3

On continuous symmetries and the foundations of modern physics

CHRISTOPHER A. MARTIN

As far as I see, all a priori statements have their origin in symmetry.

H. Weyl

Symmetry principles have moved to a new level of importance in this century and especially in the last few decades: there are symmetry principles that dictate the very existence of all the known forces of nature.

S. Weinberg

1 Introduction

It is difficult to overstate the significance of the concept of symmetry in its many guises to the development of modern physics. Indeed one could reasonably argue that twentieth-century physics with its pillar achievements of successful physical theories of spacetime/gravitation and the electromagnetic and nuclear interactions merits calling that century the ‘Century of Symmetry’. For symmetry played a key part in each of these developments. Of particular significance are symmetries described by continuous (Lie) groups of transformations. Such symmetry groups play a central role both in our current understanding of fundamental physics and in various attempts to go beyond this understanding. Today such continuous symmetries are often assigned a, if not the, fundamental role in the worldview of modern physics. The precise nature of this role, though, is not entirely unambiguous. Just what significance – physical, philosophical, and otherwise – is to be ascribed to the preeminent role of such symmetry groups in fundamental physics?

This paper aims to provide a brief and necessarily selective survey of the historical development of the place of continuous symmetries in physical theory. I discuss the central role that symmetry came to play with the development of relativity theory at the beginning of the twentieth century. After that, I discuss the central significance of symmetry considerations in quantum theory. My primary focus, though, is on the symmetry-dictates-interaction paradigm characteristic of
the gauge field theory programme, which has proved incredibly successful in describing the fundamental physical interactions. It is with this last development that considerations of symmetry have come to occupy the Platonic heights that they are often ascribed in current physics.

In discussing the role of symmetries and symmetry principles in physical theory I touch upon a number of interesting, more ‘philosophical’ issues. To what heights precisely have symmetry principles really risen, and on the basis of what? Have these come to occupy the realm of the unassailable a priori or are they ultimately of mere heuristic value? What, if anything, can be said to be the ‘physical content’ of the gauge symmetry principle purportedly lying at the heart of successful modern theory? Do such symmetry principles capture/reflect deep ontological features of the world or are they rather artifacts of our particular formal representation? A recurring theme/concern here is whether and, if so, how we are to reconcile the canonical view of gauge invariance as relating to a non-physical, formal redundancy in theory with the general belief that gauge symmetry is in fact of some deep physical significance.

2 The ‘Century of Symmetry’

2.1 Groups, invariants, and conservation

Although groups made their appearance in physics at the beginning of the nineteenth century, it was not until the detailed study of group representations in the 1920s which accompanied the developing quantum theory (discussed below) that groups truly made their way into a large part of work-a-day physics. There were, of course, major developments regarding the role of symmetry in physical theory before this. Besides the development of relativity theory, discussed in the next subsection, one development of seminal significance was the formulation of what are now called Noether’s theorems. Noether’s work incorporated three areas of mathematical physics (Kastrup, 1983, p. 115): (i) algebraic and differential invariant theory; (ii) Riemannian geometry and variational calculus in the contexts of general relativity, mechanics, and field theory; (iii) group theory and, in particular, the methods developed by Lie for solving or reducing differential equations by appeal to their invariance groups. As these areas remain central to modern physics, Noether’s theorems have come to play a central role in our current physical worldview.

Noether’s theorems grew out of her work while in Göttingen at the invitation of Hilbert and Klein. Klein had many years earlier advanced his influential Erlangen programme which concerned the application of group theory to geometry, specifically classification of geometries according to their characteristic invariants under groups of transformations. Both Klein and Hilbert had been interested in the

---

1 See Kastrup (1983) for a detailed history of the genesis and reception of Noether’s work. See also Rowe (1999).
connection between invariances and conservation laws – Hilbert, in particular, with regard to his work on formulating a relativistic theory of gravitation. Noether’s work is now considered as providing the first general treatment of this relationship.

The formal context of Noether’s theorems is Lagrangian field theory wherein equations of motion are obtained through a variational procedure (Hamilton’s principle) from the action integral, \( S = \int L \, dx \). Noether’s theorems are discussed in detail in Earman (2002) and Brading and Brown (this volume). Noether’s two theorems concern the invariances of the action integral under continuous groups of transformations of the fields. The first theorem shows that if the integral is invariant under an \( r \)-parameter (\( r \) finite) Lie group \( G_r \) of transformations then, when the equations of motion are satisfied, there exist \( r \) conserved ‘currents’. As a familiar example, this theorem relates the invariance of the action under spatial and temporal translations and rotations to the conservation of linear momentum, energy, and angular momentum respectively. The second theorem shows that if the action is invariant under an infinite dimensional (i.e. specified by \( r \) functions) group \( G_{\infty r} \) of transformations then there will be \( r \) identities holding between the Euler–Lagrange equations of motion derived from the action. These identities are commonly called generalized Bianchi identities. In electrodynamics, for example, the generalized Bianchi identities associated with the infinite dimensional group of electromagnetic gauge transformations are just the homogeneous Maxwell field equations.

There is another result of Noether’s which is perhaps equally as important to understanding the role that invariance plays in physical theory. If the action admits both \( G_{\infty r} \) and, as a rigid subgroup, \( G_r \) as symmetries, then the relations of Noether’s first theorem, from which the conservation laws are derived, are consequences of the generalized Bianchi identities associated with the second.

The Bianchi identities are tied to an underdetermination in the dynamical evolution of the field(s), there being fewer independent equations of motion than dynamic variables.\(^2\) This is where gauge symmetry enters the story. The gauge transformations are taken to relate physically equivalent situations—distinct situations evolved from a single set of initial data – and rescue determinism. Lagrangians with such local symmetries, so-called singular Lagrangians, lead, in the canonical Hamiltonian framework, to so-called constrained Hamiltonian systems. Appeal to the constrained Hamiltonian framework is an important part of treating theories with local gauge symmetry, as the Hamiltonian framework is central to (i) the canonical means of formulating quantum field theories and, relatedly, (ii) sorting out the dynamical degrees of freedom. The treatment of gauge theories in the canonical Hamiltonian formalism is discussed in Earman (this volume, Part I) and Castellani (this volume, Part IV).

\(^2\) Note that, at the same time, the equations of motion are overdetermined in the sense that there are constraints on the sets of admissible initial data.
With the advantage of hindsight, we see that the importance of Noether’s work is the establishment of systematic links between (i) invariance of the action under groups of transformations and conserved quantities, these commonly being the observables with which physics directly deals (in the quantum context, they in fact generate the symmetry transformations); and (ii) invariance of the action under infinite dimensional groups and certain specific structural properties of the associated theories.\(^3\) This link between invariance(s) of the action and other detailed features of physical theory has come to be one of the defining features of modern physical (gauge) theory.

### 2.2 Relativity, spacetime and gravitation

Not only is the Special Theory of Relativity (STR) tied directly to considerations of symmetry at the physical and formal levels, but Albert Einstein’s formulation of STR is commonly taken to signal an important shift in the overall relation of symmetry to physical theory. When one speaks of symmetry ‘principles’ this refers to a view of symmetry as lying at the heart of physical theory. Einstein’s work on special relativity with its central role for symmetry principles is understood to be the birth of this way of thinking about the significance of symmetry to physical theory.\(^4\)

With Einstein’s work we see a shift to starting from symmetries and then deducing from these various physical consequences, laws, etc. This is in contrast to taking symmetries to be interesting but after-the-fact features of some known law(s). This elevation of symmetry principles fits squarely with Einstein’s well-known regard for theories of principle, such as Euclidean geometry and thermodynamics.

