − b stands for any isodoublet of the table above; the coupling constant for each process is g. The process corresponds to the diagram of figure 3a with a coupling constant g. The basic gauge invariance of this formalism requires that the three processes 8 have the same probabilities and that the three W's are massless vector bosons.

In addition to the "SU(2)-type" couplings of equation 10 we also introduce a "hyper-electromagnetic" coupling. It is analogous to the ordinary electromagnetic one ("U(1) coupling"), but the two members a and b carry the same scalar "hypercharge" η' or η, depending on whether we consider the quark pair or the lepton pairs. This coupling does *not* distinguish right- and left-handed particles; it applies to both. We therefore get the processes (with coupling constants η' or η)

$$\begin{aligned} a &\rightleftarrows a + B^0 \\ b &\rightleftarrows b + B^0 \end{aligned} \qquad (11)$$

where B^0 is the massless quantum (vector boson) of the hyper-electromagnetic field. At very high energies we then expect the quarks and leptons to be coupled to the W field in a nonabelian way because, according to equation 10 the iso-spinor charges are transferred to the field and vice versa; but they are coupled to the B field in an abelian way via the scalar hypercharge η or η'.

This picture can be right only at very high energies. The mass of the W would show up at a lower energy. We also find there that the electromagnetic field is coupled to different charges in each isodoublet. How does Nature achieve these deviations from the symmetric theory at high energies? The current theories postulate something that is called "spontaneous symmetry breaking" at lower energies. It is caused by a new isotopic spinor field—the Higgs field. It has the following remarkable property: Its energy is such that it has a minimum not when the field is zero but when it has a finite value given by the spinor $\{\phi_0,0\}$. That would mean that the vacuum has a certain fixed direction in isospace, namely the direction of the spinor ϕ_0. At high energy this is no longer true because there the energy gained by choosing ϕ_0 instead of zero is negligible. The situation is like that of a ferromagnet, in which a direction in real space is determined as long as the energy transfers are smaller than the Curie energy. Thus at low energies the Higgs field destroys the symmetric situation described before. The effects of this destruction by the finite expectation value of the Higgs field are as follows:

▶ The hyper-electromagnetic field **B** and the W^0 field get mixed by an arbitrary mixing angle θ_W, called the Weinberg angle. The two emerging linear combinations are

$$\begin{aligned} \mathbf{Z} &= \cos\theta_W W^0 + \sin\theta_W \mathbf{B} \\ \mathbf{A} &= -\sin\theta_W W^0 + \cos\theta_W \mathbf{B} \end{aligned} \qquad (12)$$

▶ The Higgs field is coupled with the other field in such a way that W^+ and W^- acquire a mass M_W; Z gets a different mass M_Z, whereas the field A remains massless and becomes the electromagnetic field (photons).

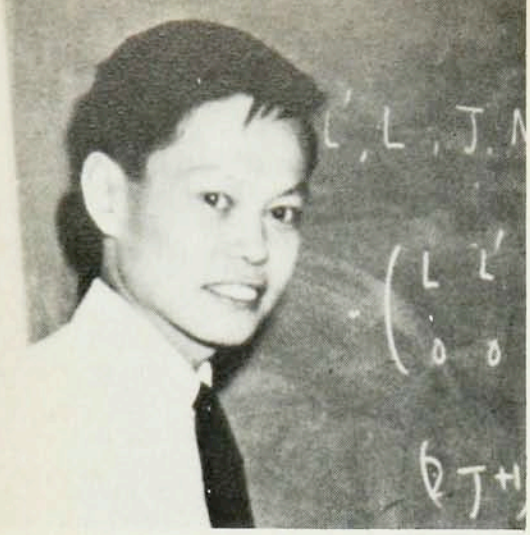
▶ The fact that W^+ and Z have large masses reduces the weak interaction effects compared to the electric ones, at low energies.

