CHINESE JOURNAL OF PHYSICS VOL. 32, NO. 6-II DECEMBER 1994

Conceptual Beginnings of Various Symmetries in Twentieth Century Physics†[1]

Chen Ning Yang

Department of Physics and Institute of Theoretical Physics, State University of New York at Stony Brook, Stony Brook, NY 11794, U.S.A.

(Received December 15,1994)

Excerpt from the Introduction by Denys Wilkinson

There is certainly no need for me to introduce Frank Yang, the towering master of symmetries who has the audacity to suggest to Nature what She might be, non-abelian. Nature, somewhat astonished, agree.

PACS. 11.30.–j– Symmetry and conservation laws.
PACS. 11.15.–q– Gauge field theories.

What I would like to discuss with you this afternoon is the conceptual beginnings of various symmetries in 20th century physics. I will write the experimental beginnings in red and theoretical beginnings in black. And you will see both kinds of origins of various symmetries, together with a very complicated and entwined pattern.

The century opened with a great bang. We all know that between 1881 and 1887 the Michelson-Morley experiment became more and more accurate. Those were experiments to detect the existence of the ether. The null result was greatly puzzling to physicists at the end of the last century. And in fact if you look into the ten-volume proceedings of the 1904 St Louis Exposition, which had a special program on the present state of the development of arts and sciences, you will find lectures by such distinguished people like Poincare, Boltzmann, and Rutherford, who was a young man at that time. And many of them expressed the view that physics was in great difficulties because almost everything that formed the fundamental concepts of physics at that time had to be revised. It turned out

† Refereed version of the invited paper presented at the First International Symposium on Symmetries in Subatomic Physics, May 16-18, 1994, Taipei, Taiwan, R.O.C.

[1] This article was edited from the recorded tape of Professor Yang's speech by W-Y. Pauchy Hwang and Keng-Hui Lin.

that revision came very rapidly. Within a few months, the 26-year-old Einstein published his paper on special relativity.

Special relativity emphasizes, in fact is built on, Lorentz symmetry, or Lorentz invariance which is one of the most crucial concepts of 20th century physics. I remember vividly in 1982 there was a conference in Erice, Italy. And Dirac was there. I was there. One day Dirac asked me, "What do you think is the most important contribution of Einstein?" I had thought about this problem before. I quickly answered, "It is general relativity." Dirac said, "Yes, general relativity was a singularly beautiful contribution." But he said, "I would choose special relativity as his most important contribution." What he meant was that special relativity through the introduction of the concept of Lorentz symmetry had exerted much more profound influence on 20th century physics than general relativity. I think that probably all of us agree with this interpretation.

Two years after 1905, Einstein started to work on general relativity, and this we know- from his memoirs which he wrote when he was approximately 70 years old. He told about the origin of his work between 1907 and 1915 which eventually gave rise to general relativity. He said that in about 1907 he thought about how special relativity came about. First, there were many years, almost a hundred years, of esperiments which gave rise to four great laws of electromagnetism. Then in 1865, or thereabouts, Maxwell wrote down the great Maxwell's equations. And then, people understood in 1905 that Maxwell equations have certain beautiful symmetries. Einstein said, he thought in 1907, "why couldn't we reverse this procedure? Why couldn't we start from symmetry and derive equations that are consistent with these symmetries and derive experiments which would be in agreement with those equations? He said that was one of the thinking in his mind in 1907 when he embarked on the difficult journey that eventually resulted in general relativity.

In today's terminology we will say that was "symmetry dictates interaction – phase one". Einstein chose a symmetry, the symmetry, or invariance, of physical laws relative to coordinate transformations. And we all know out of that concept he wrote down the beautiful general relativity equations. And out of that he derived the three fundamental experiments to test general relativity. So here we have, as a red entry nest to the Michelson-Morley esperiment, the three tests of general relativity.

Very soon after that, Einstein insisted that once you have understood gravity in terms of a field theory, namely general relativity, and since there is also another field theory already known at that time, namely electromagnetism, one should unify the two. In fact, that became the fundamental theme of his goal throughout his later life.