Einstein’s principle of relativity, one of the centrepieces of his 1905 formulation of STR, requires that it be impossible to detect one’s state of inertial motion. That is, experimentation in one inertial frame must uncover the same laws of physics as experimentation done in any other inertial frame. At base, this has the shape of a symmetry principle – the relevant transformations being those relating inertial frames of reference and the invariant being the (form of the) physical laws. In order to reconcile this invariance requirement with the known constancy of the speed of light, Einstein was forced to consider a new group of transformations relating the space and time coordinates of inertial observers. This group is now termed the inhomogeneous Lorentz group (or Poincaré group, since Poincaré had earlier written down the transformations).

The work of Minkowski elucidated the nature of the symmetry at the heart of STR. In the spirit of Klein’s Erlangen programme, he considered the geometry

---

\(^3\) For further discussion, see Brading and Brown (this volume).

Continuous symmetries

associated with the transformations advanced by Einstein. Minkowski showed that these transformations and associated invariants characterized a new and arguably even simpler geometry of the physical world – space and time were now part of a single four-dimensional geometry, (Minkowski) spacetime. With this, symmetry and its mathematical counterpart, group theory, became fixtures in fundamental physics and in its relativistic revamping in the first part of the twentieth century.

As typically construed, the relativity principle at the heart of STR is taken to be grounded in certain regularities in the underlying structure of physical events. Formally, the symmetry transformations relating equivalent inertial observers are taken to be symmetries of the absolute, i.e. non-dynamical, special relativistic Minkowski spacetime – i.e. elements of the symmetry (isometry) group of the Minkowski metric. Some, though, have raised issues with construing the relativity principle in this way. They argue that the common formulation of the principle in terms of global (symmetry) transformations of the entire spacetime (and its contents) lacks significant empirical content and, relatedly, that it overlooks that the basic empirical relativity principle is in fact defined only for isolated subsystems of the universe.

Shortly after the formulation of STR, Einstein turned his attention to developing a relativistic theory of gravitation, and considerations of symmetry and covariance once again figured prominently in his work. Symmetry and covariance are all tied up historically. One of Einstein’s primary starting points was the familiar equivalence of gravitational and inertial masses. Einstein was led to advance his principle of equivalence – the equivalence between the effects of homogeneous gravitational fields and uniformly accelerated motion. Towards accommodating this equivalence into a relativistic theory of the gravitational field, Einstein came to consider theories with wider covariance than that characteristic of STR. Through a myriad of interesting twists and turns, Einstein arrived at a principle of general covariance, the form invariance of the laws of nature under arbitrary smooth coordinate transformations. And, with considerations of symmetry playing a significant role, after much struggle Einstein formulated generally covariant gravitational field equations, which relate the distribution of matter and energy in spacetime to the presence of

---

5 See for example Anderson (1967) and Friedman (1983).

6 Intuitively, the absolute, or non-dynamical objects (fields, etc.) are those that are unaffected by interaction with the other objects of the theory, and are the same in each allowed model of the theory. Dynamical objects, on the other hand, interact with other objects, and vary from model to model.

7 See Brown and Sypel (1995) and Budden (1997), and references therein.

8 The relevant covariance in the context of STR is of course Lorentz (or Poincaré) covariance. I discuss the matter of invariance vs. covariance below. Historically, matters of symmetry/invariance and covariance were sometimes run together.

9 For discussion of the foundations of spacetime theories and, in particular, the place of general covariance and of symmetry generally, see Anderson (1967), Friedman (1983), Earman (1989), Norton (1993 and this volume), and Brown and Sypel (1995).
Christopher A. Martin

spacetime curvature.\textsuperscript{10} Importantly, the, now, dynamical spacetime metric itself comes to represent/encode the properties of the (dynamical) gravitational field. Relatedly, the familiar equivalence of inertial and gravitational masses is seen to flow from these two in fact being one and the same thing.

With the formulation of the General Theory of Relativity (GTR), symmetry played a key role in developing a new theory of a specific interaction, gravity. Though considerations of symmetry had played a role in the development and understanding of both STR and GTR, it was pointed out early on that there are important differences between these roles. As already mentioned, the symmetry at the heart of STR is grounded in the invariances (isometries) of the Minkowski spacetime. As such, the symmetry applies to all interactions taking place in the spacetime. In contrast to the symmetry or invariance requirement in STR, the principle in GTR is most often presented as strictly speaking a covariance requirement. General covariance, it is argued, is not tied to any geometrical regularity of the underlying spacetime, but rather to the form invariance (covariance) of laws under arbitrary smooth coordinate transformations. Einstein is at least partly to blame for some initial confusion regarding this distinction since he sometimes said that GTR represented a generalization of STR, that is was a ‘more’ relativistic or symmetric theory.\textsuperscript{11} Suffice it to say that considerations of symmetry/invariance and covariance are historically intertwined. Today we take it that Einstein sometimes spoke of covariance as a symmetry property in ways he perhaps should not have done.

The ascription of any physical content to a formal covariance requirement such as general covariance came under fire early on from Kretschmann. There is a long history of agreement and disagreement with Kretschmann’s claim. Some of these are discussed in Norton (this volume). Notwithstanding Kretschmann’s objection, the place of general covariance in the formulation of GTR, as a (if not the) characteristic feature of the theory, is often taken as the first example of the appeal to a new sort of symmetry principle, so-called local symmetry principles. More on such principles below. It is also worth mentioning here a related point. The contrasts between the ‘symmetries’ at the hearts of STR and GTR are an example of what gets codified in Wigner’s distinction between ‘geometrical-invariances’ and ‘dynamical-invariances’. I discuss this distinction further in section 4.1.

Weyl (1918a), which contained the first appearance of the concept of local ‘gauge’ symmetry, is a direct outgrowth of Einstein’s work on GTR.\textsuperscript{12} Weyl’s

\textsuperscript{10} The story here is much more complicated and interesting. In fact, Einstein wrestled with the implications of general covariance, at one point becoming convinced that he should drop the requirement in founding a relativistic theory of gravity. See Stachel (1989a; 1989b).
\textsuperscript{11} cf. note 7, above.
\textsuperscript{12} See also Weyl (1918b). Pauli (1921) contains a nice discussion of Weyl’s 1918 work.
starting point, ostensibly mathematical, was to consider a sort of generalization of Einstein’s work. He contended that the Riemannian geometry of Einstein’s GTR retained one vestige of Euclidean geometry—at-a-distance (*ferngeometrisches*). Although parallel transport of vectors was in general non-trivial due to spacetime curvature, the geometry of GTR did allow direct comparison of distant magnitudes. Weyl considered how one might generalize the Riemannian geometry of GTR, removing this methodological inconsistency (*Inkonsequenz*), thereby founding a purely infinitesimal geometry.\(^{13}\) Specifically, Weyl considered the geometry that results from allowing arbitrary rescalings of the local unit of length—formally rescalings of the spacetime metric—at each point of spacetime. This change of unit length, or change of ‘gauge’, is the origin of the terminology ‘gauge invariance’.\(^{14}\)

What Weyl found was that the more general geometry resulting from admitting such local changes of ‘gauge’ apparently described not only gravity but also electromagnetism. In short, the field needed to maintain invariance given the freedom to perform local gauge transformations, part of a generalized affine connection, was formally identical to the field used to represent the electromagnetic potential. Under the local changes of gauge, this field transforms as does the electromagnetic potential under electromagnetic gauge transformations. With this field, specifically a generalization of the Riemannian curvature built from it, Weyl constructed a gauge-invariant action for the unified theory. What Weyl took to be the strongest argument in favour of his new theory, however, was that electric charge conservation followed from the local gauge invariance in precisely the same characteristic way that energy–momentum conservation followed from coordinate invariance.\(^{15}\)

What Weyl refers to is that the conservation laws follow in two distinct ways in theories with local symmetries. This is traceable to the Bianchi identities holding between the coupled equations of motion, which, in turn, are due to the local gauge invariance of the action. Thus, Weyl took it that via the demand of local gauge invariance he was led to the (geometrical) unification of two previously distinct interactions, and that this brought out important similarities between the two theories.

