CHINESE JOURNAL OF PHYSICS VOL. 30, NO. 7 DECEMBER 1992

Gauge Theory-Past, Present, and Future?

David J. Gross*
Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544, U.S.A.

(Received November 3, 1992)

I review the history of our understanding of gauge symmetry, from Maxwell to Yang, to the use of Yang-Mills theory in constructing the standard model of elementary particle physics. I discuss recent attempts to go beyond the standard model, unify the seperate forces of nature and give an explanation for the origin of gauge symmetry.

I. INTRODUCTION

It is an honor and a privilege to speak at this celebration of the 70th birthday of C. N. Yang, whose work and spirit has had such an important impact on the development of physics in the second half of the twentieth century. I thought it appropriate on this occasion to review our understanding of gauge symmetry. I shall recall how this concept has evolved from Maxwell to Yang, to the use of Yang-Mills theory in constructing the Standard Model of elementary particle physics. The Standard Model of elementary particle physics is one of the major intellectual achievements of the twentieth century. In the late 1960's and early 1970's, decades of path breaking experiments culminated in the emergence of a comprehensive theory of particle physics. This theory identifies the basic fundamental constituents of matter and describes all the forces of nature relevant at accessible energies- the strong, weak and electromagnetic interactions. As reviewed at this meeting by Ting, the theory has been extensively confirmed over the last fifteen years. It appears to provide a complete description of the electromagnetic weak and strong interactions down to distances of order 10^{-17} cm. or smaller. Finally, I shall discuss recent attempts to go beyond the standard model, to unify the separate forces of nature, and in doing so, to give an explanation for the origin of gauge symmetry.

II. THE DEVELOPMENT OF GAUGE THEORY

II-1. Early History

Gauge field theory first appeared in Maxwell's formulation of electro-dynamics in 1864.

[†] Invited paper presented at the International Symposium in Honor of Professor C. N. Yang's 70th Birthday, July 10-11. 1992, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

Maxwell ’ s theory was the first field theory to appear in physics in addition to being the original gauge theory. However the symmetries of this theory were not truly appreciated for many decades. Electromagnetism contained two important symmetries, Lorentz invariance and gauge symmetry. Both went unrecognized. The full understanding of Lorentz invariance required the theory of relativity, a conceptual revolution. It was necessary both to recognize that the symmetry was present in the equations and to realize that this was a symmetry of nature. The full understanding of gauge invariance required the insights of both quantum mechanics and general relativity. Symmetry itself was not appreciated until the end of the nineteenth century. The prominent role that symmetry plays today was only established after the development of quantum mechanics, toward the middle of the twentieth century. The history of gauge invariance, its origin and development has been brilliantly reviewed by Yang. ’ I shall largely follow his exposition.

After Einstein developed his theory of general relativity, in which a dynamical role was given to geometry,.Herman Weyl ” conjectured that perhaps the scale of length would also be dynamical. He imagined a theory in which the scale of length, indeed the scale of all dimensional quantities, would vary from point to point in space and in time. His motivation was to unify gravity and electromagnetism, to find a geometrical origin for electrodynamics. He assumed that a translation in space-time dx^{μ}, would be accompanied by a change of scale or gauge*, 1 → $1 + S_{\mu}(x)dx^{\mu}$. The gauge function $S_{\mu}(x)$ would determine the relative scale of lengths, so that a function with dimension of Length ” would transform as $f(x) \rightarrow f(x) + d[\partial_{\mu} + S_{\mu}(x)]f(x)dx_{\mu}$. The hope was to identify the connection, S_{μ}, with the vector potential of electrodynamics, thus unifying this theory with gravity. This did not work.

In 1927, after the development of quantum mechanics, Fock[3] and London, ’ noticed that the term $p_{\mu} - eA_{\mu}$, when p_{μ} is replaced with $i\partial_{\mu}$, becomes $\partial_{\mu} - ie/hc\, A_{\mu}$, which looked very much like Weyl ’ s change of scale, but with a complex coefficient for the connection. Two years later Weyl[5] completed the discussion, showing how electrodynamics was invariant under the gauge transformation of the gauge field and of the wave function, Ψ, of a charged particle,

$$A_{\mu} \rightarrow A_{\mu} + \partial_{\mu}\alpha\,; \qquad \Psi \rightarrow e^{\frac{ie\alpha}{\hbar c}}\,\Psi \tag{2.1}$$

Gauge invariance was born. Accompanying the translation of charged pnrticlc there is a phase change.

Gauge symmetry, however, played almost no role in QED. It was largely rcgnrdcd as a complication and a technical difficulty that had to bc carefully handled, especially as people were struggling with the quantization of quantum electrodynamics. This is partly due to the difference between local gauge symmetry and ordinary global symmetries of nature.

* This is the origin of the term *gauge invariance*, which survived even after the theory changed completely.

II-2. The Symmetry of Gauge Invariance

There is an essential difference between gauge invariance and global symmetry such as translation or rotational invariance. Global symmetries are symmetries of the laws of nature. They imply that if an observer rotates or translates her experimental apparatus then she will record the same results. Not so for gauge transformations. They do not lead to any new transformations that leave physical measurements unchanged. Why is this?

