Symmetries in Physics

Philosophical Reflections

Edited by Katherine Brading, Elena Castellani

Book DOI: http://dx.doi.org/10.1017/CBO9780511535369

Online ISBN: 9780511535369

Hardback ISBN: 9780521821377

Paperback ISBN: 9780521528894

Chapter

7 - The interpretation of gauge symmetry pp. 124-139

Chapter DOI: http://dx.doi.org/10.1017/CBO9780511535369.008

Cambridge University Press

7

The interpretation of gauge symmetry

MICHAEL REDHEAD

1 Introduction

The term 'gauge' refers in its most general everyday connotation to a system of measuring physical quantities, for example by comparing a physical magnitude with a standard or 'unit'. Changing the gauge would then refer to changing the standard. The original idea of a gauge as introduced by Weyl in his (1918) in an attempt to provide a geometrical interpretation of the electromagnetic field was to consider the possibility of changing the standard of 'length' in a four-dimensional generalization of Riemannian geometry in an arbitrary local manner, so that the invariants of the new geometry were specified not just by general coordinate transformations but also by symmetry under conformal rescaling of the metric. The result was, in general, a non-integrability or path dependence of the notion of length which could be identified with the presence of an electromagnetic field. In relativistic terms this meant that, unacceptably, the frequencies of spectral lines would depend on the path of an atom through an electromagnetic field, as was pointed out by Einstein.

With the development of wave mechanics the notion of gauge invariance was revived by Weyl himself (1929) following earlier suggestions by Fock and by London, so as to apply to the non-integrability of the *phase* of the Schrödinger wave function, effectively replacing a scale transformation $e^{\alpha(x)}$ by a phase transformation $e^{i\alpha(x)}$. Invariance under these local phase transformations, referred to as gauge transformations of the second kind (as contrasted with constant global phase transformations of the first kind), necessitated the introduction of an interaction field which could be identified with the electromagnetic potential, a point of view which was particularly stressed by Pauli (1941). The extension of this idea to other sorts of interaction was introduced by Yang and Mills in their article (1954) (although mention should be made of the independent work of Shaw (1954) and the proposals made in an unpublished lecture by Oskar Klein in 1938). The extension to a gauge theory of gravitation was considered by Utiyama (1956). The great advantage of gauge theories was that they offered the possibility of renormalizability, but this was offset by the fact that the interactions described by gauge fields were carried by massless quanta and so seemed inappropriate to the case of the short-range weak and strong interactions of nuclear physics. In the case of the weak interactions this defect was remedied by noticing that renormalizability survived the process of spontaneous symmetry breaking that would generate effective mass for the gauge quanta, while the key to understanding strong interactions as a gauge theory lay in the development of the idea of ‘asymptotic freedom’, expressing roughly the idea that strong interactions were actually weak at very short distances, effectively increasing rather than decreasing with distance.

With this brief historical introduction we turn to consider the fundamental conceptual issues involved in gauge freedom and the closely associated idea of gauge symmetry.

2 The ambiguity of mathematical representation

As we have seen, the term gauge refers in a primitive sense to the measurement of physical magnitudes, i.e. of associating physical magnitudes with mathematical entities such as numbers. Of course the numerical measure is not unique, varying indeed inversely with the magnitude of the unit chosen. Both the unit and the measure can, with some confusion, be referred to as the gauge of the quantity, in everyday parlance.

We now want to generalize this usage by referring to the mathematical representation of any physical structure as a gauge for that structure. By narrowing down this very general definition we shall focus in on more standard definitions of gauge in theoretical physics, such as the gauge freedom of constrained Hamiltonian systems and Yang–Mills gauge symmetries.

But let us start with the most general concept.[1] Consider a physical structure P consisting of a set of physical entities and their relations, and a mathematical structure M consisting of a set of mathematical entities and their relations, which represents P in the sense that M and P share the same abstract structure, i.e. there exists a one–one structure-preserving map between P and M, what mathematicians call an isomorphism. In the old-fashioned statement view of theories, P and M could be regarded as models for an uninterpreted calculus C, as illustrated in figure 1. On the more modern semantic view, theories are of course identified directly with a collection of models such as P. We do not need to take sides in this debate. For our purposes we need merely to note that P does not refer directly to the world, but typically to a ‘stripped-down’, emasculated, idealized version of the world.

[1] The following account leans heavily on Redhead (2001).

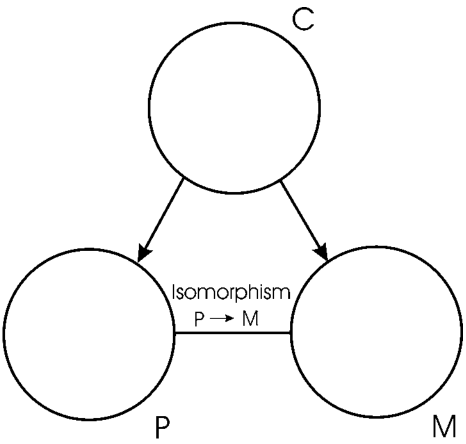

Figure 1. A physical structure P and a mathematical structure M are isomorphic models of an uninterpreted calculus C.