Given Weyl’s starting point (see Ryckman, this volume), it is perhaps not surprising that his theory faced immediate challenges from physics. Showing deference to Weyl’s mathematical ingenuity, Einstein was quick to point out that nature in fact provides us with standard rods and clocks in the form of atoms with definite spectral lines. This was in contrast to the basic assumption in Weyl’s theory that there was no absolute physical significance to be ascribed to the spacetime interval.

---

13 See Ryckman (this volume) for further discussion of the origins of local symmetry in Weyl’s work.
14 Originally ‘*MasstabInvarianz*’ and later ‘*EichInvarianz*’.
15 See Brading (2002) for a detailed discussion of Weyl’s derivation of the conservation laws.
Christopher A. Martin

Einstein argued that Weyl’s theory could not be correct since it predicts that, owing to the (in general) path-dependent gauge factor, neighbouring identical atoms would show different spectral lines depending upon their respective histories. In the next section I will continue with the historical account of the rise of gauge symmetry.

2.3 Symmetry and quantum theory

In the 1920s, on the back of important work by the likes of Rutherford, Bohr, and many others, physicists began to develop atomic physics, in the process giving birth to a new theory of matter and radiation – quantum theory. A crucial part of the rapid successes in this realm was the application of group-theoretic methods to various central physical problems. The reason for the utility (even necessity) of group theory in the development of quantum mechanics has chiefly to do with the radically different notion of the state of a physical system in the new theory. In classical mechanics, the state space is typically represented by a differentiable manifold, and observables by real-valued functions defined on the manifold. The states of quantum mechanical systems, on the other hand, are represented by vectors in an (infinite dimensional) vector space, and observables are linear mappings of these states onto one another. As such, quantum states can be added to one another yielding new states. The various symmetry groups carve up the quantum state space into subspaces invariant under the action of the associated transformations (alternatively, irreducible representations). The conserved quantities associated with the symmetries in fact ‘generate’ the associated symmetry transformations. Atomic states can then be characterized by irreducible representations of the various fundamental symmetry groups.

Thus, in the mid to late 1920s group theory was used to show that many empirically determined rules (e.g. spectral formulae, selection rules) were attributable to various symmetries of the physical system under study. For example, certain well-known spectral relations followed directly from the spherical symmetry of the hydrogen atom and the associated invariance of the physics under the rotation group. Hence, group theory provided key insights into the developing atomic theory even before there was a detailed understanding of the dynamics of the electron. This success was so complete that, referring to group theory, Weyl wrote:

We may well expect that it is just this part of quantum physics which is most certain of a lasting place… All quantum numbers, with the exception of the so-called principal quantum number, are indices characterizing representations of groups.18

16 There is more to this story: further argument in favour of his theory from Weyl and further objection by Einstein. See the correspondence and commentary in Einstein (1998).
17 Pais (1986) provides a detailed account of many of the historical developments discussed here.
By the early 1930s group theory had become a staple of the budding physics of fundamental matter. The books by Weyl (1928), Wigner (1931), and van der Waerden (1932) are universally recognized as the canonical works which set the tone and content of much of the development of (quantum) physics in that period.

Towards making contact with our main focus here, note that quantum theory introduced a new treatment of matter and with that a new physical variable to physics. The quantum matter (wave) field was described by a complex-valued function – the field had not only an amplitude but also a phase. As I discuss in the next section, this new quantum phase was a key part of a new role for symmetry in fundamental physics.

2.3.1 The ascendancy of symmetry: interaction from symmetry

Weyl (1929) is almost universally considered the origin of modern gauge symmetry principles.\(^\text{19}\) While it is sometimes said that Weyl here simply reinterpreted his earlier work, I think it is more correct to see it as substantial modification. The general idea of appealing to some sort of local symmetry requirement in arriving at electromagnetism is preserved, and the formal outline of the derivation remains the same. In his 1929 work, however, the local transformations are imaginary phase transformations of the complex-valued matter field, and the resulting theory is one of the quantum matter field interacting with the electromagnetic field. The demand of invariance under local phase transformations led Weyl to the by then well-known coupling of charged matter to the electromagnetic field.\(^\text{20}\) As in his 1918 theory, Weyl attributes a special importance to the fact that the conservation of electric charge followed from the local gauge invariance in the same characteristic way in his theory as does energy momentum conservation from coordinate invariance in GTR. This, again, is related to certain formal features of the theories tied specifically to the local invariance.

Interestingly, Weyl does not stipulate the demand of local gauge (better, phase) invariance as he had in his 1918 theory. Rather he argues that accommodating spinor fields into a general relativistic spacetime requires local phase invariance. In short, Weyl argues that one must use local tetrads in representing spinor matter fields in arbitrary curved spacetimes and that the choice of a local tetrad leaves the spinor field undetermined up to a phase factor. Weyl’s reasoning here is now seen to be misguided.\(^\text{21}\) As understood today, the phase freedom in quantum electrodynamics

\(^{19}\) Though Weyl was the first to put all the ingredients together in a nice package, numerous others had drawn attention to the same central idea, among them, London, Fock, Schrödinger, and Dirac. See O’Raifeartaigh (1997) and Vizgin (1994, esp. chapters 3 and 6).

\(^{20}\) In section 3, I look more closely at the details of this argument.

\(^{21}\) See O’Raifeartaigh (1997).
Christopher A. Martin

(QED) is not related in this way to representing spinors – which, anyway, are not inherently quantum mechanical in the first place – but to the representing of matter via elements of a complex vector space.22 Despite the questionable motivation for the central demand of local phase invariance, what Weyl does arrive at in his 1929 work is the archetype of the modern gauge symmetry principle: deriving the form of interactions from the demand of invariance under a group of local transformations – in this case, electromagnetism from the demand of invariance under the group of local \((U(1))\) phase transformations.

Jumping ahead in time, after the successful formulation of a fully renormalizable theory of QED, but in the face of many newly discovered particles and the failure of the standard methods for constructing theories of the new particles and their interactions, the 1950s and 1960s saw a low point in the quantum field theory programme – many abandoned its central tenets (even if only temporarily). While some effectively abandoned the appeal to symmetries and group theory, others turned to the further study of symmetries and conservation laws. It was in this context that the idea of local gauge invariance again emerged, in the context of describing the nuclear interactions. The seminal work here is Yang and Mills (1954) – today, the gauge theories central to much of current physics are often simply called Yang–Mills theories.23 Yang and Mills drew their inspiration from Heisenberg’s discussion of the suggestive similarities between the proton and neutron. Heisenberg had hypothesized that the neutron and the proton were two states of the same particle, the nucleon.24 The idea of the underlying \(SU(2)\) symmetry was that, in the absence of electromagnetism, the orientation of the isotopic spin would be of no physical significance – the neutron and the proton would be identical. Such transformations have come to be denoted ‘internal’ transformations since they are not tied to (‘external’) spacetime properties but rather to the intrinsic properties of the particles.

The starting point for Yang and Mills was consideration of the \(SU(2)\) symmetry of nucleon–nucleon interactions. Yang and Mills took it that this symmetry was subject to an important limitation which should be lifted. This is where local gauge invariance enters. With a global \(SU(2)\) symmetry, Yang and Mills argue, once one chooses what to call a proton/neutron at one point, one is no longer free to make independent choices at other spacetime points. Yang and Mills took this limitation to be inconsistent with the idea of a local field theory. Lifting this limitation through

---

22 Consider also that, presumably, the phase freedom is to remain in the case of zero spacetime curvature where the appeal to local tetrads (and, consequently, local phase invariance, as Weyl claims) is no longer necessary.