To understand the difference between gauge invariance let us recall Noether ' s theorem that establishes a relationship between the existence of a symmetry of the action and a conserved charge. Noether ' s theorem states that if the action is invariant under a transformation, $\delta\phi_i$ of the fields ϕ_i (which we shall assume does not change the coordinates of space-time) then there exists a conserved current,

$$J_\mu \equiv \sum_{\phi_i} \frac{\delta \mathcal{L}}{\delta(\partial_\mu \phi_i)} \delta\phi_i \; ; \qquad \partial^\mu J_\mu = 0 \tag{2.2}$$

In other words, the *charge*, Q = $\int d\mathbf{x} J_0(x)$, is time independent, commutes with the Hamiltonian and is the generator of the symmetry transformation,

$$[H, Q] = 0 \; ; \qquad [Q, \phi_i] = i\delta\phi_i . \tag{2.3}$$

Now let us consider a theory which possesses a local, gauge symmetry, say quantum electrodynamics, described by the Lagrangian,

$$\mathcal{L}_{\mathrm{QED}} = \int d^4x \Big\{ -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar\Psi \gamma_\mu (i\partial^\mu - eA^\mu)\Psi \Big\} . \tag{2.4}$$

This action is invariant under the continuous infinity of transformations, $\delta A_\mu = \partial_\mu \alpha$; $\delta\Psi = ie/hc\, \alpha\Psi$, which produces, following Noether, an infinite number of conserved currents,

$$J^\alpha_\nu = F^\mu_\nu \partial_\mu \alpha + \bar\psi \gamma_\nu \psi \alpha . \tag{2.5}$$

Now we can use the equation of motion, $\partial_\mu F^\mu{}_\nu = \bar\psi\gamma_\nu\psi$; to write the currents as,

$$J^\alpha_\nu = \partial_\mu (F^\mu{}_\nu \alpha) \rightharpoonup Q^\alpha = \int d^3x \partial^\mu (F_{\mu 0}\alpha) = \int \cdot d^3x \vec\nabla \; (\vec E \alpha) . \tag{2.6}$$

For constant a, corresponding to the global symmetry, the charge is given by Gauss ' law,

$$Q^{\alpha = const} = \text{Electric charge} = \int d\vec S \;\; \vec E . \tag{2.7}$$

However, for *a* non constant, and vanishing at spatial infinity, all the new charges vanish, $Q^{\alpha(x)}$

= 0. Even though the gauge theory contains an infinite number of new symmetries of the action, and thus an infinite number of new conserved currents, all the new charges are identically zero. Therefore, associated with a local symmetry there are no new symmetries of nature. The point of gauge symmetry is not that it leads to new charges, but rather that it constrains the form of the action. To quote Yang, gauge *symmetry dictates the form of the interaction.*

II-3. Yang-Mills Theory

In the early 1950's there was an explosion of discovery in particle physics. On the other hand theory was in a sorry state. Early attempts at constructing field theories of the nuclear or strong force were unsuccessful. There was enormous arbitrariness in constructing these theories. At this point there appeared the papers of Yang and Mills that introduced non-Abelian gauge theories.[6,7] There were two stated motivations that lay behind this historic paper. First, Yang wanted to find *aprinciple* that would enable him to select a theory and determine the interactions. The principle was that of a local symmetry. A local symmetry is much more in keeping with the lessons of field theory and relativity than a postulated global symmetry, which smells of action at a distance. The second motivation was simply to generalize the local gauge invariance of electrodynamics to the non-Abelian symmetry of isotopic-spin. Isotopic spin was the first symmetry that was evident in the strong interactions. Introduced by Heisenberg and Wigner, isospin was a good global symmetry of the strong interactions-presumably exact as long as the electromagnetic interaction could be ignored*. Yang and Mills asked whether one could one construct a theory of a gauge field coupled to the isotopic-spin current?

"*The conservation of isotopic-spin is identical with the requirement of invariance of all interactions under isotopic-spin rotation. This means* that *when electromagnetic interactions can be neglected the orientation of the isotopic-spin is of no physical significance. As usually conceived, however, this arbitrariness is subject to the following limitation: once one chooses what to call a proton, what a neutron, at one space-time point, one is then not free to make any other choices at other space-time points. It seems that this is not consistent with the localized field concept that underlies the* **usual** physical *theories. In the present paper we wish to explore* **the** *possibility* **of** *requiting all interactions to be invariant under independent rotations of the isotopic-spin at all space-time points.*"

Yang and Mills succeeded in this goal. They showed that to implement this idea one had to introduce a gauge field $B_\mu{}^a$ that was an isotopic-spin vector (which could be represented by the Hermetian matrix $\mathbf{B}_\mu \equiv 1/2\, \sigma_a B_\mu{}^a$, where σ_a are the Pauli matrices) and replace the covariant derivative $\partial_\mu - ieA_\mu$ with $1\partial_\mu - ig\mathbf{B}_\mu$, and replace the gauge transformation $A_\mu \rightarrow A_\mu$

* Ironically, we now understand that isotopic-spin symmetry, as well as $SU(3)$ X $SU(3)$ symmetry, is an accidental symmetry of the strong interactions. It arises because the light quark (up and down quarks) masses are so small, compared to the mass scale of the strong interaction and appears to have no deep significance.

$+\ \partial_\mu \alpha$ with $\mathbf{B}_\mu \rightarrow \mathbf{B}_\mu + \partial_\mu \mathbf{a} - ig[\mathbf{a}, \mathbf{B}_\mu]$. Perhaps the most surprising result was that the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ had to be replaced with a *non-linear* function of the gauge field, $\mathbf{F}_{\mu\nu} = \partial_\mu \mathbf{B}_\nu - \partial_\nu \mathbf{B}_\mu + ig[\mathbf{B}_\mu, B_\nu]$. Thus non-Abelian gauge theory was a non-linear interacting theory even in the absence of matter, since the gauge mesons were themselves charged.