Weyl was the first person who took up this task. We shall call that "symmetry dictates interaction – phase two" and that is the beginning of gauge theory. Weyl did this in the year of 1918 in several papers, but they all essentially talked about the same

thing in different ways. His fundamental point was the following. He was of course greatly influenced by Einstein and greatly influenced by Levi-Civita. And he said that, according to general relativity or according to Riemannian geometry, if you displace a vector from point A in space-time to a point B through two different routes, always performing parallel displacement on the way, you would end up at point B in two different directions in general. So he said why not apply the same idea to its length. In other words, if directions by parallel displacement could be different depending on the route, why not propose that length would also be different if you go this route or go in a different route. In other words, he proposed that there are scale changes with displacement. As you displace, the scales keep on changing. That is his fundamental idea. The terminology "gauge" derives from the fact that you are talking about "scale". The Germann word was "Eich" and that got translated in the 1920s into "gauge".

I think we need to go a little bit into the detail of this. What Weyl said was that if there are two neighboring points with the displacement dx^μ, he proposed that one institute a scale change. At the first point the scale is one and at the other point the scale is slightly bigger than one. And he proposed that it be of this form:

$$1 + s_\mu dx^\mu .$$

Now apply this idea to a function f which assumes a value f at the first point. Because f depends on the space-time, it would become

$$f + \frac{\partial f}{\partial x^\mu} dx^\mu ,$$

at the second point. If you apply also the scale change to this function, then you get

$$f + \left(\frac{\partial}{\partial x^\mu} + s_\mu \right) f dx^\mu .$$

to the lowest order of dx^μ. And what you have is an operator – the operator on function f and the operator is of the form

$$\frac{\partial}{\partial x^\mu} + s_\mu .$$

He proposed that the theory should be invariant under such a space-time dependent scale transformation. And then he showed that, because of the invariance, s_μ itself is not a physically observable quantity and it is the curl of s_μ that is observable. But we all know that A_μ is not observable but the curl of A_μ, namely $F_{\mu\nu}$, is observable. Therefore he identified s^μ with A^μ. And that was the origin of Weyl's gauge theory.

That theory ran into difficulty immediately from Einstein about which I will comment later. Weyl was discouraged for some time. But then came quantum mechanics. And in 1927, two years after quantum mechanics, Fock and London pointed out independently that

in quantum mechanics you replace the classical p_μ by $-i\hbar\partial_\mu$ and therefore the classical expression of the operator for a charged particle is

$$\frac{\partial}{\partial x^\mu} - \frac{ie}{\hbar c} A_\mu .$$

[By the way, I have looked into the literature, and not yet found the reference where the expression

$$p_\mu - \frac{e}{c} A_\mu$$

first occurred in classical electrodynamics. I believe it must have occurred around the turn of the century and it must have been either in some paper by Lorentz or by Larmor. But I have not yet found precisely where it first occurred.]

Furthmore London pointed out that in fact Weyl was correct except that his identification of s_μ with A_μ is not correct. s_μ should be identified with the expression $-\&A_\mu$. Now $\frac{e}{\hbar c}$ are just numerical numbers. You can change the scale easily to get rid of them. The important thing which we cannot get rid of is the factor $-i$. So by inserting the factor i into the definition of s_μ or into the identification between s_μ and A_μ, Weyl's gauge theory is in fact exactly correct. It is in fact necessary in quantum mechanics. Now you remember the scale change factor mentioned previously. If you now replace s_μ by $-\frac{ie}{\hbar c} A_\mu$, then of course this expression becomes just a phase factor:

$$1 + s_\mu dx^\mu \rightarrow \exp\left\{ -\frac{ie}{\hbar c} \int A_\mu dx^\mu \right\} .$$

In other words, the scale change of Weyl now becomes a phase change. So Weyl's gauge transformation idea is really a phase transformation idea. So gauge transformation should be phase transformation and gauge theory should be phase theory. But we are stucked with the misnomer of the 1920s, though a phase terminology would be more suggestive and meaningful.