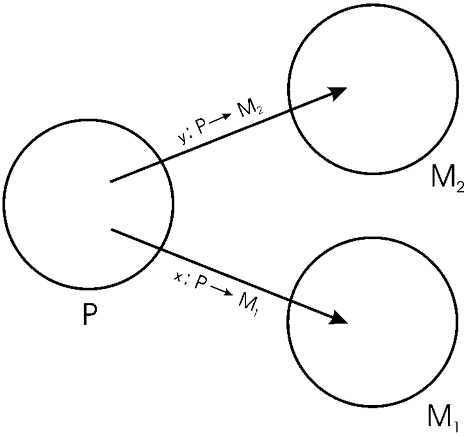

Figure 2. Ambiguity of gauge. M_1 and M_2 are distinct mathematical structures each of which represents P via isomorphisms x and y respectively.

(Only in the case of a genuine Theory of Everything would there be a proposed isomorphism between the *world* and a mathematical structure.)

In our new terminology we shall call M a gauge for P (another way of expressing the relationship between P and M, would be to say that M 'coordinatizes' P in a general sense).

In general there will be many different gauges for P. Consider, as a very elementary example, the ordinal scale provided by Moh's scale of hardness. Minerals are arranged in order of 'scratchability' on a scale of 1 to 10, i.e. the physical structure involved in ordering the hardness of minerals is mapped isomorphically onto the finite segment of the arithmetical ordinals running from 1 to 10. But of course we might just as well have used the ordinals from 2 to 11 or 21 to 30 or whatever. The general situation is sketched in figure 2, which shows two maps x and y which are isomorphisms between P and distinct mathematical structures M_1 and M_2. Of course M_1 and M_2 are also isomorphically related via the map $y \circ x^{-1} : M_1 \to M_2$ and its inverse $x \circ y^{-1} : M_2 \to M_1$.

But how can the conventional choice between M_1 and M_2 as gauges for P have any *physical* significance? To begin to answer this question we introduce the notion of a symmetry of P and its connection with the gauge freedom in the generalized sense we have been discussing.

3 Symmetry

Consider now the case where the ambiguity of representation (the gauge freedom) arises within a *single* mathematical structure M. Thus we consider two distinct isomorphisms $x : P \to M$ and $y : P \to M$, as illustrated in figure 3.

Clearly the composite map $y^{-1} \circ x : P \to P$ is an automorphism of P. This is referred to by a mathematician as a point transformation of P and by physicists as an *active* symmetry of P. The composite map $y \circ x^{-1} : M \to M$ is a 'coordinate' transformation or what physicists call a *passive* symmetry of P. It is easy to show that *every* automorphism of P or M can be factorized in terms of pairs of isomorphic maps between P and M in the way described. It is, of course, not at all surprising that the automorphisms of P and M are themselves in one–one correspondence. After all, since P and M are isomorphically related, they share the same abstract structure, so the structural properties of P represented by the symmetries of P can be simply read off from the corresponding symmetries of M.

Now the symmetries of P express very important structural properties of P, and we can see how they are related to the gauge freedom in this very important special case where the ambiguity of representation is within a *single* mathematical structure M.

The gauge freedom represented in figure 2 does not, in general, have physical repercussions related to symmetry. For example, in the case of Moh's scale of hardness, there simply are no non-trivial automorphisms of a finite ordinal scale.

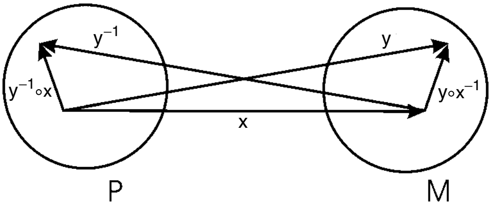

Figure 3. x and y are two distinct isomorphisms between P and M. Then $y^{-1} \circ x : P \to P$ is an automorphism of P and $y \circ x^{-1} : M \to M$ is an associated automorphism of M.

We now want to extend our discussion to a more general situation, which frequently arises in theoretical physics and which we introduce via a notion we call 'surplus structure'.

4 Surplus structure

We consider now the situation where the physical structure P is *embedded* in a larger structure M' by means of an isomorphic map between P and a substructure M of M'. This case is illustrated in figure 4.

The relative complement of M in M' comprises elements of what we shall call the surplus structure in the representation of P by means of M'. Considered as a structure rather than just as a set of elements, the surplus structure involves both relations among the surplus elements and relations between these elements and elements of M.