23 See O’Raifeartaigh (1997) for discussion of related developments, including the early role of five-dimensional theories, and related and independent work by W. Pauli and R. Shaw.

24 That the two were identical in this way was suggested by their similar masses and by the near equality of the interaction and self-interaction strengths. The differences were thought to arise from their different electromagnetic properties.
Continuous symmetries

demanding a local gauge (or isotopic spin, or $SU(2)$-phase) invariance, they were led to introduce a new field, an $SU(2)$ gauge field. This field bore the same relation to the demand of local $SU(2)$ invariance as did the electromagnetic field to the demand of local $U(1)$ invariance in Weyl’s 1929 theory. The appeal to a non-Abelian (i.e. non-commuting) symmetry group has the result that the new field satisfies non-linear equations of motion, because the gauge field carries isotopic spin and is thus self-interacting. Considering the quantization of the gauge field, Yang and Mills indicated a potential problem with putting their considerations to actual use. Though they discuss both, the question as to the mass of the $SU(2)$ gauge field and the related question of the renormalizability of the theory went unanswered in their paper. Such answers were, however, important in determining whether their theory matched experiment. Indeed, these and related matters were to plague the application of gauge theory to the nuclear interactions for some time.

Utiyama, working independently, further generalized this line of thought, cementing a ‘general rule for introducing a new field in a definite way when there exists some conservation law . . . or there is a Lie group . . . under which the system is invariant’ (Utiyama, 1956, p. 1601) Additionally, Utiyama discussed casting gravitation in this framework. Utiyama’s treatment of gauge theories was thus the most general and, as regards physics, the most comprehensive, (O’Raifeartaigh, 1997, p. 208). While the particulars of Yang and Mills’ and Utiyama’s papers pertaining to the description of the nuclear interactions turned out to be unsuccessful, they represented important generalizations of Weyl’s 1929 work. Even after nearly fifty years of intense development in physics, these works still provide the paradigm examples of describing the nuclear interactions through appeal to local symmetry principles.

The foregoing is only a relatively small (though central) part of the rise of gauge symmetry principles and gauge field theories. The generalizations of gauge field theories (GFTs) to non-Abelian symmetry groups by Yang and Mills and Utiyama nevertheless had no immediate home in physics. Most notably, the theories remained unusable due to the lurking ‘mass problem’: local gauge symmetry apparently required that the gauge field(s) be massless and, hence, long-range, in contradiction to the known short range of the strong interactions. Though there was initial appeal to GFTs given various observed conservations and similarities between certain particle

25 One can take Yang and Mills’ work to effectively generalize Weyl’s through adding a matrix index to the electromagnetic (gauge) potential.

26 Whether and in what way gravitation is a gauge theory in the Yang–Mills sense is a vexed topic that I do not broach here. See Trautman (1980), Ivanenko and Sardanashvily (1983), Weinstein (1999), and Earman (this volume, Part I).

27 Isospin symmetry is no longer seen as a fundamental symmetry but rather as an incidental symmetry: a consequence of the underlying $SU(3)$ colour symmetry of quantum chromodynamics and the fact that there are two light quarks.
interactions in the rapidly emerging experimental data, numerous other significant developments had to occur before GFTs could be put to work. The formulation of successful (quantum) GFTs occurred during an intense period of development in theoretical physics. Briefly, the key to solving the mass problem proved to be an idea imported from solid-state physics – spontaneous symmetry breaking. The idea was that the gauge particles (bosons) associated with the local symmetry requirement would acquire mass through the so-called ‘Higgs mechanism’. This mechanism gives the gauge fields mass without spoiling the gauge invariance of the underlying Lagrangian. Spontaneous symmetry breaking is discussed in Part III of this volume.

One of the most important elements of the development of successful gauge theories of the nuclear interactions was the careful study of the relationship between renormalizability and local gauge symmetry. Appealing to the spontaneous symmetry-breaking mechanism, Weinberg and Salam independently developed a unified theory of the weak and electromagnetic interactions based on the gauge group $SU(2) \times U(1)$. In this theory, what is originally a system of four massless vector bosons becomes, through the spontaneous breakdown of the ground state, a system of one massless particle (the photon), which corresponds to the unbroken $U(1)$ subgroup, and three massive particles (the $Z^0$ and the $W^\pm$) corresponding to the broken part of the group. It was not immediately clear, however, that the theory’s renormalizability was not compromised by the spontaneous symmetry breaking. In 1971 ‘t Hooft and Veltman showed that spontaneously broken gauge theories with massless bosons were renormalizable and that, in fact, gauge invariance of the theory was intimately related to the renormalizability of the theories. Renormalizability of the electroweak theory secured, the focus turned to prediction and experimentation. In the following decade, detailed predictions of, for example, the existence and properties of the weak neutral current and the $Z^0$ and $W^\pm$ bosons required by the gauge invariance were confirmed with amazing accuracy.

At the same time, ideas of local symmetry were being applied to developing a theory of the strong interactions – as it turned out, the interaction between particles making up the hadrons (baryons and mesons), quarks. Important here was the determination that non-Abelian gauge theories are asymptotically free, which is directly related to the fact that the effective coupling constant in the theory decreases with energy. And, rather than being a mere peculiarity, this proved promising as a possible

---

28 Crease and Mann (1996), Cao (1997; 1999), and Weinberg (1980) contain accounts of these developments.

29 As Gell-Mann reports (Doncel et al., 1983, p. 551), neither the renormalizability of the massless theory nor of the massive theory had been conclusively settled.

30 In fact, as was shown in the early to mid-1970s, non-Abelian gauge theories are the only asymptotically free theories in four spacetime dimensions.
Continuous symmetries

explanation for why nucleons, when bombarded with high-energy electrons, behaved as if their constituent particles were essentially free. Moreover, the increase of the coupling strength with distance in such a theory would explain why strongly interacting particles are not produced in isolation – infrared slavery, the other side of asymptotic freedom. These developments culminated in the postulation of an unbroken, internal gauge group, $SU(3)$ (the colour group), relating quark colour multiplets.\footnote{Note that $SU(3)$ also characterized the ‘eightfold way’ of Neéman and Gell-Mann. This was an earlier (and in many ways successful) classification of the hadrons (and resonances) according to an eight-dimensional, non-fundamental representation of $SU(3)$, with the transformations acting on the ‘flavour’, rather than colour, of the particles.} These quarks interact via the eight massless $SU(3)$ gauge (‘gluon’) fields – the massless vector fields required by the local $SU(3)$ symmetry (and which also carry colour charge and are, thus, self-interacting) in the Yang–Mills scheme. This led to a successful theory of the strong interactions, quantum chromodynamics (QCD). The developments discussed in this section led to the so-called Standard Model of particle physics based on the gauge group $SU(3) \times SU(2) \times U(1)$, essentially a pasting-together of QCD and the electroweak theory.

Before closing this all-too-brief history let us note that the ascendancy of gauge symmetry as a fundamental fixture of theoretical physics is, historically, to a large extent divorced from consideration of the \textit{a priori} reasonableness, or ‘deep physical basis’, or ‘physical meaning/content’ of gauge symmetry principles. Instead, this rise is founded on the hard-fought and impressive success of gauge field theories at the formal and physical level. In the end, the ascendancy of gauge symmetry principles is secured on very pragmatic bases. The eventual successes of gauge field theories of the fundamental interactions cement the heuristic value of local symmetry principles as guides in theorizing.\footnote{In the next section, I discuss how the ‘reasons’ for the success of the gauge field programme might, from our modern standpoint, be construed differently.} This might seem to constrain any real physical import of gauge principles to the context of discovery. And, along these lines, perhaps we should then count ourselves amazingly fortunate that the ‘right’ theories just happened to have such a nice structure, i.e. that seen in the theories’ tight group-theoretic structure which accompanies the characteristic symmetry/invariance. Notwithstanding this, the ‘gauge philosophy’ is often elevated and local gauge symmetry principles enshrined. Gauge symmetry principles are regularly invoked in the context of justification, as deep physical principles, fundamental starting points in thinking about why physical theories are the way they are, so to speak. This finds expression, for example, in the prominent current view of symmetry as undergirding our physical worldview in some strong sense. Next, I go on to further consider how literally and in what sense we might understand this undergirding.
3 Getting a grip on the gauge symmetry principle

3.1 Gauge principle – gauge argument – gauge heuristic

The canonical way of understanding the workings or content of local gauge symmetry principles is via the algorithm for producing interacting field theories from the demand of local gauge invariance – the ‘gauge argument’. Given the success of the gauge paradigm, some take the argument to reflect the logical order of nature. Symmetry principles as embodied in the argument are then taken to express/reflect deep features of the physical world. In order to make some remarks concerning how literally we can construe the operations of the argument, we first look at a concrete example.