▶ The coupling of the quarks and leptons to the electromagnetic field **A** is different from the coupling to the hyper-electromagnetic field **B**. Indeed it is such that the members of an isospin pair acquire the different electric charges, the ones that we usually ascribe to them.

▶ The bosons $W^\pm$ acquire an electric charge $\pm e$ that couples them to the field **A**.

▶ The weak transitions mediated by Z (no charge transfer, "neutral currents") are different from those transmitted by the $W^\pm$. The latter ones are characterized by a maximum parity violation because only the left-handed leptons and quarks are coupled to them. The Z, however, contains not only the W^0, which is coupled to left-handed particles, but also the hyper-electromagnetic field **B** that does not distinguish the handedness in its coupling.

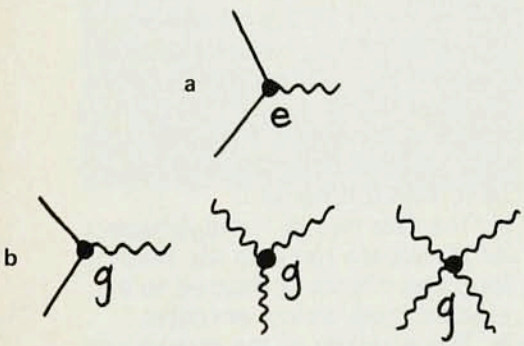
So much for the description of



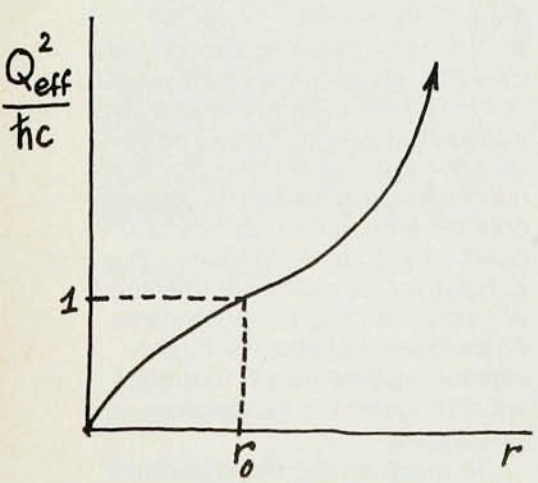
© ALAN W. RICHARDS

C. N. Yang

Fundamental diagrams *(below). (a) shows the fundamental diagram of QED. The straight lines are electron states; the wavy line is a photon state. (b) shows the three fundamental diagrams of QCD. The straight lines are quark states; the wavy lines are gluon states.* *Figure 3*



Running coupling constant *in QCD. The effective "charge"* Q_{eff} *as a function of the distance. The distance* r_0, *where* $Q_{eff} = 1$, *is of the order of the proton radius.* *Figure 4*



quantum electro-weak dynamics. The experiments have borne out the predicted consequences as far as they are accessible to today's experimentation. In particular the mixing of equations 12 could be verified and the angle θ_W determined. Several different experiments lead to the same result: $\sin^2\theta_W = 0.23 \pm 0.02$.

The most important experimental verification is still outstanding: the observation of the intermediate bosons. It is a similar situation to the one of Maxwell's theory of unification of electric and magnetic fields before Hertz's experiments. Woe to the theory if the bosons are not seen when the necessary energy and intensity for their production is reached at some of the accelerators under construction!

A questionable feature of this theory is the introduction of the Higgs field and its somewhat arbitrary couplings with other fields that are adjusted such that they produce the correct masses. The theory also requires the existence of Higgs-field particles of undetermined mass that have not yet been identified. It is hoped that a future formulation of the theory produces the effects of the Higgs field in a more elegant way and gets rid of it, as QED got rid of the vacuum filled with electrons of negative mass!

Quantum chromodynamics

The second theory that was structured as a parallel to quantum electrodynamics was "quantum chromodynamics (QCD)." It deals with the strong interactions. Since the discovery of the quark structure of hadrons one understands by "strong interaction" the forces between quarks. The nuclear force between nucleons was the previous candidate for that name. Today the nuclear force is considered as a weaker derivative of the quark–quark forces, just like the forces between atoms are weaker derivatives of the Coulomb forces between the atomic constituents.

