*Proc. Natl. Acad. Sci. USA*
Vol. 93, pp. 14256–14259, December 1996
Colloquium Paper

***This paper was presented at a colloquium entitled "Symmetries Throughout the Sciences," organized by Ernest M. Henley, held May 11–12, 1996, at the National Academy of Sciences in Irvine, CA.***

# The role of symmetry in fundamental physics

DAVID J. GROSS

Department of Physics, Princeton University, Princeton, NJ 08544

**ABSTRACT** **The role of symmetry in fundamental physics is reviewed.**

Until the 20th century principles of symmetry played little conscious role in theoretical physics. The Greeks and others were fascinated by the symmetries of objects and believed that these would be mirrored in the structure of nature. Even Kepler attempted to impose his notions of symmetry on the motion of the planets. Newton's laws of mechanics embodied symmetry principles, notably the principle of equivalence of inertial frames, or Galilean invariance. These symmetries implied conservation laws. Although these conservation laws, especially those of momentum and energy, were regarded to be of fundamental importance, these were regarded as consequences of the dynamical laws of nature rather than as consequences of the symmetries that underlay these laws. Maxwell's equations, formulated in 1865, embodied both Lorentz invariance and gauge invariance. But these symmetries of electrodynamics were not fully appreciated for over 40 years or more.

This situation changed dramatically in the 20th century beginning with Einstein. Einstein's great advance in 1905 was to put symmetry first, to regard the symmetry principle as the primary feature of nature that constrains the allowable dynamical laws. Thus the transformation properties of the electromagnetic field were not to be derived from Maxwell's equations, as Lorentz did, but rather were consequences of relativistic invariance, and indeed largely dictate the form of Maxwell's equations. This is a profound change of attitude. Lorentz must have felt that Einstein cheated. Einstein recognized the symmetry implicit in Maxwell's equations and elevated it to a symmetry of space-time itself. This was the first instance of the *geometrization* of symmetry. Ten years later this point of view scored a spectacular success with Einstein's construction of general relativity. The principle of equivalence, a principle of local symmetry—the invariance of the laws of nature under local changes of the space-time coordinates—dictated the dynamics of gravity, of space-time itself.

With the development of quantum mechanics in the 1920s symmetry principles came to play an even more fundamental role. In the latter half of the 20th century symmetry has been the most dominant concept in the exploration and formulation of the fundamental laws of physics. Today it serves as a guiding principle in the search for further unification and progress.

## The Meaning of Symmetry

Progress in physics depends on the ability to separate the analysis of a physical phenomenon into two parts. First, there are the initial conditions that are arbitrary, complicated, and unpredictable. Then there are the laws of nature that summarize the regularities that are independent of the initial conditions. The laws are often difficult to discover, since they can be hidden by the irregular initial conditions or by the influence of uncontrollable factors such as gravity friction or thermal fluctuations.

Symmetry principles play an important role with respect to the laws of nature. They summarize the regularities of the laws that are independent of the specific dynamics. Thus invariance principles provide a structure and coherence to the laws of nature just as the laws of nature provide a structure and coherence to the set of events. Indeed, it is hard to imagine that much progress could have been made in deducing the laws of nature without the existence of certain symmetries. The ability to repeat experiments at different places and at different times is based on the invariance of the laws of nature under space-time translations. Without regularities embodied in the laws of physics we would be unable to make sense of physical events; without regularities in the laws of nature we would be unable to discover the laws themselves. Today we realize that symmetry principles are even more powerful—they dictate the form of the laws of nature.

## Classical Symmetries

In classical dynamics the consequences of continuous symmetries are most evident using Hamilton's action principle. According to this principle the classical motion is determined by an extremum principle. Thus if we describe the system by a generalized coordinate $\mathbf{x}(t)$ (for example the position of a point particle in space) then the actual motion of the system, given the values of $\mathbf{x}(t)$ at $t = t_1$ and at $t = t_2$, is such that the action, $S[\mathbf{x}(t)]$, is extremal. The action is a *local* functional of $\mathbf{x}(t)$, namely it can be written as the integral over time of a function of $\mathbf{x}(t)$ and its time derivative—the Lagrangian, $S = \int_{t_1}^{t_2} dt L[\mathbf{x}(t), \mathbf{x}(t)]$. Hamilton's principle means that if $\bar{\mathbf{x}}(t)$ is the actual motion then $S[\bar{\mathbf{x}}(t) + \delta\bar{\mathbf{x}}(t)] = S[\bar{\mathbf{x}}(t)]$ for any infinitesimal variation, $\delta\bar{\mathbf{x}}(t)$, of $\bar{\mathbf{x}}(t)$ that leaves its values at $t = t_1$ and $t = t_2$ unchanged. This is often called the principle of least action, although the extremum can be either a minimum or a maximum of the action. The classical equations of motion for $\bar{\mathbf{x}}(t)$ follow from this principle.