A simple example of this surplus structure would arise in the familiar use of complex currents and impedances in alternating current theory, where the physical quantities are embedded in the wider mathematical structure of complex numbers.

Another example is the so-called S-matrix theory of the elementary particles that was popular in the 1960s, in which scattering amplitudes considered as functions of real-valued energy and momentum transfer were continued analytically into the complex plane and axioms introduced concerning the location of singularities of these functions in the complex plane were used to set up systems of equations controlling the behaviour of scattering amplitudes considered as functions of the real physical variables. This is an extreme example of the role of surplus structure in formulating a physical theory, where there was no question of identifying any physical correlate with the surplus structure.

In other examples the situation is not so clear. What starts as surplus structure may come to be seen as invested with physical reality. A striking example is the case of energy in nineteenth-century physics. The sum of kinetic and potential energy was originally introduced into mechanics as an auxiliary, purely mathematical entity,

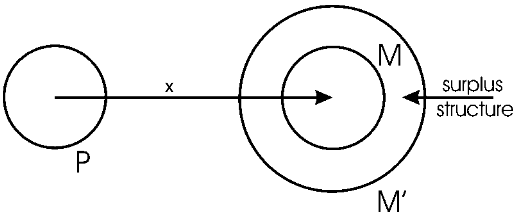

Figure 4. $x : P \to M$ is an embedding of P in the larger structure M'.

http://dx.doi.org/10.1017/CBO9780511535369.008

arising as a first integral of the Newtonian equations of motion for systems subject to conservative forces. But as a result of the formulation of the general principle of the conservation of energy and its incorporation in the science of thermodynamics (the First Law) it came to be regarded as possessing ontological significance in its own right. So the sharp boundary between M and the surplus structure as illustrated in figure 4 may become blurred, with entities in the surplus structure moving over time into M. Another example would be Dirac's hole theory of the positron, allowing a physical interpretation for the negative-energy solutions of the Dirac equation.

Ambiguities in representation, i.e. gauge freedom, can now arise via automorphisms of M' that reduce to the identity on M, i.e. the transformations of representation act non-trivially only on the surplus structure. Nevertheless such transformations can have repercussions in controlling the substructure M and hence the physical structure P. This is the situation that arises in Yang–Mills theories which we shall describe in section 6. But first we shall make a short digression to discuss the example of constrained Hamiltonian systems, of which free-field electromagnetism is a very important special case.

5 Constrained Hamiltonian systems[2]

The idea of surplus structure describes a situation in which the number of degrees of freedom used in the mathematical representation of a physical system exceeds the number of degrees of freedom associated with the physical system itself. A familiar example is the case of a constrained Hamiltonian system in classical mechanics. Here the Legendre transformation from the Lagrangian to the Hamiltonian variables is singular (non-invertible). As a result the Hamiltonian variables are not all independent, but satisfy identities known as constraints. This in turn means that the Hamiltonian equations underdetermine the time-evolution of the Hamiltonian variables, leading to a gauge freedom in the description of the time-evolution, which means in other words a breakdown of determinism for the evolution of the state of the system as specified by the Hamiltonian variables.

More formally the arena for describing a constrained Hamiltonian system is what mathematicians call a *presymplectic manifold*. This is effectively a phase space equipped with a degenerate symplectic two-form ω. By degenerate one means that the equation $\omega(X) = 0$, where X is a tangent vector field, has non-trivial solutions, the integral curves of which we shall refer to as null curves on the phase space. The equations of motion are given in the usual Hamiltonian form as $\omega(X) = dH$, where H is the Hamiltonian function. The integral curves derived from this equation

[2] The treatment of this topic broadly follows the excellent account in Belot (1998).

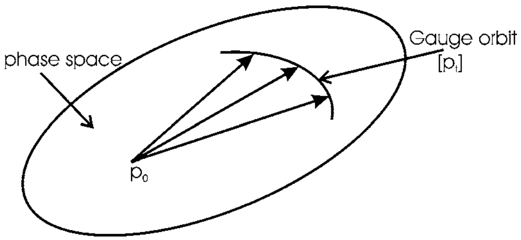

Figure 5. The indeterministic time-evolution of a constrained Hamiltonian system.

represent the dynamical trajectories in the phase space. But in the case we are considering there are many trajectories issuing from some initial point p_0, at time t_0. At a later time t the possible solutions of the Hamiltonian equations all lie on a *gauge orbit* in the phase space which is what we may call a null subspace of the phase space, in the sense that any two points on the orbit can be joined by a null curve as we have defined it. The situation is illustrated schematically in figure 5.

Instead of the initial phase point p_0 developing into a unique state p_t at a later time t as in the case of an unconstrained Hamiltonian system, we now have an indeterministic time-evolution, with a unique p_t replaced by a gauge orbit, which we denote by $[p_t]$ in figure 5. Effectively what is happening here is that the 'physical' degrees of freedom at time t are being multiply represented by points on the gauge orbit $[p_t]$ at time t in terms of the 'unphysical' degrees of freedom.