Consider a field $\Psi$ representing electrically charged matter. The free field obeys the Dirac equation which is just the Euler–Lagrange equation(s) for the Lagrangian (density) $\mathcal{L}_{\text{Dirac}} = \bar{\Psi}(i\gamma^\mu \partial_\mu - m)\Psi$. The corresponding action is clearly invariant under so-called ‘global’ $U(1)$ phase transformations: $\Psi \rightarrow e^{iq\Lambda}\Psi$ ; $\bar{\Psi} \rightarrow e^{-iq\Lambda}\bar{\Psi}$ with $\Lambda$ a constant. It follows from Noether’s first theorem that when the equations of motion are satisfied there will be a corresponding conserved current.

Consider now ‘localizing’ these phase transformations, i.e. letting $\Lambda(x)$ become an arbitrary function of the coordinates $\Lambda(x)$: $\Psi \rightarrow e^{iq\Lambda(x)}\Psi$ ; $\bar{\Psi} \rightarrow e^{-iq\Lambda(x)}\bar{\Psi}$. As it stands, the free field Lagrangian is clearly not invariant under such transformations, since the derivatives of the arbitrary functions, i.e. $\partial_\mu \Lambda(x)$, will not vanish in general. The Lagrangian must be modified if the theory is to admit the local transformations as (variational) symmetries. In particular, we replace the free field Lagrangian with $\mathcal{L}_{\text{interacting}} = \bar{\Psi}(i\gamma^\mu \partial_\mu - m)\Psi - q A_\mu \bar{\Psi}\gamma^\mu \Psi \equiv \mathcal{L}_{\text{Dirac}} - J^\mu A_\mu$, with $J^\mu = q \bar{\Psi}\gamma^\mu \Psi$. This current is in fact the conserved current associated with the global $U(1)$ invariance of the interacting theory. Towards securing local invariance we have introduced the field $A_\mu$, the gauge potential. The particular form of coupling between the matter field and this gauge potential in $\mathcal{L}_{\text{interacting}}$ is termed minimal coupling.

---

33 What I here call the ‘gauge argument’ is in fact just a convenient label for what is an amalgam of the most common features figuring in many similar such arguments in the physics literature, both popular and technical/text-book. (For an example of the former see Mills (1989) or ’t Hooft (1980), and, for the latter, see Aitchison and Hey (1989) or Quigg (1982).) Though there are certainly variations in the way the argument is presented – and, most importantly, in the overall place and significance assigned to the argument relative to other features of gauge theory – what we consider here are the most central elements, the ones to which most presentations make appeal in one way or another.

34 The following is treated in some more detail in Martin (2002).

35 Here, I suppress all spinor indices, and $\Psi$ is just the Dirac conjugate of $\Psi$, and $\gamma^\mu$ are the usual Dirac matrices.

36 The action will be invariant, and thus the Euler–Lagrange equations determined from minimizing (extremizing) it covariant, if the Lagrangian is quasi-invariant, i.e. invariant up to an overall divergence, under the transformations.

37 In general, one must be careful in making this identification since for some types of field the form of the current is modified by the coupling.
This modified Lagrangian is now invariant under the local phase transformations provided that the vector field $A_\mu$ is simultaneously transformed according to $A_\mu(x) \rightarrow A_\mu(x) - \partial_\mu \Lambda(x)$. This transformation behaviour is, of course, familiar as the covariant analogue of the well-known electromagnetic gauge transformations. This suggests the possibility of viewing the new field $A_\mu$ as representing the electromagnetic potential.

Pursuing the idea of viewing the field $A_\mu$ as representing the electromagnetic potential, we note that the Lagrangian $\mathcal{L}_{\text{interacting}}$ does not yield a fully interacting theory. Varying the Lagrangian with respect to the matter fields yields the latter’s equations of motion – the field now being coupled to the $A_\mu$ field. But, it remains to add a ‘kinetic term’ for the $A_\mu$ field itself. Such a kinetic term, in effect, imbues the field with its own existence, accounting for the presence of non-zero electromagnetic fields, for the propagation of free photons. The Lagrangian $\mathcal{L}_{\text{kinetic}} = \mathcal{L}_{\text{Maxwell}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$, with the gauge field $F_{\mu\nu}$ defined as $F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu$, gives (source-free) Maxwell’s equations. Putting this all together yields the Lagrangian for the fully interacting theory: $\mathcal{L}_{\text{Maxwell–Dirac}} = \mathcal{L}_{\text{interacting}} + \mathcal{L}_{\text{kinetic}} = \mathcal{L}_{\text{Dirac}} - J^\mu A_\mu - \mathcal{L}_{\text{Maxwell}}$. The inhomogeneous coupled equations of motion for the gauge field (the electromagnetic field) now follow from varying the full action with respect to $A_\mu$.\footnote{The homogeneous field equations follow from the local (gauge) invariance of the action, in fact being just the identities (generalized Bianchi identities) following from Noether’s second theorem.}

Finally, consider that a mass term for the vector field $A_\mu$ of standard form $\mathcal{L}_{\text{photon–mass}} = \frac{1}{2} m^2 A_\mu A_\mu$ is not gauge invariant. In keeping with the demand of local gauge invariance, the vector field $A_\mu$ (i.e. the photon) must then be massless. Local gauge invariance thus necessitates a massless photon.\footnote{This argument generalizes to fields carrying symmetries associated with arbitrary (in particular, non-Abelian) Lie groups, this yielding further interesting gauge structure. The chief difference is that the non-Abelian gauge group has the result that the gauge field ‘generated’ in the gauge argument carries its own charge and is thus self-interacting. Specifically, the associated kinetic term necessarily includes self-interactions in the form of a term proportional to the commutator of the (now, matrix-valued) gauge potentials. This commutator, of course, vanishes in the Abelian, electromagnetic case, this corresponding to the photon not carrying electromagnetic charge. The requirement of local gauge invariance similarly determines the form of the (self-)interactions in the non-Abelian case.}

3.2 Questioning the gauge logic of nature

Despite the heuristic success of the gauge argument in its historical context, one must be a little wary of any attempt to read the logical order of nature directly off the argument, through a straightforward, literal reading. Of course one might resist giving it such a reading in the first place (more in section 4.2). Here, though, we proceed along such lines, noting that this argument is often held out as embodying the purportedly deep physical import of local gauge symmetry principles.
First, the initial and all-important demand of local as opposed to global gauge invariance is anything but self-evident, and presumably it must be argued for on some basis. Historically, the arguments surrounding the ‘demand’ as such are quite thin. The most prevalent form goes back to Yang and Mills’ remarks to the effect that ‘local’ symmetries are more in line with the idea of ‘local’ field theories.\(^{40}\) Arguments from a sort of locality, and especially those predicated specifically on the demands of STR (i.e. no communication-at-a-distance),\(^{41}\) are somewhat suspect, however, and careful treading is needed. Most immediately, the requirement of locality in the STR sense – say, as given by the lightcone structure – does not map cleanly onto the global/local distinction figuring into the gauge argument – i.e. \(G_r\) vs. \(G_\infty\). Overall, the question of how ‘natural’, physically, this demand is, is not uncontroversial. This is especially so in light of the received view of gauge transformations which maintains that they have no physical significance or counterpart (more below). I will return briefly in the next section to considering possible avenues towards underwriting this initial demand.