Considering the successes of field-theoretical approaches, it is no surprise that present attempts to describe the interquark forces are also structured according to the model of quantum electrodynamics. Here is a dictionary of the analogies:

QED	QCD
electron	quarks
charge	color
photon	gluons (massless)
positronium	$\rho^0,\omega,\phi,J/\psi,\Upsilon$

Five analogs to positronium exist in QCD because five different types of quarks have been discovered up to now. Actually QED also predicts the existence of two more "positroniums," made of each of the two heavy electrons (μ,τ) and their antiparticles.

There are important differences between these two field theories, which mainly come from the different nature of the charge. In QED the charge is a scalar and remains with the fermions. The field is uncharged. In QCD, what acts as the charge is a "trivalent" magnitude ascribed to the quarks, referred to as "color." It is trivalent in the same sense in which the isotopic spin is a bivalent magnitude.

The color was introduced because three quarks were often found to be in the same quantum state. Because quarks are supposed to obey the Pauli principle, they must possess an internal quantum number capable of assuming three different values. There is a historic parallel to this: The fact that two electrons are found in the ground state of helium has contributed to the discovery of a two-valued internal quantum number—the spin.

QCD assumes that the color is the source of the field. Thus, we again face a nonabelian situation, but here the source is a trivalent "spin," whereas in quantum electro-weak dynamics we had the isotopic doublets of the pairs in the table on page 81 as sources. The consequences of QCD are also derived from a general gauge invariance with respect to the abstract "directions" of the trivalued spin. We obtain again a vector boson field whose massless quanta are the gluons. The properties of this field are analogous to the electromagnetic field. We may use terms such as "gluo-electric" and "gluo-magnetic" fields. There is one essential difference: The fields carry color charge in a similar sense as described in expressions 10. Because now we have three quark colors a, b, c, we find eight different types of gluons, arising from the following emission processes in which the quark colors may change:

$$\begin{array}{ll} a \to b + G_{a\bar{b}} & a \to c + G_{a\bar{c}} \\ b \to c + G_{b\bar{c}} & b \to a + G_{b\bar{a}} \\ c \to a + G_{c\bar{a}} & c \to b + G_{c\bar{b}} \end{array} \quad (13)$$

$$\begin{array}{ll} a \to a + G_0 & a \to a + G_0' \\ b \to b + G_0 & b \to b + G_0' \\ c \to c + G_0 & c \to c + G_0' \end{array} \quad (14)$$

Here $G_{a\bar{b}}$, etc., stands for the emitted gluon that carries double color (a, anti-b). There are eight different gluon colors. The transitions 14 give rise to colorless gluons, but invariance considerations show that there are only two: G_0, G_0', just as there is only one W^0 in equation 10. The fact that the gluons carry color charge leads to the typical nonabelian diagram in figure 3b, which indicates that gluons interact among each other.

A detailed description of QCD goes beyond the aims of this article. It may be important, however, to stress two surprising consequences of this theory, of which the second is not yet established with certainty. The first is called "asymptotic freedom." In contrast to electrodynamics, the effective coupling constant decreases when the distance decreases or when the momentum transfer increases. The coupling decreases as the inverse of the logarithm of the distance and, therefore, vanishes at infinitely close distances. For increasingly larger distances, however, the effective coupling constant does not remain finite as in QED, but seems to increase steadily. Here again we encounter an example of a "running" coupling constant but the dependence of the effective charge Q_{eff} on r is very different from the one in QED that was shown in figure 2. The situation in QCD is sketched in figure 4. The potential energy, say, between a quark and an antiquark (the analog to the Coulomb energy $-e^2/r$ between two opposite charges) probably increases linearly as ar at large distances and goes to infinity for $r \to \infty$.