A familiar example of a constrained Hamiltonian system is the case of electromagnetism described by Maxwell's equations in vacuo. Here the Hamiltonian variables may be taken as the magnetic vector potential $\vec{A}$ and the electric field $\vec{E}$ subject to the constraint $div\ E = 0$. On a gauge orbit $\vec{E}$ is constant but $\vec{A}$ is specified only up to the gradient of a scalar function. The magnetic induction $\vec{B}$ defined by $\vec{B} = curl\,\vec{A}$ is then also gauge-invariant, i.e. constant on a gauge orbit. So $\vec{A}$ involves unphysical degrees of freedom, whose time-evolution is not uniquely determined. It is only for the physical degrees of freedom represented by $\vec{E}$ and $\vec{B}$ that determinism is restored. The gauge freedom in $\vec{A}$ belongs to surplus structure in the terminology of section 4.

6 Yang–Mills gauge theories

We turn now to a still more restricted sense of gauge symmetry associated with Yang–Mills gauge theories of particle interactions. To bring out the main idea we shall consider the simplest case of non-relativistic (first-quantized) Schrödinger field. The field amplitude $\psi(x)$ (for simplicity we consider just one spatial dimension for the time being) is a complex number, but quantities like the charge density

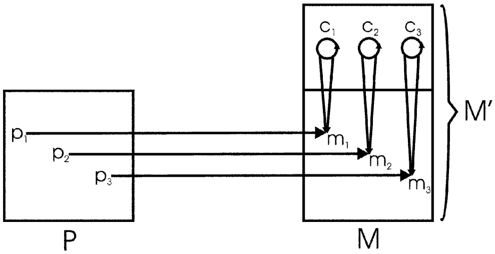

Figure 6. Gauge transformations and surplus structure.

$\phi = e\psi^*\psi$ and the current density $j = \frac{1}{2}ie\big(\psi^*\frac{d}{dx}\psi - \psi\frac{d}{dx}\psi^*\big)$ are real quantities and can represent physical magnitudes. Consider now phase transformations of the form $\psi \to \psi e^{i\alpha}$. These are known as global gauge transformations since the phase factor α does not depend on x. If we now demand invariance of physical magnitudes under such gauge transformations, then ϕ and j satisfy this requirement. But suppose we impose *local* gauge invariance, i.e. allow the phase factor α to be a function $\alpha(x)$ of x. ϕ remains invariant but j does not. In order to obtain a gauge-invariant current we introduce the following device. Replace $\frac{d}{dx}$ by a new sort of derivative $\frac{d}{dx} - iA(x)$ where A transforms according to $A \to A + \frac{d}{dx}\alpha(x)$. Then the modified current $j(x) = \frac{1}{2}ie\big(\psi^*\big(\frac{d}{dx} - iA\big)\psi - \psi\big(\frac{d}{dx} - iA\big)\psi^*\big)$ *is* gauge-invariant. But this has been achieved by introducing a new field $A(x)$ as a necessary concomitant of the original field $\psi(x)$. Reverting to three spatial dimensions, the $\vec{A}$ field can be identified (modulo the electronic charge e) with the magnetic vector potential, and the transformation law for $\vec{A}$ is exactly that described for the vector potential in the last section. The requirement of local gauge-invariance can be seen as requiring the introduction of a magnetic interaction for the ψ field.

Again we have an example here of physical structure being controlled by requirements imposed on surplus mathematical structure. The situation is illustrated schematically in figure 6. p_1, p_2, p_3 are three physical magnitudes, for example the charge or current at three different spatial locations. They are mapped onto m_1, m_2, m_3 in the mathematical structure M which is a substructure in the larger structure M'. The circles c_1, c_2, c_3 in the surplus structure represent possible phase angles associated with m_1, m_2, m_3 in a many–one fashion as represented by the arrows projecting c_1, c_2, c_3 onto m_1, m_2, m_3. Local gauge transformations represented by the arrows on the circles act independently at different spatial locations. They correspond to identity transformations on M and correlatively on P.

The $\vec{A}$ field establishes what mathematicians call a connection, correlating phases on the different circles c_1, c_2, c_3. The gauge transformations alter the connection as well as the individual phases in such a way as to maintain the gauge-invariance of the corrected ‘derivative’ $\vec{\nabla} - i\vec{A}$.

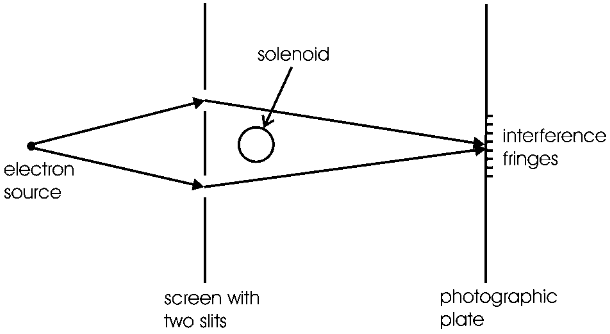

Figure 7. The Aharonov–Bohm experiment.