For now, let us assume that one can provide some sort of justification for the demand of local gauge invariance. We did not consider above the uniqueness of the minimal modification. This modification is not uniquely dictated by the demand of local gauge invariance. There are infinitely many other gauge-invariant terms that might be added to the Lagrangian if gauge invariance were the only input to the argument. In order to pick out the minimal modification uniquely, we must bring in, besides gauge invariance and knowledge of field theories generally, the requirements of Lorentz invariance, simplicity, and, importantly, renormalizability.\(^{42}\) The minimal modification is then the simplest, renormalizable, Lorentz and gauge-invariant Lagrangian yielding second-order equations of motion for the coupled system (O’Raifeartaigh, 1979). The point is simply that, in the context of the gauge argument, the requirement of local gauge invariance gets a lot of its formal muscle in combination with other important considerations and requirements.

One might argue that, at the least, other things held fixed, some requirement of formal simplicity selects the minimal modification as the unique gauge-invariant modification. In this way, the demand of local gauge invariance might be seen as going hand-in-hand with a sort of principle of simplicity. While assumptions of simplicity have certainly proved valuable (even necessary) guides in past theorizing, there is no reason to think that they provide unambiguous, let alone infallible, guides in constructing theories and/or in construing any logical order of nature.\(^{43}\) I suspect that these remarks are not likely to faze a good physicist, who might claim that the

\(^{40}\) Auyang (1995) contains a more developed argument along these lines.

\(^{41}\) See for example Ryder (1996, p. 93).

\(^{42}\) For example, a Pauli term is Lorentz invariant and gauge invariant but not renormalizable.

\(^{43}\) See Norton (2000) for some critical discussion of considerations of simplicity.
argument requires completion and not critique. My point has been only that the demand of local gauge invariance is not the sole input to the gauge argument, nor is it necessarily the most significant in any strong sense.

Setting aside the issue of the uniqueness of the gauge-invariant minimal coupling, another important point is that, in contrast to how it is often portrayed, one does not strictly speaking ‘generate’ a new interaction field in running the gauge argument.44 This gauge field, insofar as it is a physical field, is put in by hand to a large extent. The gauge potentials generated in the gauge argument (in our example, $A_\mu$) form a restricted class of all such $A_\mu$ fields – since we start with a free matter field the potentials are of course all gauge transformable to the zero field. Such potentials, however, correspond to zero $F_{\mu\nu}$ fields.45 An important physical generalization is made then in adding the kinetic term $\mathcal{L}_{\text{Maxwell}}$ for the gauge field to the Lagrangian. The generalization is from a non-physical, formal coupling of the matter field to trivial gauge fields (since $F_{\mu\nu} \equiv 0$) to the physical coupling of the matter field to non-trivial, dynamical gauge fields.46 It is this addition (and the corresponding varying of the full action with respect to the gauge potential(s)) that ‘gives physical life’ to the field. In making this generalization, one puts in by hand much of the important physics of the fully interacting theory.47 I believe that this point goes a long way towards explaining the easily acquired illusion of getting more physics out of the gauge argument than one puts in.

All this is not to say that the requirement of local gauge invariance cannot serve as a useful selection criterion in considering modifications of a free theory towards introducing interactions. It can and has served just such a purpose. As we have seen, historically, there was just such a pragmatic basis for the appeal to local symmetry principles. In any event, it is not how a straightforward, literal reading of the gauge argument might portray it: it is not the case that, unaided, the demand of local gauge invariance (i) dictates uniquely the form of the interacting theory or (ii) strictly speaking dictates the existence of, or accounts for, the origin of a new physical gauge field. In order to pick out the correct form of the theory, other considerations must enter. Also, it is not at all clear that these other considerations or requirements are in any sense inferior, conceptually or physically, to that of local gauge invariance.

44 Brown (1999) has drawn attention to these and related issues. See also the discussion in Teller (2000), Lyre (2001), and Healey (1997).

45 Strictly speaking, this is true only locally. There are potential issues in the case of non-trivial global topology as evidenced, for example, in the familiar Aharonov–Bohm effect; see Nounou (this volume).

46 Another issue worth noting is that one can argue that electric charge does not have a firm physical meaning until the physical electromagnetic field is introduced through the kinetic term.

47 I take it that it is in recognition of essentially this same point that Auyang, considering the gauge principle in electrodynamics, remarks: ‘It does not stipulate an interaction field but rules against its a priori exclusion…’ (Auyang, 1995, p. 58). Brown (1999) discusses the step in which the gauge potential is made properly dynamical as being suggested by physical phenomena such as the Aharonov–Bohm effect as well as by an action–reaction principle.
3.3 Other logics of nature

The line of thinking embodied in the gauge argument presupposes a certain logical order to nature in which gauge invariance is privileged in figuring at the very base of our theories as a sort of axiom. There are, however, other ways of thinking about why our theories are the way they are, and some of these have gauge invariance as more of an ‘output’ than an ‘input’. For example, there are arguments to the effect that various consistency requirements, mathematical and/or physical, require theories of, for example, self-interacting spin-one particles to be of Yang–Mills form with its characteristic group properties.\(^{48}\) Such arguments clearly paint the gauge invariance of physical theory in a different light than does the canonical view.

A more prominent approach that effectively turns the canonical view on its head is that of placing renormalizability (or, alternatively, perturbative unitarity) at the base of fundamental theory. It can be shown that the requirement of renormalizability (respectively, perturbative unitarity) requires that theories have the characteristic form of (spontaneously broken) Yang–Mills gauge theories.\(^{49}\) As renormalizability can be tied directly to a theory’s being well behaved, in the sense that it make sensible predictions for quantities of direct physical interest, one could reasonably argue that gauge invariance is but a feature of the class of well-behaved quantum theories that happen to correctly describe the physics at hand.

Interestingly, renormalizability has itself arguably been superseded in a certain sense. According to the currently prominent effective field theory programme, the familiar renormalizable (quantum) field theories are actually low-energy approximations to some more fundamental underlying theory. Besides the finite number of familiar renormalizable terms, such effective theories (rather, the actions) necessarily contain an infinite number of non-renormalizable terms. However, as long as at high energies there is some well-defined underlying local, dynamical theory (e.g. strings, loops, etc.), then at much lower energies these non-renormalizable interactions will be highly suppressed and thus calculationally insignificant, though, strictly speaking, not physically absent.\(^{50}\) That is, the low-energy ‘residue’ will in fact ‘look like’ a renormalizable theory.\(^{51}\)

---

\(^{48}\) See Wald (1986) and Deser (1970; 1987), Weinberg (1995, chapters 5 and 8), in contrast with the flow of the gauge argument, starts from a quantum theory of massless spin-one fields and arrives at the gauge-invariant coupling to matter.


\(^{50}\) Formally, this result goes by the name of the decoupling theorem. Note that, according to this way of thinking, gravity is just such a non-renormalizable interaction – gravity ‘shows up’ despite being highly suppressed because there are no negative gravitational charges.