A symmetry of a classical system is a transformation of the dynamical variable $\bar{\mathbf{x}}(t)$, $\bar{\mathbf{x}}(t) \rightarrow \mathcal{R}[\bar{\mathbf{x}}(t)]$, that leaves the action unchanged. If follows that the classical equations of motion are invariant under the symmetry transformation, since if $\bar{\mathbf{x}}(t)$ is an extremum of the action, and $\mathcal{R}$ generates a symmetry of the action, then $\mathcal{R}[\bar{\mathbf{x}}(t)]$ is also an extremum. The symmetry can then be used to derive new solutions. Thus, if the laws of motion are invariant under spatial rotations, then if $\mathbf{x}(t)$ is a solution of the equations of motion, say an orbit of the earth around the sun, then the spatially rotated $\mathbf{x}(t)$, is also a solution. This is interesting and sometimes useful.

A more important implication of symmetry in physics is the existence of conservation laws. For every global continuous symmetry—i.e., a transformation of a physical system that acts

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

the same way everywhere and at all times—there exists an associated time independent quantity: a conserved charge. This connection went unnoticed until 1918, when Emmy Noether proved her famous theorem relating symmetry and conservation laws. Thus due to the invariance of the laws of physics under spatial transformations momentum is conserved, due to time translational invariance energy is conserved and due to the invariance under a change in phase of the wave functions of charged particles electric charge is conserved. It is essential that the symmetry be continuous; namely that it is specified by a set of parameters that can be varied continuously, and that the symmetry transformation can be arbitrarily close to the identity transformation (which does nothing to the system). The discrete symmetries of nature (all of which are approximate symmetries), such as time reversal invariance or mirror reflection, do not lead to new conserved quantities.

One can give a simple geometrical argument that illustrates the connection between symmetry and conservation laws. Consider the motion of a particle described by $\mathbf{x}(t)$, from $\mathbf{x}_i$ to $\mathbf{x}_f$. Assume that the action is invariant under spatial translations. If so the action for the actual path, $S[\mathbf{x}(t)]$ will be equal to the action for the displaced path—i.e., $S[\mathbf{x}(t)] = S[\mathbf{x}(t) + \mathbf{a}]$. Now consider the path of motion from $\mathbf{x_i}$ to $\mathbf{x_i} + \mathbf{a}$ to $\mathbf{x}_f + \mathbf{a}$ to $\mathbf{x}_f$. If $\mathbf{a}$ is very, very small (infinitesimal) then, according to Hamilton's principle, the action along this path is the same as the original action. Thus the difference of the two vanishes, and since the action is additive we have

$$S[\mathbf{x}_i \to \mathbf{x}_i + \mathbf{a} \to \mathbf{x}_f + \mathbf{a} \to \mathbf{x}_f] - S[\mathbf{x}_i \to \mathbf{x}_f] = S[\mathbf{x}_i \to \mathbf{x}_i + \mathbf{a}] + S[\mathbf{x}_f + \mathbf{a} \to \mathbf{x}_f] = 0.$$

The action along the infinitesimal path from $\mathbf{x}_i$ to $\mathbf{x}_i + \mathbf{a}$ must be proportional to $\mathbf{a}$, i.e., $S[\mathbf{x}_i \to \mathbf{x}_i + \mathbf{a}] \equiv \mathbf{p}_i \cdot \mathbf{a}$, which defines the *momentum* $\mathbf{p}$. Similarly the action along the infinitesimal path from $\mathbf{x}_f + \mathbf{a}$ to $\mathbf{x}_f$ is given by $S[\mathbf{x}_f + \mathbf{a} \to \mathbf{x}_f] \equiv -\mathbf{p}_f \cdot \mathbf{a}$ (the minus sign is because the path runs in the opposite direction). Consequently the *momentum* $\mathbf{p}$ is conserved, namely it is a time independent constant along the path of motion.

## Symmetry in Quantum Mechanics

In quantum theory, invariance principles permit even further reaching conclusions than in classical mechanics. In quantum mechanics the state of a physical system is described by a ray in a Hilbert space, $|\Psi\rangle$. A symmetry transformation gives rise to a linear operator, $R$, that acts on these states and transforms them to new states. Just as in classical physics the symmetry can be used to generate new allowed states of the system. However, in quantum mechanics there is a new and powerful twist due to the linearity of the symmetry transformation and the superposition principle. Thus if $|\Psi\rangle$ is an allowed state than so is $R|\Psi\rangle$, where $R$ is the operator in the Hilbert space corresponding to the symmetry transformation $\mathcal{R}$. So far this is similar to classical mechanics. However, we can now superpose these states—i.e., construct a new allowed state: $|\Psi\rangle + R|\Psi\rangle$. (There is no classical analogue for such a superposition; of say the superposition of two orbits of the earth.)

