96-19

The Evolution of the Gauge Principle

L. O'Raifeartaigh

In 1936 Yukawa made his famous suggestion that the strong nuclear interactions were mediated by mesons. Since then Yukawa's idea has been realized in a more fundamental and universal way than he could have imagined, as *all* the known fundamental interactions are mediated by vector mesons and their self-interactions and interactions with matter are completely determined by group considerations.

The theory which has brought about this remarkable change is gauge theory, which, like gravitation, has a deep geometrical significance. Only in electromagnetism was the existence of a gauge structure obvious from the beginning and even there its physical and geometrical significance were slow to be appreciated. For gravity and the weak and the strong nuclear interactions respectively, the existence of a gauge structure was hidden by the emphasis on the metric, by spontaneous symmetry-breaking, and by quark-gluon confinement. Indeed, as described by the author in a forthcoming book [1], the discovery of gauge theory was a slow and tortuous process. The purpose of this article is to give a brief resume of that process.

Brief Description of Gauge Theory

The basic idea of gauge theory is that a local field theory which is invariant with respect a rigid (space-time independent) continuous symmetry group G, remains invariant when the symmetry group becomes space-time dependent $G \to G(x)$, provided the ordinary derivatives ∂_μ in the Lagrangian are replaced by covariant derivatives $D_\mu = \partial_\mu + A_\mu(x)$, where the A_μ are vector (radiation) fields. This replacement is called the gauge principle and the derivatives are called covariant because they transform covariantly with respect to $G(x)$ i.e. $D_\mu \to g^{-1}(x)D_\mu g(x)$ where $g(x) \in G$. The radiation fields (gauge potentials) A_μ are the fields that realize the Yukawa proposal and in practice they are the well-known electromagnetic and gravitational potentials, and the gluon and $W^\pm_\mu$, Z^o fields of the strong and electroweak interactions respectively. The gauge potentials A_μ are not themselves covariant but the fields strengths $F_{\mu\nu} = [D_\mu, D_\nu]$ constructed from them are. Accordingly, the kinetic energy density for the radiation fields is constructed from the $F_{\mu\nu}$ and, except for gravity (to be discussed later) takes the simple form tr $(F_{\mu\nu}F^{\mu\nu})$.

Electromagnetic Gauge Principle in Classical Physics

Gauge theory first came to light in classical electromagnetism when it was realized that, in contrast to the Newtonian gravitational and electrical forces, the magnetic force was not the gradient of a scalar potential ϕ but the curl of a vector-potential $\vec{A}$, a result whose relativistic generalization is

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \,. \qquad (1)$$

This equation defines the gauge-potential A_μ only up to gauge transformations of the form $A_\mu \to A_\mu + \partial_\mu \alpha$ where $\alpha(x)$ is any differentiable scalar function. The *gauge principle* in classical electrodynamics is the statement that in the presence of an electromagnetic field a particle of electric charge e changes its momentum from p_μ to $p_\mu + eA_\mu$. In particular its relativistic Hamilton-Jacobi equation takes the form

$$(P_\mu + eA_\mu)^2 + m^2 = 0\,, \qquad (2)$$

where m is the mass.

Gravitation and Gauge Theory

Einstein's theory of gravitation was the inspiration for modern gauge theory in two ways. First it inspired the Levi-Civita generalization [2] of Riemannian geometry which, through its development into fibre bundle theory, provided the mathematical structure. Second, it inspired Weyl's attempt [3] to combine electromagnetism and gravitation, which, though unsuccessful in its own right, paved the way for a full understanding of electromagnetic gauge structure and of its non-abelian generalization.

Levi-Civita's observation was that, although Riemannian geometry was metrical, its covariance required only parallel transfer i.e. the existence of a derivative ∇_μ which was covariant with respect to general coordinate transformations. Accordingly a more general geometry could be developed by replacing ∇_μ by a more general covariant derivative $D_\mu = \partial_\mu + \Gamma_\mu$ in which the Christoffel symbol is replaced by a more general connection Γ_μ. The development of Levi-Civita's idea into fibre bundle theory [4] remained almost unknown to physicists and it was only long after the completion of gauge and fibre-bundle theories that their relationship became clear.