Two ways of dealing with the surplus structure inherent in gauge theories suggest themselves. Firstly, we might just fix the gauge by some arbitrary convention,[3] but then we have lost the possibility of expressing gauge transformations which lead from one gauge to another. Alternatively, we might try to formulate the theory in terms of gauge-invariant quantities, which are the physically 'real' quantities in the theory. Thus instead of the gauge potential, the $\vec{A}$ field in electromagnetism, we should employ the magnetic induction $\vec{B}$, specified by the equation $\vec{B} = curl\,\vec{A}$.

However, this manoeuvre has the serious disadvantage of rendering theory non-local! This is most clearly seen in the Aharonov–Bohm effect[4] in which a phase shift occurs between electron waves propagating above and below a long (in principle infinitely long) solenoid. The experiment is illustrated schematically in figure 7.

The magnetic induction is, of course, confined within the solenoid, so if it is regarded as responsible for the phase shift, it must be regarded as acting non-locally. On the other hand the vector potential extends everywhere outside the solenoid, so if invested with physical reality its effect on the electron phases can be understood as occurring locally. This is an argument for extending physical reality to elements which originated as elements of surplus structure.

However, just as in the case of free electromagnetism discussed in the previous section, the time-evolution of the vector potential is indeterministic since it is only specified up to the unfolding of a, in general, time-dependent gauge transformation. To restore determinism we must regard the gauge as being determined by additional 'hidden variables' which pick out the One True Gauge; this seems a highly *ad hoc* way of proceeding as a remedy for restoring determinism. This is indeed a quite general feature of Yang–Mills gauge theories.[5]

[3] In some pathological cases this may not be consistently possible, a phenomenon known in the trade as the Gribov obstruction.

[4] The interpretation of the Aharonov–Bohm effect has occasioned considerable controversy in the philosophical literature. See, in particular, Healey (1997), Belot (1998), and Leeds (1999).

[5] For a detailed discussion see a paper presented by Lyre at the Fifth International Conference on the History and Foundations of General Relativity, 8–11 July 1999, University of Notre Dame, Notre Dame, Indiana (E-Print: gr-qc/9904036).

7 The case of general relativity

The general arena for Yang–Mills gauge theories is provided by the notion of a fibre bundle. Speaking crudely a fibre bundle can be thought of as being constructed by attaching one sort of space, the fibre, to each point of a second sort of space, the base space, so that *locally* the structure is just the familiar Cartesian product.

We can effectively redraw figure 6 in a way that brings out the bundle structure, as illustrated in figure 8.

The local gauge group changes the phases according to the action of the $U(1)$ group. A cross-section of 'parallel' or constant phase is specified by the connection field, i.e. the gauge potential.

In the case of general relativity (GR) we are dealing with the bundle of tangent spaces at each point of the spacetime manifold, or more appositely the frame bundle, specifying the basis (or frame) for the tangent space at every point. The gauge group is now the group of general 4-dimensional frame transformations, usually denoted by $GL(4, \mathbb{R})$. If consideration is restricted to Lorentzian frames the gauge group reduces to the familiar Lorentz group $SO(1, 3)$ (or one might want to consider $SL(2, \mathbb{C})$, the covering group of $SO(1, 3)$, if spinor fields are to be introduced). There are now two ways to go. Stick with the Lorentz group, and introduce a connection field to define parallel transport of frames from one point of spacetime to another. This was the original approach of Utiyama (1956). But it has been claimed repeatedly in the literature that if one wants to generalize classical relativity, so as to allow for torsion in the spacetime manifold, it is necessary to introduce an affine structure into the fibres (to be sharply distinguished from an affine connection on the bundle), so the local symmetry group becomes the *inhomogeneous* Lorentz group, i.e. the Poincaré group. Of course, this can be done from a purely mathematical point of view, but does not really make any *physical* sense at all. The translation subgroup effectively changes the origin, i.e. the point of attachment of the tangent space to the spacetime manifold, so inhomogeneous frame transformations correspond picturesquely to sliding the tangent space over the base space, but that is *not* what

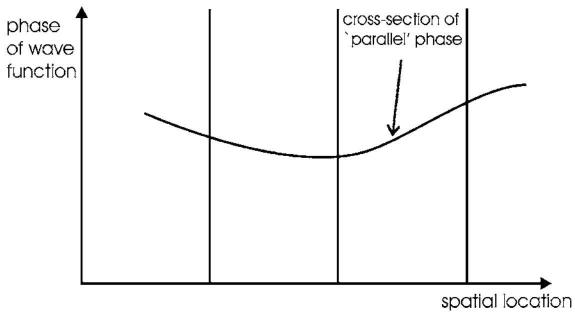

Figure 8. Fibre bundle structure of Yang–Mills gauge theory corresponding to figure 6.

local gauge transformation are supposed to do – they move points around in the fibre at a *fixed* point on the base space. I refer the reader to Ivanenko and Sardanshvily (1983) or Göckeler and Schücker (1987), who support, in my view correctly, the view that we do not need an affine bundle at all in order to extend GR to the Einstein–Cartan $\mathbb{U}_4$ theory incorporating spin and torsion.