\(^{51}\) Relatedly, Wilczek (2000, p. 4) writes that as only asymptotically free theories make good physical sense, and as the only asymptotically free theories in four spacetime dimensions are Yang–Mills theories, ‘the axioms of gauge symmetry and renormalizability are, in a sense, gratuitous. They are implicit in the mere existence of non-trivial interacting quantum field theories’.
Such a view would appear to change completely what we take to be ‘fundamental.’ In light of the effective field theory programme, the appeal of ascribing any deeply fundamental significance to the gauge structure of our theories, especially as resulting from the operations of some deep physical gauge principle, is further diminished.\textsuperscript{52} This gauge structure, it could reasonably be argued, is but a direct consequence of (i) the empirical fact that there exist interacting spin-one particles in nature and (ii) the assumption that our theory of such particles is the residue of some more fundamental underlying theory (for example string theory). Under these assumptions, we arrive at the familiar Yang–Mills gauge theories. So, according to this way of thinking, we might avoid altogether any appeal to gauge symmetry principles in describing the shape or content of current theory. Gauge invariance might be taken as an incidental, albeit interesting feature of the only possible theories at current energies.\textsuperscript{53, 54}

4 The physical content of the gauge symmetry principle

The history and the nature of relevant physics seems to call out for nothing more than a heuristic reading of the gauge argument and of the role of gauge symmetry principles. Yet one might go on to further consider the question as to the physical content of the demand (or featuring) of local gauge symmetry in successful physical theory. Presumably, the hope is in some way to give or find for gauge symmetry principles some physical ‘oomph’. This hope is understandable given the central role often ascribed to these principles and, at the least, their historical role in the development of successful fundamental theories. Also, perhaps with a physically grounded local gauge principle, the gauge argument would be physically underwritten, and possibly construed as something more than a (as it turned out) successful heuristic. In this section, we consider further the received view of gauge invariance and the – I think – associated tension regarding what physical significance to ascribe to gauge symmetry principles. First, I make some brief remarks on some subtleties involved in describing gauge theories.

Aside: On describing gauge theories

Because of the characteristic gauge invariance, describing the basics of formalizing and interpreting gauge theories is a subtle, sometimes tricky business. There are

\textsuperscript{52} Froggatt and Nielsen (1991, chapter 7) discuss a ‘random dynamics’ programme wherein one desideratum is the derivation or explanation of (the usually assumed) Lorentz invariance and gauge invariance from underlying randomness at the fundamental level: effectively, a lack of assumption about what is going on.

\textsuperscript{53} Though symmetry figures prominently into most attempts to go beyond the Standard Model, the search for a new fundamental theory unifying all forces is saddled with the task of determining what, in all likelihood, are radically new ideas and fundamental physical principles governing the theory (see Weinberg, 1999).

\textsuperscript{54} For further discussion of effective field theories and of the importance of scale considerations, see Castellani (this volume, Part IV).
numerous distinct formalisms for describing gauge theories.\textsuperscript{55} Moreover, there are associated interpretational matters to navigate.\textsuperscript{56} One branch of the discussion to receive particular attention from philosophers of physics concerns the physical interpretation of the electromagnetic gauge potential.

Reasonable requirements on the interpretation of a gauge theory in a fixed formalism are that it render the theory deterministic and local in some appropriate sense. Yet indeterminism threatens theories with gauge symmetries since, in the canonical framework, these theories do not possess well-posed initial value problems.\textsuperscript{57} For further discussion, see Earman (this volume, Part I) and Wallace (this volume). This, though, is where interpretation necessarily enters. For example, if, contrary to the norm, one takes the electromagnetic vector potential to represent a physically real, though unobservable, field, electromagnetism will indeed be indeterministic.\textsuperscript{58} The indeterminism owes to the gauge freedom in the electromagnetic potential. We might instead give a gauge-invariant interpretation, taking the physical state as specified completely by the gauge-invariant electric and magnetic field strengths. In this case, electromagnetism is deterministic since the gauge invariance that threatened determinism is in effect washed away from the beginning.

However, in the case of non-trivial spatial topologies, the gauge-invariant interpretation runs into potential complications. The issue is that in this case there are other gauge invariants. So-called holonomies (or their traces, Wilson loops) – the line integral of the gauge potential around closed loops in space – encode physically significant information about the global features of the gauge field. The problem is that these gauge invariants, being ascribed to loops in space, are apparently non-local. But, coming full circle, providing a local description requires appeal to non-gauge-invariant entities such as the electromagnetic potential, whose very reality is in question according to the received understanding.

The context for this discussion is the interpretation of the well-known Aharonov–Bohm (A–B) effect: the manifestation in a particular experimental context of this further gauge-invariant observable, the holonomy, with its apparently non-local nature.\textsuperscript{59} The interpretation of the A–B effect is discussed in more detail by Nounou (this volume). Healey (1997) argues that there is an important analogy to be drawn between the analyses of the (quantum) non-locality evidenced in the A–B effect

\textsuperscript{55} Creutz (1983), for example, distinguishes four formal notions of gauge theory. See also Earman (this volume, Part I).

\textsuperscript{56} Formalism and interpretation are, in some sense, symbiotic concepts, and consequently certain interpretational choices might suggest or be suggested by certain formalisms.

\textsuperscript{57} As mentioned above, there are fewer equations of motion than independent variables. See section 2.1.

\textsuperscript{58} Belot (1998) lays out and critically discusses the various formal and interpretational options here.

\textsuperscript{59} See Anandan (1983) who argues that from both the physical and mathematical points of view, the holonomy contains all the relevant (gauge-invariant) information. Specifically, the connection can be constructed (up to gauge transformation) from a knowledge of the holonomies. Formalizing gauge theories in terms of holonomies associated with (non-local) loops in space appears, though, to require a revamped conception of the notion of a physical field (see Belot, 1998).
and the violation of Bell-type inequalities. Both, Healey takes it, involve possible violations – depending importantly on one’s chosen interpretation of quantum mechanics – of one or the other of two distinct conditions, separability and local action (locality). While acknowledging certain disanalogies, Healey (1999) maintains that similarities in the quantum mechanical explanation of the two phenomena highlights a certain non-separability in the physical world. Local accounts, Healey argues, require non-separable electromagnetic and, more generally, gauge fields. Maudlin (1998) contends that the above analogy depends critically on Healey’s (gauge-invariant) interpretation of gauge theories. Taking the vector potential to represent a real physical field, properly described by only ‘one true gauge’, Maudlin argues that we can provide a local and separable account of the A–B effect. One immediate epistemological contention is that nothing in the physics can reveal this one true gauge that properly represents the real gauge potential. Leeds (1999) defends Aharonov and Bohm’s original interpretation of the A–B effect in terms of a local interaction of matter and a ‘real’ gauge potential, the defence underwritten by appeal to and a near-literal reading of fibre bundles.  

We have certainly not heard the final word on these matters. What the foregoing discussion does indicate (and what should not be all too surprising) is that assessing the physical content of gauge symmetry and gauge symmetry principles is intertwined with a host of other issues – issues surrounding fundamental matters of formalization, interpretation, possibly even epistemological and metaphysical predilection. Let us now continue to consider the general issue of the physical content of gauge symmetry principles.

### 4.1 Gauge invariance: profundity, redundancy, or both?

There are numerous ways of classifying the various symmetries figuring in physical theories: continuous vs. discrete; internal vs. external; global vs. local; physical vs. mathematical; geometrical vs. non-geometrical; geometrical vs. dynamical; universal vs. special, etc.  

The received way of characterizing the domain of gauge symmetries is that gauge symmetry concerns the covariance of the fundamental equations of motion for specific interactions, and that the covariance is tied to a certain descriptive freedom related to the presence of non-physical and, therefore, redundant or ‘surplus’ quantities in the theory. The basic idea is that in describing the physics we introduce too much, and the symmetry under the covariance group effectively rids the theory of the non-physical excess.  

---

60 See also Cao (1995) and Auyang (1995) for related views. See also Nounou’s discussion of the fibre bundle formalism in this volume.