The consequences of these relations are most unusual. It follows that single quarks cannot exist as free particles. Because the effective charge would become infinite at large distances, the energy necessary to isolate a quark from its partners in a hadron would be infinite. An isolated quark would be surrounded by a field that does not decrease with the distance. Obviously, no isolated quarks (or gluons) can exist in Nature if these conclusions are confirmed. Only systems whose total color charge is zero can exist in isolation. In the spin analogy to color, it would mean that the spins of the constituents must be opposed to each other and form a state of zero spin (singlet). In the trivalent case, three quarks are needed so that their colors add up to zero, or a quark–antiquark pair. Hence hadrons consist of either three quarks or of a quark–antiquark pair, because the antiquark has the complementary color to the quark. (This property justifies the use of the term "color". The three fundamental colors add up to white, and so do a color and its complementary one.)

The fact that hadrons carry no net color charge emphasizes the previously mentioned parallel between the nuclear force and the forces between atoms. Atoms are electrically neutral but when they approach each other, their structure is sufficiently altered that attraction occurs through resonance (Van der Waals forces) or through formation of new quantum states (chemical force). The same would happen when color-neutral nucleons approach each other.

Here we encounter a new situation: The elementary constituents—quarks and gluons—can only exist in bound states, never as single free particles. It should be noted that this paradoxical situation most probably follows (it has not yet been proved beyond a doubt) from a field theory that is a generalization of QED. In the latter, of course, fermions and bosons do exist as free particles; moreover, the system of free particles is the natural limit reached when the coupling constant goes to zero. This limit does not exist in QCD except for very small distances, the opposite situation to that of free particles.

One may ask why a similar situation—the impossibility of isolated particles—does not occur in the case of the weak interaction, which is also a nonabelian field theory. The answer lies in the fact that the symmetry of the isospin space is broken by the Higgs field at low energies (which means low momentum transfers and large distances) whereas the symmetry of the color space does not seem to be broken. Indeed, the mass M of the field quanta (a consequence of the Higgs field) prevents the fields from spreading over distances larger than h/M. Isolated particles do not have infinitely strong fields in QEWD.

Unsolved problems

The development of quantum field theory since its inception half a century ago is most impressive. Today we have the means to calculate electromagnetic effects with incredible accuracy; two new field theories were created that seem reasonably appropriate to deal with the strong and weak interactions, the new forces of nature that were discovered during this half century. These forces are more complicated than the electromagnetic ones and exhibit different properties, such as charge-carrying fields, symmetries broken by vacuum fields, and forever confined particles. The fact that they nevertheless can be described by field theories is an indication that the concepts of those theories play an important role in natural phenomena. Certainly the language of field theory is used by Nature. There exist today attempts to bring together into one unified theory not only the weak and electromagnetic interactions but also the strong ones. These attempts use quantum electroweak dynamics as a model, to bring the SU(2) doublets of the weak forces and the SU(3) triplets of the color variety into one super group with new types of intermediate bosons. They are encouraged by the fact that the strong coupling constant decreases towards higher energies so that one might imagine a very high energy (10^{15} GeV) at which the electro-weak and strong coupling constants merge to one universal parameter. The differing values at lower energies are again caused by symmetry-breaking fields of the Higgs type.

It is by no means clear as to whether these attempts will turn out to be successful or not. In this so-called "grand unification" scheme, the Weinberg angle is no longer arbitrary and seems to come out close to the observed value. It also predicts transitions between quarks and leptons. For example, the u quarks, each having the charge $\frac{2}{3}$, end up as a positron (charge 1) and an anti-d quark (charge $\frac{1}{3}$). Thus a proton (a uud combination) can decay into a π^0 (a $d\bar{d}$ combination) and a

positron. The proton would have a finite lifetime! Such transitions would be very slow because they would be mediated by some of those new intermediate bosons that are supposed to have masses near the characteristic energy of 10^{15} GeV. The lifetime of the proton should be of the order of 10^{32} years. If the numerous ongoing experiments to measure such lifetimes turn out to be successful, the ideas of field theory would win a new victory, and a unification of the three forces of Nature would be in sight. This still would leave gravity alone. The characteristic energy at which quantum effects become important in gravity is given by the mass of the particle pair, whose gravitational potential energy at a distance r is equal to the quantum energy $\hbar c/r$. It is of the order of 10^{19} GeV. This is about 1000 times higher than the characteristic energy of the grand unification attempt.