So, there is considerable confusion between the Lorentz group and the Poincaré group as to the appropriate Yang–Mills gauge group for GR and its generalizations, but it is also often claimed that general coordinate transformations (the subject of general covariance) provide the gauge group of GR! The following comments are intended to clarify what is going on here. Firstly, it should be noted that general coordinate transformations do not in general constitute a group from the global point of view, since in general they cannot be defined globally. But there is a globally defined symmetry group, which is an invariance group of GR, namely the diffeomorphism group, *diff*, which from the local point of view is the active version of local coordinate transformations. From the bundle point of view described above, elements of *diff* move points around in the base space, which is just the spacetime manifold. This is not directly connected with gauge freedom in the more specialized sense we have defined, that is to say either in the Yang–Mills sense or as arising in the theory of constrained Hamiltonian systems as described in section 5 above. To link up with the latter notion, we need to exhibit GR in a canonical formulation, sometimes referred to as the $(3 + 1)$ approach to GR as compared with the 4-dimensional approach of the more familiar covariant formulation. In the $(3 + 1)$ approach the configuration variables are the 3-geometries on a spatial slice at a given coordinate time. (The collection of all possible 3-geometries is what is often referred to as *superspace*.) The Hamiltonian (canonical) variables satisfy constraints, indeed the Hamiltonian itself vanishes identically. The gauge freedom arises essentially as a *manifestation* of the diffeomorphism invariance of the 4-dimensional covariant formulation, in the $(3 + 1)$ setting. In this setting there are two sorts of gauge motion, one sort acting in the spatial slices and corresponding to diffeomorphisms of the 3-geometries, the other acting in time-like directions and corresponding to time-evolution of the 3-geometries.

The fact that time-evolution is a gauge motion, and hence does not correspond to any change at all in the 'physical' degrees of freedom in the theory, produces the famous 'problem of time' in canonical GR! Crudely this is often referred to under the slogan 'time does not exist!' In a Pickwickian sense the indeterminism problem for constrained Hamiltonian systems is solved because time-evolution itself lies in a gauge orbit rather than cutting across gauge orbits, as in figure 5. The solution of the problem of time (which plagues attempts to quantize canonical GR), must involve in some way identifying some combination of the *physical* degrees of freedom with an *internal* time variable. But exactly how to do this remains a matter of controversy among the experts in canonical approaches to quantum gravity.[6]

8 BRST symmetry

In the path integral approach to general (non-Abelian) gauge theories, a naive approach would involve integrating over paths which are connected by gauge transformations. To make physical sense of the theory, the obvious move is to 'fix the gauge', so that each path intersects each gauge orbit in just one point. However, early attempts to derive Feynman rules for expanding the gauge-fixed path integral in a perturbation expansion led to an unexpected breakdown of unitarity.[7] This was dealt with in an *ad hoc* fashion by introducing fictitious fields, later termed ghost fields, which only circulated on internal lines of the Feynman diagrams in such a way as to cure the unitarity problem, but could never occur as real quanta propagating along the external lines of the diagrams. So getting rid of one sort of surplus structure, the unphysical gauge freedom, seemed to involve one in a new sort of surplus structure associated with the ghost fields.

The whole situation was greatly clarified by the work of Fadeev and Popov (1967) who pointed out that when fixing the gauge in the path integral careful consideration must be given to transforming the measure over the paths appropriately. The transformation of the measure was expressed in a purely mathematical manoeuvre as an integral over scalar Grassmann (i.e. anticommuting) fields which were none other than the ghost (and antighost) fields!

The effective Lagrangian density could now be written as the sum of three terms, $\mathcal{L}_{eff} = \mathcal{L}_{gi} + \mathcal{L}_{gf} + \mathcal{L}_{ghost}$, where $\mathcal{L}_{gi}$ is a gauge-invariant part, $\mathcal{L}_{gf}$ is a non-gauge-invariant part arising from the gauge fixing, and $\mathcal{L}_{ghost}$ is the contribution from the ghost fields.