11-4. Non-Integrable Phase Factor

The third phase in the evolution of the gauge field concept was the understanding that the gauge field is more appropriately understood as a non-integrable phase factor (P stands for path ordering),

$$\mathcal{P} e^{\frac{ie}{\hbar c} \int_A^B \mathbf{B}_\mu(\mathbf{x}) dx^\mu}, \tag{2.8}$$

that accompanies a finite translation (from A to B) of a charged object. This had been already stressed by Dirac in 1931,[8] was evident in the Bohm-Aharanov effect (1959),[9] and was forcefully developed by Yang (1974),[10] and Yang and Wu in 1971.[11] As emphasized by Yang the vector potential is an over complete specification of the physics of a gauge theory but the gauge covariant field strength underspecifies the content of a gauge theory. The Bohm-Aharanov effect is the most striking example of this, wherein there exist physical effects on charged particles in a region where the field strength vanishes. The complete and minimal set of variables necessary to capture all the physics are the non-integrable phase factors.

This description is most clear in the lattice formulation of gauge theory, due to Wegner, Wilson and Polyakov.[12] In this formulation the gauge field is introduced as a matrix that connects neighboring sites ($U_\mu(x)$, connecting a matter field at sitex to one at the neighboring site $x + \mu a$) and the action is the trace of the product of these link variables around the lattice plaquettes. Matter variables, in contrast, sit at sites on the lattice.

Gauge theory exemplifies one of the most profound mysteries of nature and the source of our greatest enjoyment in the exploration of nature. As we probe deeper to reveal the microscope simplicity of nature we require deeper and deeper mathematical structures-deep, interesting, beautiful and powerful structures. To quote Dirac,

"It seems to be one of the fundamental features of nature that fundamental physical laws are described in terms of great beauty and power. As time goes on it becomes increasingly evident that the rules that the mathematician finds interesting are the same that Nature has chosen."

It is extremely gratifying to realize that some of the most profound discoveries of math and physics are isomorphic-a realization which has lead in recent years to measured cross fertilization of fields-with great benefit to both. Recently, in the work of Atiyah, Donaldson, Singer, Witten and others these connections have been fruitfully cxplorccl.

III. APPLICATION OF YANG-MILLS THEORY TO THE STANDARD MODEL

The application of gauge theories to particle physics was a long tricky process, long because it was so tricky. The whole process took almost twenty years. Why did it take so long? Part of the reason is that the elucidation of new symmetries is no easy matter. It is not enough to invent a new symmetry one must also explain how the new symmetry is broken or hidden. After all, if it were not hidden or broken then it would be evident-not a new symmetry.

In the case of the electroweak interactions the issue was how to break the gauge invariance. If unbroken the gauge bosons are necessarily massless. The fact that such particles, aside from the photon, do not exist in nature was the major stumbling block for Yang-Mills theory. This problem clearly bothered Yang in 1954. It took much courage to confront Pauli and others, who immediately objected that this prediction was blatantly false. Yang and Mills, however, pointed out that charged massless vector bosons, in a nonlinear theory, had singular interactions for low momentum. Therefore they might not be massless. Although this is true it took many, many years to understand. The germ of the mechanism was already present in the BCS theory of superconductivity, but it awaited the insights of Higgs, Brout, Englert and Kibble[12] to explain to particle theorists how the symmetry of Yang-Mills theory may be apparently broken-yet no massless vector mesons need emerge. After this was understood models of the electro-weak theory soon followed.

The application of Yang-Mills theory to the strong interactions-the original motivation for the theory—was even trickier. The constituents of hadrons (quarks and gluons) as well as the conserved charges (color) were all hidden by confinement. The idea that hadrons might be composed of quarks emerged, in the work of Gell-Mann and Zweig in 1964,[14] from the approximate flavor $SU(3)$ symmetry of the strong interactions. Originally however these were regarded as mathematical devices without physical reality and it was only with the deep inelastic scattering experiments in 1969 that they acquired some dynamical substance. The color degree of freedom also emerged, in a rather confused fashion, from considering the spectrum of hadrons in crude quark models, in the work of Greenberg and Han and Nambu.[15] Finally the attempt to understand the scaling behavior of the deep inelastic scattering cross sections, which pointed to pointlike constituents within hadrons, lead to the discovery of asymptotic freedom[16,17] and the singling out of the color gauge theory of the strong interactions – QCD.[16]

X11-1. Dynamics of gauge fields

One of the most remarkable features of the standard model, which lies at the heart of the dynamics of this theory, is the profound difference in the dynamics of charge renormalization in Abelian and in non-Abelian gauge thcorics.

Charge renormalization is nothing more (certainly in the case of QED) than vacuum polarization. The vacuum or the ground state of a relativistic quantum mechanical system can be thought of as a medium virtual particles. In QED the vacuum contains virtual electron-positron pairs. If a charge, e_0, is put in this medium it polarizes it. Such a medium with virtual electric dipoles will screen the charge and the actual, observable, charge e, will differ from e_0 as e_0/ε, where ε is the dielectric constant. Now ε is frequency dependent (or energy, or distance dependent). To deal with this one can introduce the notion of an effective coupling e(r), which governs the force at a distance r. As r increases there is more medium that screens, thus e(r) decreases with increasing r, and correspondingly increases with decreasing r! The β-function, which is simply minus the derivative of $\log[e(r)]$ with respect to log(r), is therefore positive.

The Higgs mechanism can be regarded as arising from a vacuum that is a perfect dielectric. This is due to the fact that the Higg's field forms a condensate, filling the vacuum with a dielectric medium, that causes the vacuum to be a *perfect dielectric* with $\varepsilon = \infty$. This mechanism is in many ways the least satisfactory and the least well explored part of the standard model. We eagerly await the completion of the SSC which will fully explore the physics of the Higgs mechanism.