The gauge character of gravitation itself was displayed by Weyl [5] using the so-called Vierbein formalism, which he developed to its present form from the original version introduced by Einstein and Wigner to describe distant parallelism and curved space spinors respectively. In terms of the Vierbein $e^a_\mu(x)$ the gravitational gauge potential is defined as

$$A^a_{\mu b}(x) = e^a_\tau(x)\nabla_\mu e^\tau_b(x) \quad \text{where} \quad g_{\mu\nu}(x) = e^a_\mu\ (x)\ e^b_\nu(x)\ \eta_{ab}\ , \qquad (3)$$

where η_{ab} is a flat Minkowski metric. It is easy to verify that A_μ is a space-time vector but that it transforms as connection, i.e. according to

$$A_\mu(x) \to L^{-1}(x)\,(\partial_\mu + A_\mu(x)\,L(x) \qquad (4)$$

with respect to any 'internal' local Lorentz transformations $L(x)$ that leave the Minkowski metric invariant. The main difference

L. O'Raifeartaigh is at Dublin Institute for Advanced Studies, 10, Burlington Road, Dublin 4; E-mail: lor@stp.dias.ie

between gravitation and other gauge theories is that gravitation couples *universally* through the energy momentum tensor and that (because of the metric) the kinetic term may be linear rather than quadratic in the field strength, which is the Riemann tensor.

Electromagnetic Gauge Principle in Quantum Physics

It was only when electromagnetism was introduced into quantum theory that the depth of the article and geometrical significance came to be appreciated. It was actually introduced from two quite different sources, as follows:

The first source was Weyl's unsuccessful attempt [3] to unify electromagnetism and gravity by attaching a non-local electromagnetic scale factor to the gravitational metric i.e. by letting

$$g_{\mu\nu}(x) \rightarrow \hat{g}_{\mu\nu}(x) = \exp\left[\frac{\mathrm{e}}{\mathrm{k}}\int^{x} A_{\tau}(y)\, dy^{\tau}\right] g_{\mu\nu}(x), \quad k = \text{constant} \quad (5)$$

In fact the word gauge (German Eich) originates in the fact that electromagnetism would then re-scale the metric. In an addendum to Weyl's paper Einstein had shown that his idea was untenable because it implied that atomic energy levels would be variable, in contradiction with experiment. But Weyl's idea was later resurrected as follows:

In 1922, Schrödinger [6] observed that in certain circumstances the quantity $e \int A{\cdot}dy$ in (5) was proportional to $\int p.dq$, which is quantized to nh, and thus by a suitable (imaginary) choice of the constant k, namely $i/\hbar$ the Weyl scale factor could become unity. After the invention of wave mechanics London [7] built on this observation to propose that Weyl's scale factor should be changed to a phase factor and applied to the wave function. In other words he proposed that (5) be changed to

$$\psi(x) \rightarrow \exp\left[\frac{ie}{\hbar}\int^{x} A_{\tau}(y)\, dy^{\tau}\right]\psi(x) \qquad (6)$$

The second source was the Hamilton-Jacobi equation (2). By combining the gauge principle $p_\mu \rightarrow p_\mu + eA_\mu$ with the quantum-mechanical correspondence principle $p_\mu \rightarrow \frac{\hbar}{i}\partial_\mu$ Schrödinger [8] converted (2) into the so-called Klein-Gordon equation

$$(D_\mu D_\mu + m^2)\psi(x) = 0 \text{ where } D_\mu = \partial_\mu + \frac{ie}{\hbar}A_\mu \qquad (7)$$

Because (7) did not produce the correct relativistic corrections to hydrogen spectrum, Schrödinger did not have full confidence in it, and used it only as a device for introducing the magnetic field. The spectral defect was, of course, remedied when Dirac applied the quantum-mechanical gauge principle $\partial_\mu \rightarrow D_\mu$ to his *spinor* equation. Indeed the symbiotic relationship between the Dirac equation and the gauge principle constitutes the strongest known support for the validity of the principle. One sees by inspection that the Schrödinger-Dirac formalism is actually the differential version of the London formalism, and the subtle distinction between the two versions was to emerge later in the form of the Aharonov-Bohm effect.