With this development, Weyl came back in 1929 and systematized all these developments inserting the factor i now. And he said that there are two kinds of gauge transformations. "The second kind" when you add a gradient to A_μ:

$$A_\mu \rightarrow A'_\mu = A_\mu + \partial_\mu \alpha .$$

And he called the phase transformation,

$$\psi \rightarrow \psi' = \exp\left(\frac{ie\alpha}{\hbar c} \right) \psi ,$$

"gauge transformation of the first kind". And this summary is the kind of thing which physicists of my generation learned, not so much from Weyl's original papers, but from the

various papers by Pauli in the 1930s and 1940s where my generation of graduate students learned field theory from. By the way this paper of Weyl is also the paper that first proposed the two component theory of neutrino which Weyl did write down and immediately rejected as "non-physical" because it does not conserve parity.

Now I want to make two comments. The first comment is about a little-known paper of 1922 of Schrödinger's. Schrodinger wrote this paper in 1922, which was after Weyl's original paper on gauge theory, but before quantum mechanics. And he looked at the gauge-theory scale factor of Weyl's.

$$1 + s_\mu dx^\mu \rightarrow \exp\left\{7 \int A_\mu dx^\mu\right\}.$$

He put a factor 7 there and he applied this idea to Bohr's old quantum theory. He said this leads to a remarkable property of the old Bohr's orbits. What was the remarkable property? Let's take a circular orbits of Bohr of radius of r. Bohr's quantum condition was this:

$$\oint \boldsymbol{pdq} = nh,$$
$$p2\pi r = nh,$$
$$pr = n\hbar.$$

This is Bohr's condition. Schrodinger then said let's talk about the factor in the exponent of Weyl's scale change.

$$\oint A_\mu dx^\mu = \int \phi dt = \frac{e^2}{r} \cdot \frac{2\pi r}{v} = \frac{2\pi e^2}{v},$$
$$\frac{e^2}{r^2} = m\frac{v^2}{r} \rightarrow \frac{e^2}{V} = mvr = pr;$$
$$\oint A_\mu dx^\mu = 2\pi pr = (2\pi\hbar)n.$$

Thus Weyl's integral is equal to $(2\pi\hbar)$n. So Schrodinger pointed out that in fact the Bohr's classical condition is the same condition as having the Weyl's exponent equal to integral multiples of ***h, a*** fact he called remarkable.

This paper was mostly forgotten. One of the reasons that it was forgotten is because Schrodinger never mentioned this 1922 paper in his great papers of 1926 -- Schrodinger wrote six great papers in 1926 which gave rise to wave mechanics. But if you look at the 1922 paper, he did remark that the factor 7 in this exponent could be imaginary. If he had pursued this idea, he would have invented quantum mechanics in 1922.

This remarkable historical fact was pointed out by Raman and Forman, who are historians of science. They wrote a very interesting article called "Why was it Schrodinger

who invented wave mechanics". And their point was that Schrodinger had written this 1922 paper and therefore he was very much acquainted with the concept of phase around the Bohr's orbits. And that, they said, was how Schrodinger got on to wave mechanics. This theory of the history of wave mechanics proposed by Raman and Forman was later confirmed, because Hanle found a letter written by Schrodinger to Einstein, in which he said in November 1925 that the paper by de Broglie was very much similar to what he wrote in 1922 and he was now exploring further developments in this direction. If you want to read more details about this, I had written a paper in the 1987 Centennial Celebration Volume for Schrodinger and in that paper I explain all these things in detail, with additional comments of mine on why perhaps Schrodinger failed in 1926 to mention his 1922 paper.