There are many indications that we understand only a partial aspect of what is going on. Here is an incomplete list of questions that are still unanswered:

▶ Is the renormalization procedure sound? So far it can only be carried out in successive perturbation steps. Can it be applied to a theory with an arbitrarily large coupling constant? The answer to this question may save or condemn field theory. A better understanding of the strong coupling limit (small distances in QED, large distances in QCD) may result in a satisfactory solution to the problems of infinities and of confinement or it may reveal fundamental shortcomings.

▶ The large value of the effective coupling constant of quantum chromodynamics at small momentum transfers causes serious problems as to the nature of the vacuum itself. The field fluctuations may turn out to be very large and may require new conceptions of the nature of the vacuum.

▶ Is the present interpretation of the electro-weak interactions correct? Do the intermediate bosons and the Higgs field really exist? These are questions that will soon be answered by experiments.

▶ The present theories contain arbitrary constants. In QED it is the coupling constant $e^2/\hbar c$ at large distances and the masses of the different electrons. Today three such electrons are known, but there may be more. There is

no way visible at present to explain how their mass values may emerge from the field theories. Moreover, the question remains why there is only one value of the electric charge (the quark charges are simple rational fractions of it) but several mass values seemingly without any simple relations.

In the electro-weak interaction there are two coupling constants between fermions and intermediate bosons, both of the order $e^2/\hbar c$. The Weinberg angle determines the ratio between the two. Furthermore, we find arbitrary coupling constants with the Higgs field that are chosen in order to yield the correct mass for the particles.

In QCD the situation is worse in respect to the mass problem because we deal with many different types of quarks, each having its own mass value. The coupling constant problem, however, is less difficult in QCD, if it turns out for sure that we deal with a running coupling from 0 at very small distances to infinity at large ones. Such a theory does not contain a fixed value at large distance, like $e^2/\hbar c$. But it contains length r_0 (of the order of 10^{-13} cm) at which the running coupling constant is near unity. We expect the composite quark systems to be of that size, and their masses to be of the order $\hbar/r_0 c$, in particular when the masses of the constituent quarks can be negligible compared to that mass. This is indeed the case for those hadrons that are made of u and d quarks. Therefore QCD has the advantage of containing the proton mass as a basic ingredient. (In our description of Nature we expect three intrinsic magnitudes to appear that determine the units of our measuring system. Their values do not require any explanation. These units may well be h, c, and the length r_0 as defined above.) But there is no indication whatsoever how the masses of the heavier quarks are determined by field theory. The theory does not even allow us to hope that the mass problem may be answered by strong coupling effects at small distances. Asymptotic freedom excludes any such effects.

The importance of the mass problem may be illustrated as follows. We have no explanation for the mass of the electron, that is for smallness of the ratio $(1836)^{-1}$ between the electron mass and the proton mass. (The latter may be considered as the natural unit defined by QCD.) The small value of this ratio determines the properties of everything we see around us. It is the precondition of molecular architecture, of the fact that the positions of atomic nuclei are well defined within the surrounding electron clouds. Without it there would be no materials and no life. We have no idea about the deeper reasons for the smallness of that important ratio.