The superposition principle means that we can construct linear combinations of states that transform simply under the symmetry transformations. Thus superimposing all states that are related by rotations we obtain a state $|\Phi\rangle = \Sigma_R R|\Psi\rangle$ that is rotationally invariant, the singlet representation of the rotation group. Namely

$$R|\Phi\rangle = \sum_{R'} RR'|\Psi\rangle = \sum_{R''} R''|\Psi\rangle = |\Phi\rangle.$$

For example the lowest energy state, the ground state of the Hydrogen atom is such a rotational invariant singlet state. Other superpositions of rotated states will yield other irreducible representations of the symmetry group. Indeed any state can be written as a sum of states transforming according to irreducible representations of the symmetry group. These special states can be used to classify all the states of a system possessing symmetries and play a fundamental role in the analysis of such systems. Consequently the theory of representations of continuous and discrete groups plays an important role deducing the consequences of symmetry in quantum mechanics. With the tools of group theory many consequences of symmetry are revealed. For example, the selection rules that govern atomic spectra are simply the consequences of rotational symmetry.

Quantum mechanics also revealed a new kind of symmetry, that of exchange of identical particles. This lead to a classification of all elementary particles as either bosons, whose wave function is invariant under interchange of two identical particles, or fermions, whose wave function changes sign when two identical particles are interchanged. The quantum statistics of such particles is different, with profound implications for their behavior in aggregate.

In relativistic quantum mechanics the implications of symmetry are greater. Here the symmetry group is the Poincaré group, of space-time translations, rotations, and boosts to moving frames. The analysis of the representations of this group leads to a complete classification of physical irreducible representations–elementary particles:

(*i*) Massive representations: $M > 0$. These irreducible representations are labeled by the mass and the spin $J$, which is quantized in half-integer units, $J = 0, 1/2, 1, \ldots$.

(*ii*) Massless representations: $M = 0$. In this case the only finite dimensional representations of this group are one dimensional. These are labeled by a single helicity, $\lambda$, that is half-integer. An example of such a representation is the left-handed neutrino, which only has one helicity state with $\lambda = 1/2$. (If we include parity then irreducible representations contain both positive and negative helicities, $\pm\lambda$.) This group theoretic analysis makes it clear that massless spinning particles are fundamentally different from massive particles. This difference has profound implications for dynamics; indeed, it requires that massless spinning particles be described by gauge theories.

## Symmetry Breaking

The secret of nature is symmetry, but much of the texture of the world is due to mechanisms of symmetry breaking. The are a variety of mechanisms wherein the symmetry of nature can be hidden or broken. The first is explicit symmetry breaking where the dynamics is only approximately symmetric, but the magnitude of the symmetry breaking forces is small, so that one can treat the symmetry violation as a small correction. Such approximate symmetries lead to approximate conservation laws. Many of the symmetries observed in nature are of this sort, not really symmetries of the laws of physics at all, but—for what appears sometimes to be accidental reasons—approximate symmetries for a certain class of phenomena. The isotopic symmetry of the nuclear force is an example of an approximate symmetry; good due to the small values of the up and down quark masses and the weakness of the electromagnetic force.

A more profound way of hiding symmetry is the phenomenon of *spontaneous symmetry breaking*. Here the laws of physics are symmetric but the state of the system is not. This situation is common in classical physics. The earth's orbit is an example of a solution of Newton's equations that is not rotationally invariant, although the equations are. Consequently, for an observer of the solar system, the rotational

invariance of the law of gravitation is not manifest. The particular orbit is picked out by the asymmetric initial conditions of the planet. Thus this mechanism for hiding symmetries of physics is related to the asymmetry induced by asymmetric initial values. In quantum mechanics the situation is different. Quantum mechanical systems with a finite number of degrees of freedom (an adequate description of atomic physics, for example) always have a symmetric ground state. Classically a particle under the influence of gravity can sit anywhere on a flat surface; yet the quantum mechanical state of lowest energy (the ground state) is a superposition of all these classical allowed states. Thus the quantum mechanical particle is everywhere at one—a state that exhibits the translational invariance of the laws of motion.