The second comment I would like make about this episode of development concerns the following historical fact. The 1918 paper of Weyl was submitted to the Berlin Academy. Before he submitted the paper, he had preliminary exchanges with Einstein, saying that he had now found some way to express electromagnetism, and Einstein encouraged him. When the paper came, when the preprint came to the Berlin Academy, Planck and Nernst who were the editors showed the preprint to Einstein. Einstein looked at it and said that this is wrong. Well... It's very fortunate that Planck and Nernst, unlike the editors of the Physical Review Letters today, did not reject the paper. And they allowed the paper of Weyl to be published and asked Einstein to write a postscript to it. They in fact even asked Weyl to write a post-postscript. And all three were published back to back.

What did Eistein say? Einstein had a very good objection. He said, if Mr. Weyl is right, and if there are two rulers starting at the same point, and you make one ruler go this way, the other go that way, they would expand or shrink differently. So by the time they arrive at this same point, they would not have the same length. So you cannot standardize rulers and therefore there can be no physics. This is of course a devastating argument.

In the post-postscript Weyl hand-wave violently but it is clear that he has no real way to explain away Einstein's objection.

In 1929, as I told you before, the factor i was inserted and Weyl came back and wrote his improtant paper. But nobody to my knowledge came back to Einstein's objection. Neither Einstein nor Weyl, nor Planck, nor anybody had come back to this objection. But let us do ask what happens to Einstein's objection in view of the insertion of i. Well, in view of the insertion of i, the ruler in going from here to there, acquires a phase which is path dependent. But two rulers with different phases still have the same length, so there's no problem any more.

But we could ask the next question, "Is this phase difference detectable ?" To detect a phase difference, you must do some interference experiment, and everybody knows you cannot interfere two rulers, at least not yet. But you can interfere two electrons. SO if you

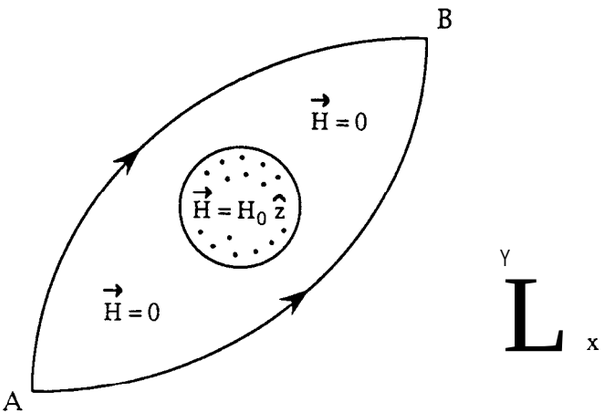

FIG. 1.

have two electrons which go through different paths (Fig. 1), their phase difference can be measured. In particular, if you put a solenoid inside the loop formed by the two paths, you can control that phase, and that is of course the famous Aharonov-Bohm experiment. But Aharonov-Bohm in writing their paper in 1959 did not know about Einstein's objection. Now about this history, I have written an article in the Centennial Celebration of Hermann Weyl, which was published in 1986 by Springer.

Now we change gear to discussions of various different types of symmetries. Before quantum mechanics, there were already known quantum numbers. Quantum numbers in atomic physics were greatly discussed in the first twenty some years of the century. These numbers, n, I, m, spin, and parity, all were esperimentally found because of contacts with the experimental reality. (A red entry marking another experimental beginning!) It was only after quantum mechanics, between 1927 and 1931, through the influence of Weyl and of Wigner, that group theory came into the picture in a big way. In particular SO_3 and SU_2 and the inversion operator were discussed explicitly both by Weyl and by Wigner.

One day I had asked Wigner, "Who first used the word 'parity'?" He couldn't remember and I also did not find in the literature where the word "parity" first occured. But if you look into Condon and Shortley ["The Theory of Atomic Spectra" (Cambridge, London, and MacMillan, New York, 1935)] which was a very important book in atomic physics, it is clear that by 1935 parity was a very important quantum number already.

I believed that the next very important development which makes symmetry considerations something of daily importance to physicists was all these very complicated and beautiful discussions of LS coupling, jj coupling in such books like Condon and Shortley

in early 1930s. These discussions prompted Racah, during the War, to greatly develop the Clebsch-Gordon coefficient type of things related to SU_2 and these Racah-coefficient development later of course became a very important part of comtemporary physics. But when I was a graduate student in the late 1940s, Racah coefficients were just something in the distant horizon. None of the graduate students, my corntemporaries at Chicago, and none of the faculty members at Chicago were really familiar with Racah coefficients.