The nature of charge normalization is quite different in a non-Abelian gauge theory, instead of charge screening we have anti-screening! The easiest way to understand this is by considering the magnetic screening properties of the vacuum. In a relativistic theory one can calculate the dielectric constant, ε, in terms of the magnetic permeability, μ, since $\varepsilon\mu = 1$ (in units where c = velocity of light = 1). In classical physics all media are diamagnetic. This is because, classically, all magnets arise from electric currents and the response of a system to an applied magnetic field is to set up currents that act to decrease the field (Lenz's law). Thus $\mu < 1$, a situation that corresponds to electric screening or $\varepsilon > 1$. However, in quantum mechanical systems paramagnetism is possible. This is the case in non Abelian gauge theories where the gluons are charged particles of spin one. They behave as permanent color magnetic dipoles that align themselves parallel to an applied external field increasing its magnitude and producing $\mu > 1$. We can therefore regard the anti-screening of the Yang-Mills vacuum as paramagnetism!

QCD is asymptotically free because the anti-screening of the gluons overcomes the screening due to the quarks. The arithmetic works as follows. The contribution to ε (in some units) from a particle of charge q is $-q^2/3$, arising from ordinary dielectric (or diamagnetic) screening. If the particle has spins (and thus a permanent dipole moment γs) it contributes $(\gamma s)^2$ to μ. Thus a spin one gluon (with γ = 2, as in Yang-Mills theory) gives a contribution to μ of

$$\delta\mu = \left(-\frac{1}{3} + 2^2\right) q^2 = \frac{11}{3} q^2 \; ; \tag{3.1}$$

whereas a spin one-half quark contributes,

$$\delta\mu = -\left(-\frac{1}{3} + \left(2 \times \frac{1}{2}\right)^2\right) q^2 = -\frac{2}{3} q^2 \tag{3.2}$$

(the extra minus arises because quarks are fermions). In any case, the upshot is that as long as there are not too many quarks the anti-screening of the gluons wins out over the screening of the quarks.

IV. BEYOND THE STANDARD MODEL

The standard model is extraordinarily successful. Nonetheless we are not satisfied with the present situation. Experimentalists are frustrated with having to merely confirm theory and both theorists and experimentalists long for new discoveries. More importantly the standard model contains the seeds of its own destruction. First, the very success of the standard model prompts us to ask new questions; questions that were unthinkable before.

The progress of physics is measured by the nature of the questions we ask. First we ask *what,* what are the phenomena; then we ask how, how does it work and then finally we ask *why?* In elementary particle physics we are now at this last stage. The standard model certainly deals with the what and how questipns but it is powerless to answer most of the why questions.

Why are there three families of quarks and leptons, why is $m_e/m_\mu \sim 1/280$, why the strange pattern of quark masses and mixing angles, why is the fine structure constant given by a = 1/137, why is the gauge group SU_3 x SU_2 x U_1, why, why, why? These questions relate to the nineteen parameters of the standard model that have to be fixed by experiment. Surely our understanding of nature will not be complete until we understand the origin of these dimensionless numbers.

In addition we are beginning to ask even more profound questions. Why does fermionic matter exist at all? Unlike the gauge bosons of a Yang-Mills theory fermions are not required by symmetry. Supersymmetry, a beautiful extension of ordinary space-time symmetry, offers a possible answer to this question in addition to many other virtues. It is hoped that supersymmetry could be discovered at the next generation of particle accelerators-say the SSC. Finally, one of the most interesting questions is why is nature dominated by gauge interactions?

There is another reason why it is now evident that the standard model is not the final fundamental theory. It is not just that it cannot answer the above questions but we now realize that we do not know, once again, the principle that determines the interactions. Gauge symmetry and renormalizability are not sufficient. Any theory which can be expressed in terms of local fields at low energy (compared to some cutloff, say M_{planck}) and which is gauge symmetric will be described at these energies by

$$\mathcal{L} \approx \mathcal{L}_{\text{standard model}} + O\left(\frac{1}{M^2_{Planck}}\right)(\bar{\psi}\delta\psi)^2 + \cdot \tag{4.1}$$

The extra, non-rcnormalizable terms, are highly suppressed at energies small compared to the cutoff energy. They will give rise to small, but very interesting (proton decay, etc.) effects. However if we ncglcct these effects the Lagrangian is just the most general Lagrangian with dimcnsionless couplings consistent with the gauge symmetries. This is precisely the standard model with its arbitrary parameters. Thus the success of the standard model offers few clues as to the nature of the high energy theory, save for the fact that this theory must contain gauge symmetry. To be more predictive we must know the theory, and the principle that dictates this theory, which is operative at the energy scale A.

Indeed, the standard, conventional way of answering such why questions and of achieving greater unification is to study physics at smaller distances or greater energies. Since we have a good theory of low energy physics, we have a platform from which we can try to extrapolate to higher energies. This effort has been going on for fifteen years or so. Such extrapolations were considered by theorists shortly after the establishment of the standard model. In particular the discovery of asymptotic freedom, which indicated that the strong force decreases as the energy increases, made it possible to imagine that all forces come together at high energy. The most important thing that emerged early in these investigations was the realization that if we are going to unify all the interactions together, the natural distance or energy scale is very far from present day experiments. It is probably at an energy of 10^{16} to 10^{19} GeV, which is close to the energy where gravity becomes a strong force. In other words, the simple extrapolation of what we know seems to imply that nothing fundamentally new will happen until we get way above present energies, until we go to energies so large or distances so small that gravity, otherwise ignorable as a very weak force, becomes important.

The beautiful experiments described by Ting have spectacularly confirmed the standard model. They are so precise that one can test whether this extrapolation really makes sense. This has been looked at very carefully by the experimentalists, who find that the couplings are very close to coming together. In fact there is a very slight discrepancy, which some take as evidence of supersymmetry. In a supersymmetric theory there are additional matter multiplets which affect the screening and anti-screening properties of the vacuum.