A bonus of both approaches was that D_μ turned out to be the kind of covariant derivative used in fibre bundle theory and thus *the quantum-mechanical gauge principal acquired a geometrical meaning*. The reaction of Weyl to these developments was enthusiastic. Indeed he not only adopted the idea but in [5] went a step further and proposed that gauge theory be regarded as a *principle* for determining interactions rather than a simple symmetry. In the case of electromagnetism and gravitation this proposal was something of a luxury since these theories already existed but for the nuclear interactions it turned out to be indispensable.

The Nuclear Interactions: Non-Abelian Gauge Theory

For many years the nuclear forces seemed to have no direct connection with gauge theory, although isospin symmetry suggested that there might be some connection. As is now well-known, isospin gauge theory was successfully formulated by Yang and Mills [9] in the 1954 paper which is rightly regarded as the seminal paper on nonabelian gauge theory. What is perhaps not so well-known is that this paper was the result of earlier thinking by Yang based on Weyl's ideas and that the Y-M structure was arrived at independently by other physicists, motivated in different ways. Here a brief sketch of the various approaches will be given.

Klein's 1938 Premonition

The first relevant work [10] was that of Oscar Klein. Inspired by Yukawa's proposal and by isospin Klein generalized the Kaluza-Klein (KK) 5-dimensional theory of electromagnetism and gravitation by making the (unorthodox) assumption that the $g_{\mu 5}$ components of the metric tensor formed a 2 × 2 matrix,

$$g_{\mu 5}(x) = \beta\begin{pmatrix} A_\mu(x) & B_\mu(x) \\ \bar{B}_\mu(x) & A_\mu(x) \end{pmatrix}, \qquad \beta = \text{constant} \qquad (8)$$

with $A_\mu(x)$ identified as the electromagnetic potential and the $B_\mu(x)$ fields as Yukawa mesons. He also allowed the charge fields to have an exponential dependence on x_5, identifying the electric charge as ∂_5. The surprising feature was with these assumptions, the usual 5-dimensional Einstein theory decomposed into what we would now call an $SU(2)$ Y-M theory in a gravitational background. The non-linear parts of the B fields were produced by the electromagnetic covariant derivatives $D_\mu B$ and the non-linear part of the A-field by the commutator $[g_{\mu 5}, \partial_5 g_{\mu 5}]$. Klein himself does not seem to have been aware of the $SU(2)$ gauge structure. Indeed he proceeded to assign masses to the B-fields, which violated $SU(2)$ gauge invariance but not his own assumptions.

A serious objection to Klein's theory (raised by Møller) was that, with no neutral B-field, it violated charge independence. In a remarkable reply Klein pointed out that this defect could be remedied by using *four* gauge field in (8), which amounts to extending the gauge group from $SU(2)$ to $SU(2) + U(1)$! He also suggested that a similar structure might be valid for the weak interactions!

Pauli's Dimensional Reduction

By the early fifties the pi-meson and many other short-lived particles had been found and the need for a fundamental theory to take account of them became acute. Inspired by a talk on the subject given by Pais in 1953 Pauli constructed [11] a differential-geometric theory of the strong interactions to *see how it would look*, as he said. Pauli's idea was to generalize the KK theory more systematically, by assuming that the extra dimensions formed a 2-dimensional sphere rather than a circle and by identifying the gauge fields with components of the Christoffel connection rather than the metric. In fact Pauli's approach was the prototype of modern [12] dimensional reduction. The advantage of his approach is that it produces the field strengths automatically as part of the Riemann tensor. However, when the fermion spectrum proved unsatisfactory Pauli lost interest and did not publish his results.