$\mathcal{L}_{eff}$ no longer, of course, has the property of gauge-invariance, but it was discovered by Becchi, Rouet, and Stora (1975) and independently by Tyutin (1975) that $\mathcal{L}_{eff}$ does exhibit a kind of generalized gauge symmetry, now known as BRST symmetry, in which the non-invariance of $\mathcal{L}_{gf}$ is compensated by a suitable transformation of the ghost fields contributing to $\mathcal{L}_{ghost}$.

To see how this comes about we consider the simplest (Abelian) case of scalar electrodynamics. The matter field ψ satisfies the familiar Klein–Gordon equation. Under the local gauge transformation $\psi \rightarrow \psi e^{i\alpha(x)}$, where x now stands for the 4-dimensional spacetime location x^{μ}, the gauge-invariance of the Lagrangian for the free field is restored by using the corrected derivative $\partial \rightarrow \partial_{\mu} - iA_{\mu}$, where

[6] For a comprehensive account of canonical quantum gravity and the 'problem of time' reference may be made to Isham (1993).
[7] Cf. Feynman (1963).

the gauge potential A_μ can be identified, modulo the electronic charge, with the electromagnetic 4-potential. A_μ transforms as $A_\mu \to A_\mu + \partial_\mu \alpha(x)$. The field strength $F_{\nu\mu} = A_{\nu,\mu} - A_{\mu,\nu}$ is gauge-invariant and measures the curvature of the connection field A_μ in the geometrical fibre bundle language. All that we have done here is just a relativistic generalization of the discussion already given in section 6.

To formulate the BRST transformation we consider a 5-component object

$$\Phi = \begin{pmatrix} \psi \\ A_\mu \\ \eta \\ \omega \\ b \end{pmatrix}$$

where ψ is the matter field, A_μ the gauge potential which we have already introduced above, η is the ghost field, ω the antighost field, and b is what is usually termed a Nakanishi–Lautrup field.

η and ω are anticommuting (Grassmann) scalar fields. The fact that they violate the spin-statistic theorem, which would associate scalar fields with commuting variables, emphasizes the unphysical character of the ghosts and antighosts.

We have then

$$\omega^2 = \eta^2 = 0.$$

The BRST symmetry is defined by

$$\Phi \to \Phi + \epsilon s \Phi$$

where ϵ is an infinitesimal Grassmann parameter and

$$s\Phi = \begin{pmatrix} i\eta\psi \\ \partial_\mu \eta \\ 0 \\ b \\ 0 \end{pmatrix}.$$

The first two components of $s\Phi$ comprise just the infinitesimal version of a gauge transformation with the arbitrary spacetime function $\alpha(x)$ replaced by the ghost field η. But ϵ is a constant so the BRST transformation is a curious hybrid. It is in essence a non-linear rigid fermionic transformation, which contains within itself, so to speak, a local gauge transformation specified by a dynamical field, namely the ghost field.

What is the role of the Nakanishi–Lautrup field? By incorporating this field the transformation is rendered nilpotent,[8] i.e. it is easily checked that $s^2\Phi = 0$. But this means that s behaves like an exterior derivative on the extended space of fields.

[8] The original BRST transformation failed to be nilpotent on the antighost sector.

This in turn leads to a beautiful generalized de Rham cohomology theory in terms of which delicate properties of gauge fields, such as the presence of anomalies, the violation of a classically imposed symmetry in the quantized version of the theory, can be given an elegant geometrical interpretation.[9]

But now we can go further. Instead of arriving at the BRST symmetry via the Fadeev–Popov formalism, we can forget all about gauge symmetry in the original Yang–Mills sense, and impose BRST symmetry directly as the fundamental symmetry principle. It turns out that this is all that is required to prove the renormalizability of anomaly-free gauge theories such as those considered in the standard model of the strong and electroweak interactions of the elementary particles.

But we may note in passing that for still more recondite gauge theories, further generalizations have had to be introduced.[10]

1. In a sense the ghosts compensate for the unphysical degrees of freedom in the original gauge theories. But in some cases the ghosts can 'overcompensate' and this has to be corrected by introducing ghosts of ghosts, and indeed ghosts of ghosts of ghosts etc!
2. For the more general actions contemplated in string and membrane theories the so-called Batalin–Vilkovisky antifield formalism has been developed. This introduces partners (antifields) for all the fields, but the antifield of a ghost is not an antighost and the anti (antighost) is not a ghost!

9 Conclusion

As we have seen, there are three main approaches to interpreting the gauge potentials.

The first is to try and invest them with physical reality, i.e. to move them across the boundary from surplus structure to M in the language of figure 4. The advantage is that we may then be able to tell a local story as to how the gauge potentials bring about the relative phase shifts between the electron wave functions in the Aharonov–Bohm effect, but the disadvantage is that the theory becomes indeterministic unless we introduce *ad hoc* hidden variables that pick out the One True Gauge.