If experimentalists can extrapolate to such energies then certainly theorists can. They have been doing so for the last decade or so in an attempt to find an answer to the why questions posed by the standard model. But how are we to extrapolate so far in the absence of experimental clues? This poses a serious problem for the theorists. We face the question of how to discover the truly new physics if the energy at which the new physics shows up is 17 orders of magnitude bigger than present experimental investigation. Usually, we theorists, have had the luxury of being presented by our experimental friends with new discoveries, new paradoxes and new phenomena, which made it easy for us to discover new theories and new explanations. Following Yang we should search for need new symmetries to dictate new theories. But it is not easy to invent new symmetries. It requires discovering new degrees of freedom, as well as new dynamical mechanisms for hiding the symmetry, otherwise the symmetry would not be new, it would bc obvious to everyone, an old symmetry.

Can wc succeed in making this extrapolation? You can easily givc arguments both pro and con? The arguments against success are easy-history teaches us that without direct experimental clues and tests theorists tend to go wrong. In favor we have the fact that we are lucky to possess a good starting point for extrapolation. We have a very comprehensive theory of low energy physics that seems to work very well. It is not easy to extend such a theory without contradiction, so consistency is a guide. We also have the incredible luck of knowing an important bit of Planck mass physics--namely gravity, which turns off like a power of the energy at low energies but fortunately couples to large objects. Indeed, much of the motivation to construct unified theories is based on the desire to combine gravity with the other things that we know from low energies. Gravity is our only direct handle on the Planck mass physics. Finally, we can be lucky.

IV-l. Extra dimensions

One of the most amusing ideas for new symmetries is that of extra dimensions. The idea that there are more than three dimensions (right-left, forward-backward and up-down) is an old idea. It goes back to 1921, shortly after Einstein ' s theory of relativity, when it was invented by a Polish mathematician called Kaluza.[18] In modern language, Kaluza said the following: Einstein tells us that space-time is a dynamical property of the world, so let us imagine that there are actually five space-time dimensions, one more spatial dimension Imagine that the dynamics of space-time are such that this dimension is not a straight infinite line, but rather a small, little circle of size, perhaps 10^{-33} centimeters, the characteristic dimension of gravity. In that case, we would never know that there is an extra dimension; at every point, we could move right-left, forward-backward or up-down, or around the little circle as well. A low energy physicist, who cannot do experiments at very high energies, of order the Planck length, would never see these extra dimensions.

You might then ask, " If you can ' t see them why hypothesize them? " What Kaluza discovered is that there is some effect of gravity in five dimensions that persists even if one of the dimensions is a very small little circle. He discovered that, the momentum of particles in the fifth dimension, which is conserved and quantized in integer units, can be thought of as the electric charge and the remnant of the gravitational forces in the fifth dimension appear to a low energy physicist as electro-magnetic interactions between these charges. This, in fact, was the first attempt to unify electromagnetism and gravity. In this scheme we can see how gauge invariance could arise from a larger symmetry-that of general coordinate invariance of a higher dimensional theory of gravity.

The generalization of this to the non-Abelian case was first made by Oscar Klein in 1938.[19] He tried to construct a theory of the nuclear (weak-strong) force by generalizing the Kaluza-Klein theory in 5 dimensions to 6 dimensions. Assuming Einsteinian gravity in six-dimensional space-time, and assuming that the background space-time had the structure of M_4 x S_2 he showed that this theory would look, to a low energy observer, like four-dimensional general relativity plus SU_2 Yang-Mills theory. Although he derived some of the equations of Yang-Mills theory he did not appreciate gauge symmetry. To get rid of the massless charged vector mesons he added a mass term by hand, thus destroying gauge symmetry.

This idea has been generalized and extended in the last decade with the hope of using it to unify all of the forces of nature, all of which are gauge interactions similar to electromagnetism, together with gravity. Even more, it offers the hope of explaining all other forces as consequences of gravity. To do so, of course, one has to imagine more than five or six dimensions, perhaps ten. But one can imagine with equal ease a world of ten dimensions, with nine spatial dimensions in which six of these are curled up into little circles so that they are unobservable, except for the remnants of gravity which would appear to us as nuclear, weak and electromagnetic interactions.

However attempts to construct unified theories based on higher dimensional gravity, in fact supergravity, in D = 10 or 11 dimensions have failed. This failure is an indication that it is not easy to construct consistent unified theories.

V. STRING THEORY

String theory also offers an explanation of the origin of gauge symmetry, but much more. In string theory both general coordinate invariance and gauge invariance arise from string symmetry. String theory is a radically conservative extension of the laws of physics. Particles are replaced by extended one-dimensional objects- strings, but all other principles of physics are left unchanged.

We used to think that the proton was an elementary point like particle and then, we learned that at distances of a Fermi (or 10^{-13} centimeters), it has structure. In fact, it is made out of quarks. At present energies we can only explore these distances, and the quarks look point-like. Many people have wondered whether when we look at shorter distances, we will not see that each quark is made out of three preons or subquarks. But history does not always repeat itself. String theory says that if we look at a quark with a good microscope, that can see distances of 10^{-33} centimeters, we will not see smaller constituents, but rather the quarks will look to us like a little closed string.

I cannot explain here how string theory works in detail, but I would like to emphasize that the way we have constructed string theory is a natural generalization of the way we construct theories of particles. For example, in classical physics particles move, as time evolves, along trajectories that are such so as to have minimal length. In other words, of all possible motion the actual motion is the one for which the path traversed has the smallest possible length. In flat space a particle, if there are no other particles around, will therefore move in a straight line. The dynamics of strings is constructed by generalizing this same principle to extended objects. We say that strings as well, as they evolve in time, move along a trajectory in such a way that the area of the tube they span is as small as possible. Based on that principle one can construct both the classical and the quantum mechanical description of the propagation of strings.