Yang-Mills Theory

The next contribution to gauge-theory was that of Yang and Mills. This development is so well-known that it requires little description and it is perhaps better to let Yang [13] describe it.

While a graduate student in Kunming and in Chicago, I had thoroughly studied Pauli's review articles on field theory. I was very much impressed with the idea that charge conservation was related to the invariance of the theory under phase changes, an idea, I later found out, due originally to Weyl. I was even more impressed by the fact that gauge invariance determined all the electromagnetic interactions. While in Chicago I tried to generalize this to isotopic spin interactions by the procedure later written up in [9] equations (1) and (2) [covariant derivative]. Starting from these it was easy to get equation (3) of [9] [transformation law for the gauge potential]. Then I tried to define the field strengths $F_{\mu\nu}$ by $F_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ which was a natural generalization of electromagnetism. This led to a mess, and I had to give up. But the basic idea remained attractive, and I came back to it several times in the next few years, always getting stuck at the same point... As more and more mesons were discovered and all kinds of interactions were being considered, the necessity to have a principle for writing down interactions became more obvious to me. So while at Brookhaven [in the summer of 1953] I returned once more to the idea of generalizing gauge invariance. My office mate was R.L. Mills who was about to finish his Ph.D ... We worked on the problem and eventually produced [9]. We also wrote an Abstract for the April 1954 meeting of the AMS in Washington. Different motivations were emphasized in the two papers. The formal aspect of the work did not take long and was essentially finished by February 1954. But we found that we were unable to conclude what the mass of the gauge particles should be. We toyed with the dimensional argument that, for a pure gauge theory, there is no quantity with the dimension of mass to start with, and therefore a gauge particle must be massless. But we quickly rejected this line of reasoning.

Thus Yang's ideas can be traced to the 1929 paper of Weyl via Pauli's Handbuch article. An interesting footnote [13] is that when Yang presented the Y-M theory at Princeton in February 1954 Pauli objected so much to his statement that he did not yet know the mass of the gauge field, that it was only Oppenheimer's intervention that allowed the talk to proceed.

Shaw's Independent Construction

Meanwhile the Y-M theory was being re-discovered independently by Ronald Shaw [14], presently professor at Hull University but then a post-graduate student at Cambridge. According to Shaw himself [15a] his research started in the summer of 1952 and the part on Y-M theory was completed early in 1954. However, because of the mass problem he presented his results to his supervisor Abdus Salam in a rather dismissive way and it was only when Salam heard of the Y-M paper that he advised Shaw to publish (which he did not).

One of the interests of Shaw's contribution is the motivation. As Shaw [15b] writes *I am absolutely astonished how much gauge fields have come to the fore in recent years. The idea seemed to me at the time as completely obvious: I had been reading (in 1953) some manuscript of Schwinger's in which he introduced the electromagnetic interaction in this way – he used real spinors and so had SO(2), rather than U(1), invariance and the generalization to SU(2) invariance seemed to shout itself out!* Later Shaw [15c] adds*...the idea arose in a flash, directly from reading some preprint of Schwinger's ... At any rate it seemed to me an obvious idea to replace the SO(2) $\simeq$U(1) of electromagnetism by SU(2) (of Kemmer, etc.) isospin and see what would happen. But I was disappointed that Nature (no suitable m = 0 particle) seemed to reject the idea.*

Utiyama's General Theory

In 1956 there appeared in Physical Review a paper by Utiyama [16]. Because of the date this paper is often dismissed as a simple generalization of Y-M theory to arbitrary simple groups. But this is grossly unfair. Utiyama not only arrived at his results quite independently, but included gravity and had completed his paper by April 1954.