In systems with an infinite number of degrees of freedom, however, global symmetries may be realized in two different ways. The first way is standard: the laws of physics are invariant and the ground state of the theory is unique and symmetric, as is the case for quantum mechanical systems with a finite number of degrees of freedom. However in systems with an infinite number of degrees of freedom a second mode is possible, in which the ground state is asymmetric. Such spontaneous symmetry breaking is responsible for the existence of crystals (that break translational invariance), magnetism (in which rotational invariance is broken), superconductivity (in which the phase invariance of charged particles is broken), and the structure of the unified electro-weak theory and more. Indeed the spontaneous symmetry breaking of global and local gauge symmetries is a recurrent theme in modern theoretical physics. The search for new symmetries of nature is based on this possibility, for a new symmetry that we discover must be somehow broken otherwise it would have been apparent long ago; it would have been an old symmetry.

Spontaneous broken symmetries have consequences. Although the symmetry is not manifest it has implications. Thus for every broken global symmetry there exist fluctuations with very low energy. These appear as massless particles. Examples are sound waves in solids, spin waves in magnetics and pions in nuclear physics.

Associated with spontaneous symmetry breaking is the phenomenon of symmetry restoration. If one heats a system that possesses a broken symmetry it tends to be restored at high temperature. Thus a ferromagnetic material can be magnetized at low temperature (or even at room temperature) with all the little atomic magnets aligned in the same direction. This is a state of broken rotational symmetry. As the temperature increases the atoms vibrate more and more. Finally when the temperature is greater than a certain critical value the fluctuations win out over the forces that tend to align the atomic magnets and the average magnetization vanishes. Above the critical temperature the system exhibits rotational symmetry. Such a transition from a state of broken symmetry to one where the symmetry is restored is a phase transition. We believe that the same phenomenon occurs in the case of the symmetries of the fundamental forces of nature. Many of these are broken at low temperatures. Very early in the history of the universe, when the temperature was very high, all of these symmetries of nature were presumably restored. The resulting phase transitions, as the universe expanded and cooled, from symmetric states to those of broken symmetry have important cosmological implications.

## Gauge Symmetry

The traditional symmetries discovered in nature were *global* symmetries, transformations of a physical system in a way that is the same everywhere in space. Global symmetries are regularities of the laws of motion but are formulated in terms of physical events; the application of the symmetry transformation yields a different physical situation, but all observations are invariant under the transformation. Thus global rotations rotate the laboratory, including the observer and the physical apparatus, and all observations will remain unchanged. Gauge symmetry is of a totally different nature. Gauge symmetries are formulated only in terms of the laws of nature; the application of the symmetry transformation merely changes our description of the same physical situation, does not lead to a different physical situation.

Gauge symmetry first appeared in Maxwell's electrodynamics. Here the physical observables are the electric and magnetic fields, **E** and **B**. It was discovered early on that one could simplify the equations by introducing a vector potential $A_\mu$, in terms of which both the electric and magnetic fields could be expressed. Yet this description was not unique, one could perform a gauge transformation, $A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu\phi(x)$, without changing the values of **E** and **B.** This symmetry was regarded for decades as rather artificial. As Wigner, one the pioneers of symmetry in this century put it,

> This gauge invariance is, of course, an artificial one, similar to that which we could obtain by introducing into our equations the location of a ghost. The equations must then be invariant with respect to changes of coordinates of that ghost. One does not see, in fact, what good the introduction of the coordinate of the ghost does.

This attitude toward gauge invariance has changed dramatically in the last two decades. Gauge theories have assumed a central position in the fundamental theories of nature. They provide the basis for the extremely successful standard model, a theory of the fundamental, nongravitational forces of nature—the electromagnetic, weak, and strong interactions. To be sure gauge invariance is a symmetry of our description of nature, yet it underlies dynamics. Gauge invariance forces the existence of special particles, gauge bosons. These are massless spin one particles that are associated with the vector potential and mediate the forces. Thus the $SU(3) \times SU(2) \times U(1)$ gauge symmetry of the standard model implies the existence of eight gluons that mediate the strong interaction, three gauge bosons, the $W^{\pm}$ and the $Z$ bosons, that mediate the weak interactions and the photon of light. As Yang has stated: *Symmetry dictates interaction.* The first example of this was general relativity where Einstein employed the symmetry of space-time under local changes of coordinate to determine the laws of gravity.