The second approach is to try and reformulate the whole theory in terms of gauge-invariant quantities. But then the theory becomes non-local. In the case of the Aharonov–Bohm effect this can be seen in two ways. If the phase shift is attributed to the gauge-invariant magnetic induction this is confined *within* the solenoid whereas the experiment is designed so that the electron waves propagate *outside* the solenoid. Alternatively we might try to interpret the effect not in terms

[9] See Fine and Fine (1997) for an excellent account of these developments.

[10] Weinberg (1996, chapter 15) may be consulted for further information on these matters.

of the $\vec{A}$ field itself which of course is not gauge-invariant but in terms of the gauge-invariant holonomy integral $\oint \vec{A} \cdot d\vec{l}$ taken round a closed curve C encircling the solenoid. (This by Stokes theorem is of course just equal to the flux of magnetic induction through the solenoid.) But if the fundamental physical quantities are holonomies, then the theory is again clearly 'non-local', since these holonomies are functions defined on a space of loops, rather than a space of points.

Furthermore, with this second approach, the principle of gauge invariance cannot even be formulated since gauge transformations are defined by their action on non-gauge-invariant quantities such as gauge potentials, and in the approach we are now considering the idea is to eschew the introduction of non-gauge-invariant quantities altogether!

So this leaves us with the third approach. Allow non-gauge-invariant quantities to enter the theory via surplus structure. And then develop the theory by introducing still more surplus structure, such as ghost fields, antifields and so on. This is the route that has actually been followed in the practical development of the concept of gauge symmetry as we have described in the previous section.

But this leaves us with a mysterious, even mystical, Platonist-Pythagorean role for purely mathematical considerations in theoretical physics. This is a situation which is quite congenial to most practising physicists. But it is something which philosophers have probably not paid sufficient attention to in discussing the foundations of physics. The gauge principle is generally regarded as the most fundamental cornerstone of modern theoretical physics. In my view its elucidation is the most pressing problem in current philosophy of physics. The aim of the present paper has been, not so much to provide solutions, but rather to lay out the options that need to be discussed, in as clear a fashion as possible.

References

Becchi, C., Rouet, A., and Stora, R. (1975). 'Renormalization of the Abelian Higgs–Kibble model'. *Communications in Mathematical Physics*, **42**, 127–62.

Belot, G. (1998). 'Understanding electromagnetism'. *The British Journal for the Philosophy of Science*, **49**, 531–55.

Fadeev, L. D., and Popov, V. N. (1967). 'Feynman diagrams for the Yang–Mills field'. *Physics Letters B*, **25**, 29–30.

Feynman, R. P. (1963). 'Quantum theory of gravity'. *Acta Physica Polonica*, **24**, 697–722.

Fine, D., and Fine, A. (1997). 'Gauge theory, anomalies and global geometry: the interplay of physics and mathematics'. *Studies in History and Philosophy of Modern Physics*, **28**, 307–23.

Göckeler, M., and Schücker, T. (1987). *Differential Geometry, Gauge Theories, and Gravity*. Cambridge: Cambridge University Press.

Healey, R. (1997). 'Nonlocality and the Aharonov–Bohm effect'. *Philosophy of Science*, **64**, 18–40.

Isham, C. J. (1993). ‘Canonical quantum gravity and the problem of time’. In *Integrable Systems, Quantum Groups, and Quantum Field Theories*, ed. L. A. Ibort and M. A. Rodriguez, pp. 157–287. Dordrecht: Kluwer.

Ivanenko, D., and Sardanshvily, G. (1983). ‘The gauge treatment of gravity’. *Physics Reports*, **94**, 1–45.

Leeds, S. (1999). ‘Gauges: Aharonov, Bohm, Yang, Healey’. *Philosophy of Science*, **66**, 606–27.

Pauli, W. (1941). ‘Relativistic field theories of elementary particles’. *Reviews of Modern Physics*, **13**, 203–32.

Redhead, M. L. G. (2001). ‘The intelligibility of the universe’. In *Philosophy at the New Millennium*, ed. A. O’Hear. Cambridge: Cambridge University Press.

Shaw, R. (1954). *The problem of particle types and other contributions to the theory of elementary particles*. Ph.D. thesis, Cambridge University.

Tyutin, I. V. (1975). ‘Gauge invariance in field theory and statistical mechanics’. *Lebedev Institute Reprint N 39*.

Utiyama, R. (1956). ‘Invariant theoretical interpretation of interaction’. *Physical Review*, **101**, 1597–607.

Weinberg, S. (1996). *The Quantum Theory of Fields, Vol. 2: Modern Applications*. Cambridge: Cambridge University Press.

Weyl, H. (1918). ‘Gravitation und Elektrizität’. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, pp. 465–80.

(1929). ‘Elektron und Gravitation’. *Zeitschrift für Physik*, **56**, 330–52.

Yang, C. N., and Mills, R. L. (1954). ‘Conservation of isotopic spin and isotopic gauge invariance’. *Physical Review*, **96**, 191–5.