Unlike point particle theories we do not possess until now a more fundamental principle on which to base string theory. We do not know the Hamiltonian or the Lagrangian of string theory. All we have are these rules for constructing the probability amplitudes for the propagation of strings in a semi-classical expansion about some background. Nonetheless, this has already produced some amazing consequences.

At first people studied the modes of vibration of both closed strings and open strings and looked for their properties, i.e. the masses and quantum numbers of the natural vibrations of these strings. The remarkable thing that they discovered was that closed strings always contained a particle that could be identified with the graviton, the quantum of gravity, and that open strings always contained a particle that could be identified with gauge mesons. This came out of the theory without having to be put in by hand. In fact, it was very embarrassing because originally string theory was constructed as a theory of nuclear force. As such there was no room for gravity or electromagnetism. It is only with the revival of string theory in the 1980's, as a unifying theory of everything, that this feature is very welcome. The other remarkable, and originally embarrassing, feature of string theory was that these theories were only consistent if one imagined that space time was 26 dimensional (Later, for the so-called supersymmetric superstrings, the dimension of space time had to be ten). Again, as a theory of the nuclear force this is absurd, but it is quite tolerable in the context of a unified theory of gravity.

The biggest difference between particles and strings appears when we come to interactions, to the forces that exist between particles or strings. We can think about interactions between particles in terms of the trajectories that describe their motion by saying that when two particles (say A and B) meet at the same point they have some probability of turning into a third particle (C), and then that third particle with some probability can turn into two particles (D and E). Thus we have a scattering process, where particle A scatters off particle B to produce particles D and E. The interaction, therefore, is all concentrated at the point where the trajectories meet, a singular point of the diagram, or graph, that describes their space-time evolution. The introduction of such an interaction at a point is an ad hoc and highly non unique procedure, which is one of the reasons there are so many particle theories in the world. The situation is much more appealing in the case of strings.

How do strings interact? We would like to let strings interact locally as well by having two strings come together and when they touch at a point become a third string. We clearly can describe this by the so-called pants diagram. Think of horizontal slices through your pants and you will see that this describes the time history of two strings, coming together and forming a third string. Howevcr there is no particular point, no singular point, where the strings join togcthcr. Unlike thc particle picture there is no point that you can pick out and say " this is where thc interaction took place." Thc surface is completcly smooth. It is essentially because of this (ADD) phenomenon that the dynamics of strings is uniquely dctcrmincd. This is an indication of a very large symmetry of string theory, a symmetry which leads to the natural emergence of gravity and the other gauge interactions of nature.

VI. THE ORIGIN OF GAUGE SYMMETRIES IN STRING THEORY

The emergence of gauge symmetries in string theory is one of the most attractive features of the theory and one of the most important elements of its potential to provide a unified theory of physics. I shall describe, in some detail, how gauge interactions emerge in the *heterotic string,*[20] the most interesting and potentially useful of the various superstring theories.

Previously known string theories are the bosonic theory in 26 dimensions (the Veneziano model) and the fermionic, superstring theory in 10 dimensions (an outgrowth of the Ramond-Neveu-Schwarz string). The heterotic string theory was constructed as a chiral hybrid of these.

Free string theories are constructed by first quantization of an action given by the invariant area of the world sheet swept out by the string, or by its supersymmetric generalization. For the bosonic string the action is

$$S = -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \sqrt{g}\left[\eta_{\alpha\beta} g^{ab}\partial_a x^\alpha \partial_b x^\beta\right]. \tag{6.1}$$

where $x^\alpha(\sigma,\tau)$ labels the space time position of the string, embedded in some D dimensional manifold (a = 1, 2,$\cdots$, D), with σ,τ labeling the world sheet that the string sweeps out. It is possible to construct other above two dimensional sigma-model is conformally invariant. For the moment we take the big space to be flat, so that η_{ab} is the Minkowski metric. This is essentially a choice of vacuum for the quantum string theory and will have to be modified in order to describe the real world where one will be interested in non flat D dimensional manifolds. The reparametrization invariance of the action (in σ,τ) permits one to choose the metric of the world sheet to be conformally flat and to identify the timelike parameter of the world sheet, τ, with, say, light cone time. In this light cone gauge the theory reduces to a two-dimensional free field theory of the physical degrees of freedom-the transverse coordinates of the string, subject to constraints. This procedure is valid however only in the critical dimension of 26 for the bosonic string and 10 for the fermionic string. In other dimensions of space time the existence of conformal anomalies imply that the conformal degree of freedom of the internal metric does not decouple. If it is ignored there is a breakdown of world sheet reparametrization invariance.

In the critical dimension the physical degrees of freedom, being massless two-dimensional fields, can be decomposed into right and left movers, i.e. functions of $\tau-\sigma$ and $\tau+\sigma$ respectivcly. If we consider only closed strings then the right and left movers never mix. This separation is maintained even in the presence of string interactions, as long as we allow only orientable world sheets on which a handedness can be defined. This is because the interactions between closed strings are constructed, order by order in perturbation theory, by simply modifying the topology of the world sheet on which the strings propagate. In terms of the first quantized two-dimensional theory no interaction is thereby introduced; the right and left movers still propagate freely and independently as massless fields. Thus, there is in principle no obstacle to constructing the right and left moving sectors of a closed string in a different fashion, as long as each sector is separately consistent, and together can be regarded as a string embedded in ordinary space-time. This is the idea behind the movers of the fermionic superstring with the left movers of the bosonic string theory of closed and orientable strings, since one can clearly distinguish an orientation on such a string. In some sense the heterotic string is inherently chiral; indeed we do not have the option, present in other closed string theories, of constructing a left-right symmetric theory.