For that matter when group theory was first introduced in atomic physics, it had been strongly resisted by physicists. Some of you may have heard of the term "the group pest". If you read the introduction to Weyl's book on group theory, you would find that there had been people-who regarded group theory as a pest. Also there was a famous story that when Slater found that you could use determinant wave functions and get rid of the permutation group, it was said that he has slaughtered the group pest.

The origin of the concept of charge conjugation started theoretically with Dirac's article in the proceedings of the Royal Society [Proc. Roy. Soc. A126, 360 (1930)]. This was the article after he had proposed the Dirac equation. And when he was cornered about the negnative energy states he invented the idea of the negnative sea. Oppenheimer and Weyl had both worked on this and eventually it was Kramers [Proc. Acad. Aust. 40, 814 (1937)] and Furry [Phys. Rev. 51, 125 (1937)] in 1937 that formalize the concept of charge conjugation invariance. The experimental verification, or experimental realization, that charge conjugation invariance is for real started with the 1932 discovery of the positron and the 1955 discovery of the antiproton.

Time reversal invariance has a complicated history. It originated with a paper by Kramers in 1930 [Proc. Acad. Aust. 33, 959 (1930)] in which he pointed out that in many atomic systems with various interactions with an odd number of electrons you would have a doubling, a necessary doubling, of each energy level. This was quickly pointed out to be a consequence of time reversal invariance by Wigner [Nachr. Akad. Wiss. Göttingen Math-Phyk. 1932, p. 546]. This is a very subtle and important paper. It is subtle because it was the first place where it was pointed out that the time reversal operator is not a unitary operator, but an anti-unitary operator. This fact became an isolated concept and not generally accepted. If you read the 1941 Reviews of Modern Physics article by Pauli, you will find that Pauli had not yet accepted the Wigner's interpretation of the true meaning of time reversal invariance. Pauli did not have the anti-unitary operator concept for time reversal invariance. Applications of time reversal invarince came in 1951 with the paper by Lloyd [Phys. Rev. **81**, 161 (1951)] which is relatively unknown at that time and still relatively unknown today. But that was the first time where it was pointed out that because of time reversal invariance of the system the relative phase of matrix elements with the same initial state and final state, for example like an E2 transition and an M1 transition, can be

determined. Indeed, as we gather moreexperience, we realize that time reversal invariance is most powerful for determining relative phases.

The CPT theorem has a complicated history, too. Schwinger in his papers in early 1950s had clearly some hunches on the consequences of the CPT theorem, but he never formulated this theorem. It was Lüders who first called it a theorem except that it was not general. And then Pauli in his last important paper in the Neils Bohr Festschrift stated and proved the CPT theorem [in N. Bohr and the Development of Physics (Pergamon, 1955)]. The CPT theorem was later given a more intrinsic proof by Jost in his article [Helv. Phys. Acta. 30, 409 (1957)] when he connected the CPT theorem with analytical continuation in field theories, The real application of the CPT theorem in an intrinsic way was first in the paper by Lee, Oheme and me after the parity nonconservation experiment of 1957, and more explicitly in the analysis of the $K^0-\bar{K}^0$ decays of 1964.

The conservation of isotopic spin was the mathematical scheme first discussed in a kind of formalistic way without having any real reference to the physical world. [Heisenberg, Z. Physik 77, 1 (1932).] Later on in 1936, Breit, Condon, and Present [Phys. Rev. 50, 825 (1936)] analyzed the $p-n$ and $p-p$ interactions and pointed out that they are equivalent in the same states. That led Cassen and Condon [Phys. Rev. 50, 846(1936)] and Wigner [Phys. Rev. 51, 106(1937] to their papers of 1936 and 1937 which formally instituted the concept of isotopic spin conservation.