Furthermore, the realization that gauge symmetry is based on the fiber bundle, a sophisticated geometrical concept, has provided a deep and beautiful geometrical foundation for gauge symmetry. The fiber bundle is a beautiful mathematical construction that combines an internal space together with space-time, to form a unified geometrical object which exhibits the gauge symmetry. It has also been understood, a century after Maxwell discovered his equations, that the vector potential is not just an artificial construct, but has direct observable meaning. This is most evident in the Bohm–Aharonov effect, wherein an electron beam propagates in a region where there is no electromagnetic field, yet due to the geometry the vector potential is nonvanishing. Due to the change of phase that accompanies a charged particle moving in a background vector potential one can observe interference effects directly. Thus the vector potential is primary.

Indeed today we believe that global symmetries are unnatural. They smell of action at a distance. We now suspect that all fundamental symmetries are local gauge symmetries. Global symmetries are either all broken (such as parity, time reversal invariance, and charge symmetry) or approximate (such as isotopic spin invariance) or they are the remnants of spontaneously broken local symmetries. Thus, Poincaré invariance can be regarded as the residual symmetry of the Minkowski vacuum under changes of the coordinates.

## The Origin of Symmetry

Why is nature symmetric? There are at least two views. The first is based on the paradigm of condensed matter systems where unexpected and new symmetries often occur, although they are not present in the fundamental laws. The prime example is the appearance of symmetry in the behavior of long-range fluctuations of a system undergoing a second-order phase transition. Here one has the phenomenon that at the fundamental, short distance or high energy, level there is no symmetry. Rather the symmetry emerges dynamically at large distances.

Could this be the reason for the "fundamental symmetries" that we observe in nature? Could they be dynamical consequences of an asymmetric physics? I believe not. The lesson of the history of physics in this century points to the opposite conclusion. As we explore physics at higher and higher energy, revealing its structure at shorter and shorter distances, we discover more and more symmetry. This symmetry is usually broken or hidden at low energy. I like to think of the first paradigm as *Garbage in—Beauty out*, and the second as *Beauty in—Garbage out*. At the fundamental level nature, for whatever reason, prefers beauty and is marvelously inventive in inventing new forms of beauty. If this is the case then it provides us with an important tool for the exploration of nature. When searching for new and more fundamental laws of nature we should search for new symmetries.

## New Symmetries

Current theoretical exploration in the search for further unification of the forces of nature, including gravity, is largely based on the search for new symmetries of nature. Theorists speculate on larger and larger local symmetries and more intricate patterns of symmetry breaking in order to further unify the separate interactions. Most exciting is the speculation concerning new kinds of symmetry, which could explain some of the most mysterious features of nature. Foremost among these is *supersymmetry* that has the ability to unify bosons and fermions into a single pattern, to unify matter and force, and to help explain the mysterious fact that the mass scale of atomic and nuclear physics is so much smaller than the scale determined by gravity (the hierarchy problem).

Supersymmetry is a profound and beautiful extension of the geometric symmetries of space-time to include symmetries generated by fermionic (anticommuting) charges. We can describe supersymmetry by saying that space-time is to be replaced by super-space-time, which has new coordinates in addition to the usual coordinates of space and time, that we denote by $\theta_i$, $i = 1, 2, \ldots$. The new feature of these coordinates is that they are *anticommuting* numbers, i.e., $\theta_1\theta_2 = -\theta_2\theta_1$. Supersymmetric physics is formulated in this superspace. Thus all fields are function of $\mathbf{x}$ and $t$ and the $\theta_i$ values. Super-symmetry is then a set of continuous transformation, rotations, of all of the coordinates of superspace. These symmetries contain the usual relativistic symmetries of spacetime, but in addition new symmetries, with new consequences. The most important consequence is that for every particle with spin $J$ there must be another with spin $J \pm 1/2$. If the symmetry were exact these would be degenerate in mass. This is not what is observed in nature—thus supersymmetry must be broken. But, as we have learned, this is no problem; most symmetries are broke. If the scale of breaking is the scale of the standard model then this symmetry could explain the hierarchy problem. Furthermore it would then be visible at energies which are just now becoming accessible. We eagerly await the experimental discovery of the signs of this (broken) symmetry at the next generation of particle accelerators—a discovery of new dimensions of space-time.

Finally, in recent years we have begun to seriously explore a new kind of theory based on a radical extension of the conceptual framework of local quantum field theory: string theory. String theory is the most ambitious candidate for a unified theory of all the interactions that naturally embodies in a consistent fashion quantum gravity. It contains within it all the familiar symmetries which we have discovered play a role in nature. It indeed appears to have all the ingredients we need to derive or explain the standard model. In addition, there are hints within the theory that it embodies new and strange symmetries that we are now trying to understand.

Thus, once again we are embarked on a new stage of exploration of fundamental laws of nature, a voyage guided largely by the search for and the discovery of new symmetries.

Supported in part by the National Science Foundation under Grant PHY90-21984.