The physical degrees of freedom of the right-moving sector of the fermionic superstring consist of eight transverse coordinates: $x^i(\tau - \mathrm{a})$ $(i = 1, \cdots 8)$; and eight Majorana-Weyl fermionic coordinates: $S^a(\tau - \sigma)$ (a $= 1, \cdots 8$). The physical degrees of freedom of the left-moving sector of the bosonic string consist of 24 transverse coordinates: $x^i(\tau + \mathrm{a})$ and $x^I(\tau + \mathrm{a})$ $(i = 1, \cdots 8, I = 1, 16)$. Together they comprise the physical degrees of freedom of the heterotic string. The eight transverse right and left movers combine with the longitudinal coordinates to describe the position of the string embedded in 10 dimensional space. The extra fermionic and bosonic degrees of freedom parametrize an internal space.

The light cone action that yields the dynamics of these degrees of freedom can be derived from the manifestly covariant action, and one can easily quantize it. The only new feature that enters is the compactification and quantization of the extra 16 left-moving bosonic coordinates. It is this compactification, on a uniquely determined 16 dimensional compact space, that leads to the emergence of Yang-Mills interactions. The extra 16 left moving coordinates of the heterotic string can be viewed as parametrizing an " internal " compact space T. This interpretation should not be taken too literally, in fact one can equally well represent these degrees of freedom by 32 real fermions. The question then arises as to the nature of the internal manifold T. This should be a dynamical question since a string theory is a theory of gravity, thus the choice of a background spacetime is a dynamical issue-the background must be a solution of the string equation of motion. The remarkable feature of the heterotic string is that the space T is completely determined by a special 16 dimensional torus (the maximal torus of $E_8 \times E_8$ or Spin $32/Z_2$). That a torus is a solution is reasonable since a torus is simply a flat space with periodic boundary conditions; however, there are many 16 dimensional tori. What picks out the special torus is the requirement that the coordinates of T are left moving (i.e. functions of $\tau + \mathrm{a}$). Consider the expansion of the internal coordinates $x^I(\tau + \mathrm{a})$

$$x^I(\tau + \sigma) = X^I + P^I\tau + L^I\sigma + oscillators. \tag{6.2}$$

The momentum P^I is quantized (since X^I lives on a compact domain) in units of 1/R (where R is a radius of T). The term $L^I\sigma$ must go around the torus (in some direction) an integer number of times, as σ goes from 0 to π, so that X^I will be a periodic function of σ. Therefore L^I must equal an integer multiple of a radius of T in some direction. A string configuration with non-vanishing L^I represents a soliton, i.e. a string that winds around the torus some number of times, and its winding number is a topological, conserved charge. Since X^I is a function of $\tau + \sigma$ then L^I must equal P^I. This clearly restricts the form of the torus. It obviously means that $R \approx 1$ (in our units this means $1/M_{Planck}$). But the form of T is further constrained, since for a general torus it is not possible to identify the winding numbers, which span a lattice Γ, (T = R_{16}/Γ), with the momenta, P^I, which 1'ie on the lattice Γ^* dual to Γ. In fact the full consistency of the heterotic string theory requires that $\Gamma = \Gamma^*$, i.e. that the lattice defining the torus be self dual. If this is not satisfied then " modular invariance " breaks down. This means that the amplitudes, at the one loop level, develop anomalies which break invariance.

Now, there exist very few self dual lattices of the appropriate type, in fact they only exist in 8N dimensions! In 8 dimensions there is one self dual lattice-Γ_8, the root lattice of E_8. In 16 dimensions there exist two self dual lattices, Γ_8 x Γ_8 and Γ_{16}, where Γ_{16} is the weight lattice of *Spin* $32/Z_2$. Thus modular invariance, i.e. invariance of the two dimensional world sheet under global diffeomorphisms, is responsible for the emergence of these two gauge groups.

Let us now examine the massless particles of the heterotic string. Since the theory contains gravity and since it possesses one supersymmetry (the one that transforms the supersymmetric right movers) it will contain the massless multiplet of ten dimensional, N = 1, supergravity. We would also expect, in analogy to the standard Kaluza-Klein mechanism, that a compactified string theory will contain massless vector mesons associated with the isometries of the compact space.

For T, whose isometry consists of sixteen $U(1)$'s, this would yield the *16* gauge bosons of $U(1)^{16}$. A remarkable feature of closed string theories is that for special choices of the compact space there will exist extra massless gauge bosons, which are massless solitons. These are string configurations that wind around the internal torus and have non-vanishing momenta = winding number = $U(1)^{16}$ charge. These combine with the Kaluza-Klein gauge bosons to fill out the adjoint representation of a simple group whose rank equals the dimension of T. For the two allowed choices of T these produce the gauge bosons of G = $Spin(32)/Z_2$ or E_8 x E_8. Thus the heterotic string produces Yang-Mills theory in a novel fashion. The neutral gauge boson, that generate $U(1)^{16}$, orignate by the Kaluza-Klein mechanism, whereas the charged gauge-bosons are stringy solitons which wind around the compactified internal space.

The heterotic string theory has, by now, been developed to the same stage as other superstring theories. Interactions have been introduced, and shown to preserve the symmetries and consistency of the theory, radiative corrections calculated and shown to be finite. In many ways it now appears as the simplest of all superstring theories. It surely provides a most satisfactory explanation for the emergence of specific gauge interactions and offers much phenomenological promise.

VII. THE PROSPECTS FOR STRING THEORY

There are three significant achievements of string theory. First, string theory is a consistent logical extension of the conceptual structure of physics. Second, it produces a finite and consistent theory of quantum gravity. Finally, it might describe the real world.