Now about nonconservation. At this conference I learned for the first time the notation of P-slashed, $\not{P}$. Lee and I wrote a paper in 1956 proposing that parity nonconservation in weak interactions might be an explanation for the $\theta-\tau$ puzzle. To tell the truth, neither of us thought this was likely to be the real explanation because why should Nature not respect a perfectly beautiful symmetry. Immediately after we wrote the paper we got into statistical mechanics and we were deeply involved in very complicated statistical mechanics discussions. If we had realized that it is the right proposal, we certainly would not have taken that long excursion into statistical mechanics. And we all know that within half a year Wu, Ambler, and their collaborators (1957) experimentally found that indeed in beta decay parity was not conserved and thatwas followed within a few days by the Lederman-Garwin experiment which showed beautifully that in the $\pi\rightarrow\mu\rightarrow$ e decay sequence parity was violated twice.

Then, in 1964, we all know that Christenson, Cronin, Fitch, and Turlay showed to the whole physics community by their beautiful experiment, a decisive esperiment, that CP was also not conserved. After the late 1950s, symmetry discussions became a dominant theme in physics. And there were beautiful discussions of SU_3 by Cell-Mann, Neeman, and others, which wasbeautifully confirmed by the finding of Ω^- at Brookhaven.

Now we come to "symmetry dictates interaction – phase three" – non-abelian gauge

fields. Weyl's formulation, or Pauli's review article, has told all physicists that there is something called gauge invariance for electromagnetism. But it was used only as a check whether a calculation is correct. In fact in the late forties all the postdocs knew that after a complicated theoretical presentation, you can ask a very smart question, inquiring if the result is gauge invariant. But that was, in some sense, at that time the only use of it.

Mills and I realized that gauge invariance, when properly generalized, could also be used to determine the field equation itself. We thought it was a very beautiful development. That was 1954. So we wrote it up. If you read our article, you will see that we did not really know how to connect it with reality. And the main problem was with the mass of the gauge field. We knew that there was no charged massless gauge field. And yet we did not know how to put mass into it. In 1971 to 72, the renormalizability of non-abelian gauge fields was proven by 't Hooft and Veltmann. Since non-abelian gauge theories are rather complicated with non-linear terms, the proof of renormalizability was a very difficult task. Another development which came a little bit later was the realization that all gauge fields, abelian or non-abelian, are related to mathematical concepts of fiber bundles. That realization brings into physics topological concepts in field theory, which became very important in later developments.

A most important development was the development of the concept of symmetry breaking. Higgs mechanism is one of them. The idea is that you can still have mathematical symmetry but you no longer have physical symmetry. And that is a beautiful idea. When the symmtry breaking idea was married to non-abelian gauge fields, that produced the Standard Model. And when the idea of confinement came, that produced QCD. So now we have the situation that all fundamental interactions other than gravity are gauge fields.

Is gravity a gauge field ? Everybody would say yes. But precisely how it is a gauge field is unclear. This is both because the gravitational field is a more complicated thing and also because the marriage of gravity with quantum mechanics has not yet been perfected.

In 1973, there was the discovery of the idea of supersymmetry, a beautiful idea which creates additional symmetries between fermions and bosons. Fermions and bosons, both mathematically and physically, look very very different. So when I first was told about supersymmetry, I didn't believe it. I said that maybe they had checked to only the second order, or the third order, and it certainly would break down at some unknown higher order. When I understood it, I realized that indeed there is something beautiful there. However, there is no experimental test so far for supersymmetry. Another beautiful idea was supergravity developed in 1976. And these ideas all are in the direction of exploring additional symmetry ideas in order to help us to resolve the still unanswered questions.

If you look at the history of 20th century physics, you will find that the symmetry concept has emerged as a. most fundamental theme, occupying center stage in today's theoretical physics. We cannot tell what the 21st century will bring to us but I feel safe to say that for the next ten or twenty years many many theoretical physicists will continue to try variations on the fundamental theme of symmetry at the very foundation of our theoretical understanding of the structure of the physical universe.