▶ Our present view of elementary particles is plagued by the following problem: Nature as we know it consists almost exclusively of u and d quarks (the constituents of protons and neutrons), and of ordinary electrons; all important interactions are mediated by photons, intermediate bosons and gluons. But there definitely exist higher families of particles, such as the heavier quarks and the heavier electrons. These additional particles are very short-lived or give rise to short-lived hadronic entities. They appear only under very exceptional circumstances that are realized during the early instances of the big bang, perhaps in the center of neutron stars, and at the targets of giant accelerators. What is their role in Nature, why do they exist? Rabi exclaimed when he heard of the first of those "unnecessary" particles, the muon: "Who ordered them?" Again, field theory does not seem to contain the answer to this question. Are they, perhaps, an indication of a deeper internal structure within the quarks and leptons? Are they the excited states of systems made of more elementary units held together by more elementary forces? Will the quantum ladder, the progression from atoms to nuclei, to nucleons and to quarks, ever reach an end?

We will find out sooner or later whether field theory is able to clear up some of these outstanding problems. It may be that a very different approach will be required to solve the questions for which field theory so far has failed to provide answers. Nature's language may be much wider than the language of field theory. We have not yet been able to make sense of much of what Nature says to us.

Looking back over a lifetime of field theory, it seems obvious that we have learned much since 1927, but there is a great deal more that is still shrouded in darkness. New ideas and new experimental facts will be needed to shed more light upon the deeper riddles of the material world.

* * *

Parts of this article appeared in the proceedings of a symposium on the history of particle physics held in May 1980 at Fermilab and also in the 1979 Bernard Gregory Lectures, CERN Report No. 80-03 (1980).

References

1. There exist two interesting studies about this subject: A. Pais, The Early History of the Electron 1897–1947 in Aspects of Quantum Theory, A. Salam, E. Wigner, eds., Cambridge University Press, 1972; S. Weinberg, Notes for a History of Quantum Field Theory, Daedalus, Fall 1977.
2. P. A. M. Dirac, Proc. Roy. Soc. **109**, 642 (1926); **114**, 243 (1927).
3. W. Pauli, Phys. Z. **20**, 457 (1919).
4. A. Einstein, Physica **5**, 330 (1925).
5. A. Pais, Rev. Mod. Phys. **51**, 861 (1979).
6. E. Fermi, Zeits. f. Phys. **29**, 315 (1924).
7. C. V. von Weizsäcker, Z. Phys. **88**, 612 (1934).
8. E. J. Williams, Phys. Rev. **45**, 729 (1934).
9. W. Heisenberg, Z. Phys. **90**, 209 (1934).
10. J. R. Oppenheimer, W. Furry, Phys. Rev. **45**, 245 (1934).
11. W. Pauli, V. F. Weisskopf, Helv. Phys. Acta **7**, 709 (1934).
12. J. R. Oppenheimer, Phys. Rev. **35**, 461 (1930).
13. V. F. Weisskopf, Zeits. f. Phys. **89**, 27; **90**, 817 (1934).
14. F. Bloch, A. Nordsieck, Phys. Rev. **52**, 54 (1937).
15. R. Serber, Phys. Rev. **48**, 49 (1935).
16. E. Uehling, Phys. Rev. **48**, 55 (1935).
17. H. Euler, Ann. d. Phys. V **26**, 398 (1936).
18. V. F. Weisskopf, Kgl. Dansk. Vid. Selsk. **14**, no. 6 (1936).
19. E. C. G. Stueckelberg, Ann. d. Phys. **21**, 367 (1934).
20. E. C. G. Stueckelberg, Helv. Phys. Acta **9**, 225 (1938).
21. W. Lamb, R. Retherford, Phys. Rev. **72**, 241 (1947).
22. H. A. Kramers, Nuovo Cim. **15**, 108 (1938).
23. N. Kroll, W. Lamb, Phys. Rev. **75**, 388 (1949).
24. J. B. French, V. F. Weisskopf, Phys. Rev. **75**, 1240 (1949).
25. O. Klein in *New Theories in Physics*, Conf. Proc. (Warsaw, 1938), Institut International de la Cooperation Intellectuelle, ed., M. Nijhoff, The Hague (1939), page 77.
26. S. Coleman, Science **206**, 1290 (1979).
27. C. N. Yang, R. Mills, Phys. Rev. **96**, 190 (1954). □