There have been very few times in the history of physics where consistent, logical and non-trivial extensions of the framework of physics have been successfully made. The best examples are the theories of relativity, quantum mechanics, and now perhaps string theory. Relativity is an extension of classical physics to the realm where the the velocity of light, c, must be regarded as finite; quantum mechanics is an extension to the realm where Planck's constant, h, is not zero. Perhaps string theory completes the trio of fundamental dimensional constants by extending classical physics to the regime where the Planck mass cannot be taken to be infinite. As I explained above string theory is a conservative extension of the logical framework of physics. It changes nothing save the attempt to base physics on point particles. From then on string theories have been developed in a totally conservative fashion, without relinquishing any of the traditional structure of physics. String theory automatically contains gravity. This is to be expected once the graviton appears in the spectrum of the string, since according to general principles the only consistent interactions of a massless spin two meson must be those of general relativity, at least at low energies. Since strings contain both gravitons and gauge particles they yield a theory which contains and reduces to ordinary Einsteinian gravity and Yang-Mills theory at low energies. More than that, strings, for the first time, provide us with a consistent and well behaved quantum mechanical theory of gravity, thus providing us, at the very least, with a viable model of quantum gravity.

Finally, the reason string theories are so exciting is that they offer attractive, realistic theories of the world. The heterotic string also naturally contains quantum gauge theories of precisely the type that we need to describe low energy physics. It has classical solutions that look very much like the standard four dimensional model, with the correct gauge group and matter multiplets. Of course, the potential of string theory for providing a unique description of the real world is far from being realized. But there is no aspect of the real world that we have so far observed that is not contained in some sense in string theory.

Whcrc wc do we stand in string theory? We have hit upon this theory in an accidental fashion. To date all we really understand are the rules for doing perturbation theory about a given classical background and partial rules for constructing consistent classical backgrounds. Some people think that it might take decades, if not longer, to fully explore the structure of this theory. That wc are far from a full understanding is clear if we notc that one main problem is to arrivc at an elucidation of the logical structure of string theory. We still lack a unifying principle that can guide us. Most of the advances that have taken place so far have occurred almost by accident. Einstein developed relativity by having an idea — the principle of equivalence — and then he constructed his equations. String theory has largely developed in the opposite direction, by discovering mathematical structures and then grope towards the physical concepts. Presumably, enormous advances will be required in order to obtain greater dynamical control and in order to make calculations and testable predictions. What is missing is a deep understanding of the conceptual framework from which the symmetries and properties of the theory emerge. There are many, many hints that the ultimate formulation of the theory will be extraordinarily rich and deep, but most likely it will look very different from our present, rather primitive, understanding.

VIII. CONCLUSIONS

In conclusion I reiterate again that the primary lesson of physics in this century is that the secret of nature is symmetry. The most advanced form of symmetries we have understood are local symmetries-general coordinate invariance and gauge symmetry. In contrast we do not believe that global symmetries are fundamental. Most global symmetries are approximate and even those that, so far, have shown no sign of being broken, like baryon number and perhaps even CPT, are likely to be broken. They seem to be simply accidental features of low energy physics. Gauge symmetry, however is never really broken-it is only hidden by the asymmetric macroscopic state we live in. At high temperature or pressure gauge symmetry will always be restored.

We search now for a synthesis of these two forms of symmetry, a unified theory that contains both as a consequence of a greater and deeper symmetry of which these are the low energy remnants. We have hints that string theory, or its generalizations, many achieve this; but we have a long road to follow.

So what have we learned is that *Symmetry dictates interactions* and *Yang dictates symmetry.*

REFERENCES

* Supported by NSF Grant PHY90-21984

1. C. N. Yang, in *Five Decades of Weak Interactions,* edited by N. Chang, New York Academy of Science, 1977.
2. H. Weyl, in *Raum, Zeit und Materie*, 3rd edit. Springer Verlag. Berlin-Heidelberg. New York 1920.
3. V. Fock, Z. Phys., 39,226, 1927.
4. F. London, Z. Phys., 42,375, 1927.
5. H. Wcyl, Z. Phys., 56, 330, 1929.
6. C. N. Yang and R. Mills, Phys. Rev. 95,631, 1954.

7. C. N. Yang and R. Mills, Phys. Rev. 96, 191, 1954.

8. P. A. M. Dirac, Proc. Roy. Soc. **A133**, 60, 1931.

9. Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485, 1959.

10. C. N. Yang, Phys. Rev. Lett. 33,445, 1974.

11. C. N. Yang and T. T. Wu, Phys. Rev. **D12**, 3845, 1975.

12. F. Wegner, J. Math. Phys. **10**, 2259, 1971; K. Wilson, Phys. Rev. **D10**, 2445, 1974; A. Polyakov.

13. P. Higgs, Phys. Lett. 12, 132 (1964); R. Brout and Englert, Phys. Rev. Lett. 13, 321 (1964); T. Kibble, Phys. Rev. **155**, 1554 (1967).

14. M. Gell-Mann, Phys. Lett. 8, 214 (1964); G. Zweig, CERN Report No. TH401, 4R12 (1964) (unpublished).

15. 0. W. Greenberg, Phys. Rev. Lett. 13,598 (1964); M. Y. Han and Y. Nambu, Phys. Rev. **139B**, 1006 (1965).

16. D. J. Gross and F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973).

17. H. D. Politzer, Phys. Rev. Lett. **30**, 1346 (1973).

18. Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Phys. Math. Klasse, 966, 1921.

19. 0. Klein, in New Theories in Physics International Institute of Intelectual Cooperation, Paris, 1939.

20. D. Gross, J. Harvey, E. Martinec, and R. Rohm, Phys. Rev. Lett. 54, 502 (1985), Nucl. Phys. **B256**, 253 (1985).