So why did the paper not appear until 1956? According to Utiyama himself [17] the reason was the following: In the years up to 1954 his study of the gauge principle in electromagnetism and gravity had convinced him of its universality. Accordingly, on receiving in January 1954 an invitation to spend a sabbatical at the Princeton Institute he set about formulating his ideas and by April had essentially completed the 1956 paper. However, he did not publish it, partly because of the impending sabbatical and partly because he felt that it had been badly received when presented at a workshop in Kyoto that June. (Actually it appears to have engendered a lively discussion, some of it quite favourable. The main criticism was that he was abandoning the Yukawa tradition of deducing the theory from the phenomenology rather than the reverse). It was only when he arrived in Princeton in September 1954 that Utiyama heard about the Y-M paper. The discovery depressed him so much that he did not read it carefully, and only a year later did he realize that Y-M had considered only *SU*(2) and had not considered gravity. At that stage he decided to translate his manuscript and send it for publication.

Utiyama's approach to gauge theory was the most comprehensive of the five considered here. He was the first to realize the universality of the gauge concept and to suggest explicitly that, in spite of appearances, the nuclear interactions might be gauge interactions. He has sometimes been criticised for trying to relate Y-M theory and gravity too closely but in his defence says [18]: *If I am allowed to refer to my work on the general gauge theory, I should like to stress that it is my paper which first showed clearly that the theory of gravitation could fall into the framework of the gauge theory, which would be a first step toward the grand unified theory from the modern viewpoint. Even more importantly, my paper pointed out that fields carrying a fundamental force – either gravity of electromagnetism – must in fact be those termed connections in mathematics, which are now called gauge fields. That the concept of connections is indispensable in establishing a theory of interactions was the basic assertion that I wanted to make.*

Concluding Remarks

The formulation of gauge theory was not, of course, the end of the gauge story. Indeed how the gauge principle was later found to be applicable to both nuclear interactions is another long tale, well beyond the scope of this article. What I hope to have given here is some feeling for how the thinking about gauge principle developed over the years and how original ideas may develop independently.

References

[1] *The Dawning of Gauge-Theory* (Princeton Univ. Press, in print).

[2] C. Levi-Civita, Rend. Circ. Mat. Palermo **42**, 73 (1917).

[3] H. Weyl. Sitz. Ber. Preuss. Akad. Wiss. 465 (1918).

[4] C. Ehresmann, Coll. de Top., Brussels 29 (1950), S. Kobayashi and K. Nomizu, *Foundation of Differential Geometry*, Vol **I** & **II** (Interscience, I, 1963; II, 1969).

[5] H. Weyl, Zeit. f. Physik **56**, 330 (1929).

[6] E. Schrödinger, Zeit. f. Physik **12,** 13 (1922).

[7] F. London, Zeit. f. Physik **42**, 375 (1927).

[8] E. Schrödinger, Ann. d. Physik **79, 80, 81** (1926).

[9] C.N. Yang and R.L. Mills, Phys. Rev. **96**, 191 (1954).

[10] O. Klein, 1938 Kazimierz Conf. on New Theories in Physics, reprinted in 1988 Kazimierz Conf. eds. Z. Ajduk et al. (World Scientific, Singapore, 1989).

[11] W. Pauli, Letters to Pais, to appear in [1].

[12] T. Appelquist et al., *Modern Kaluza-Klein Theories* (Addison-Wesley, 1987).

[13] C.N. Yang, *Selected Papers with Commentary* (Freeman, New York 1983).

[14] R. Shaw, Ph.D. Thesis, Cambridge University (Sept. 1955).

[15] R. Shaw, (a) Letter to Kemmer, August 1982 (b) Letter to Kemmer. May 1982 (c) Letter to Mills, February 1985.

[16] R. Utiyama, Phys. Rev. **101**, 1597 (1956).

[17] R. Utiyama, *Butsurigaku Wa Dokomade Susundaka* (How Far Has Physics Progressed) (Iwanami Shoten, Tokyo, 1983).

[18] R. Utiyama, *Ippan Gauge Ba Ron Josetsu* (Introduction to the General Gauge Field Theory) (Iwanami Shoten, Tokyo, 1987).