61 Note that not all authors classify the same symmetries in the same ways.

62 This view originated with Dirac. See Castellani (this volume, Part IV).
symmetries can be seen against the backdrop of the more general ‘received view’ of symmetry and invariance and their relation to physical theory, due to E. Wigner. Not only are Wigner’s views lucidly and authoritatively expressed, but they are representative of (even responsible for) the modern view in many respects.

Generally, an invariance principle, according to Wigner, expresses that two systems with the same initial conditions relative to two putatively equivalent coordinate systems develop in the same way, i.e. according to the same physical laws. Understanding the precise import of invariance principles is, however, a trickier business than this. For, as Wigner (and many to follow in his footsteps) notes, there are different types of symmetries and associated invariance principles. This in turn is reflected in a more intricate relationship between invariances, laws, and the physical world. In particular, certain sorts of invariances, such as gauge invariance, do not, according to Wigner, fit this general characterization.

Wigner maintains that there are two qualitatively and fundamentally different sorts of symmetries and associated invariance principles: geometrical and dynamical. At a general level, the primary difference is that the former concern the invariance of all the laws of nature under geometric transformations tied to regularities of the underlying spacetime, while the latter concern the form invariance (i.e. covariance) of the laws governing particular interactions under groups of transformations not tied to spacetime. Both of these types of symmetries posit/embody a certain structure to some set of physical laws in placing restrictions on their possible forms. However, Wigner maintains that the geometrical principles are properly construed as concerning directly the physical events subsumed under these laws. For this reason, he takes it that geometrical invariances are properly physical invariance principles, their being a direct statement about physical events. For example, the geometrical invariance of physical laws under spatial translations is, Wigner says, properly viewed as a statement to the effect that correlations amongst events (i.e. laws) depend only on the relative distances between these events and not on any ‘absolute position’ of the events in the spacetime. The physical features of spacetime (the regularities of its defining structure in fact) ‘undergird’ geometric invariances. Dynamic invariance principles, on the contrary, concern directly the laws of theory and cannot be formulated directly in terms of physical events. Thus, for

---

63 See Wigner (1967b) and the collected works, Wigner (1992); see also Houtappel, Van Dam, and Wigner (1965) and the introduction to this volume.

64 Redhead (1975) continues this symmetry taxonomy, further dividing the dynamical symmetries into those having heuristic potential and those that are merely accidental. The former refers to symmetries used as constraints on theory building and the latter to dynamical accidents having no fundamental physical significance. As Redhead notes, this division is apparently not fixed, there being examples of changes in status – gauge symmetry he takes to be an accidental symmetry which came to be a heuristic symmetry. Kosso (2000) further discusses the distinction between fundamental and accidental symmetries.

65 Pace Wigner, one can bring dynamical invariance principles closer to geometrical invariances in this regard by making appeal to the mathematics of principal fibre bundles. Gauge transformations are then automorphisms of an enlarged geometrical space of sorts. This, though, might be taken to require that one take the bundle structure quite literally at an ontological level. See Leeds (1999) and Cao (1995).
example, the (dynamic) gauge invariance characteristic of electrodynamics is a statement about the specific laws governing electromagnetic interactions and nothing else, Wigner takes it. Briefly stated, there is no regularity in physical events that undergirds such symmetries. The geometrical invariances, concerning as they do directly the physical world, are thus physical in a way that dynamical invariances simply are not.

At root, the above distinction rests on what Wigner takes to be a fundamental difference in the nature of the respective symmetry transformations. The crux of the issue is Wigner’s holding that proper invariance principles (i.e. physically meaningful ones) concern transformations which it is, at least in principle, possible to conceive of actively (Wigner, 1967a, p. 45). In particular, physically meaningful invariance transformations for Wigner are those transformations that can be taken as relating (equivalent) physical observers. This, Wigner takes it, constrains the physically meaningful invariance principles to be the familiar geometric ones associated with the (improper) Lorentz group. Wigner maintains that the transformations figuring into dynamical invariance principles cannot be viewed actively, but rather must be conceived of passively, as mere changes of description.

For Wigner, dynamical symmetry principles are not properly physical since the associated transformations cannot be taken to relate equivalent physical observers and (relatedly) do not characterize any regularity of physical events per se. Rather than relating different physical situations (taken to be equivalent from the point of view of the two observers), they relate redescriptions or re-coordinatizations of one and the same physical situation. So, for example, Wigner (1967a, p. 22) likens appealing to the vector potential and associated gauge invariance to introducing the coordinates of a ghost into our physics and then noting that nothing physical depends on transforming the coordinates of the ghost – a descriptive freedom of no real physical consequence. In the end, Wigner stresses that we should speak only of dynamical ‘invariance’ (or better, perhaps, ‘covariance’) and not dynamical ‘symmetry’: ‘I do not like the idea of gauge invariance being a symmetry principle’ (Wigner, 1984, p. 729).

The basis of this view is straightforward. Wigner’s theory of theories quite sensibly takes observables, specifically probability functions, as fundamental. Laws are in effect nothing but convenient ways of encompassing the various probability distributions for observable outcomes. It immediately follows that dynamical symmetry principles are not physical. The associated transformations change nothing physical since they correspond to the identity transformation on observables. And, since they affect nothing physical, there is no question of viewing the transformations actively, as relating equivalent physical observers.

Redhead (1975) provides a somewhat similar account. Key to the account is the notion of ‘surplus structure,’ essentially mathematics used in the formulation
of some theory that outstrips the physics of the theory. In effect, Redhead gives Wigner’s ghost a name. Redhead considers the familiar distinction between physical and mathematical symmetries, the former a proper subset of the latter. The need for such a distinction arises from the fact that symmetries are not formulation independent. The idea is that mathematical symmetries are features of some particular choice of formal description, whereas physical symmetries are present in all formalisms and hence characterize the underlying physics itself. Within this framework, Redhead characterizes gauge symmetries as non-trivial mathematical symmetries that are, however, physically trivial. Gauge transformations affect only surplus structure and are the identity transformation on that part of the mathematics necessary in representing the actual physics at hand. Redhead (this volume) further discusses these matters, providing further insight into the interesting role of gauge invariance in physical theory.

4.1.1 Received view, received tension

The received view, in effect, presupposes a gauge-invariant formulation. It is not surprising then that gauge transformations are physically impotent, since any potential physical significance of the characteristic gauge symmetry has been washed away from the start. This way of carving out the domain of gauge symmetries, though, seems to pose the following question: how is it that symmetries having to do with a non-physical, formal/descriptive freedom come to have any substantive physical import as it is generally thought gauge symmetries do? In this way, the received view can be taken to sow the seeds that, left unattended, can grow into a sort of tension – a tension between what we might call the ‘redundancy of gauge’ and the ‘profundity of gauge’. On the one hand, as just discussed, we have the view that dynamical symmetry principles are void of physical content, the symmetry transformations being purely mathematical changes of description. On the other hand, as we have also discussed, gauge invariance is often invoked as a supremely powerful, beautiful, deeply physical, even undeniably necessary feature of current fundamental physical theory. Redhead (this volume) calls attention to just this tension, remarking that we may be led to ascribe a ‘mysterious, even mystical, Platonist-Pythagorean role for purely mathematical considerations’.

4.1.2 A generalized Kretschmannian objection

This tension is already familiar, at least in outline, from the wranglings over the physical content of general covariance and its place in the foundations of GTR. I mentioned above Kretschmann’s objection to the ascription of physical content to

---

66 This tension was once lucidly (even if unknowingly) indicated to me by a working particle physicist when in discussion he likened gauge invariance to ‘a phantom that is somehow real’.
general covariance (see Norton, this volume). The tension just discussed might be construed as resulting from what is, in effect, a generalized Kretschmannian objection. The general challenge is to say how it is that the mathematical requirement of covariance can have any deep physical significance. Any theory of appropriate type, the generalized objection might go, can be made gauge invariant (read covariant) if one is willing to add whatever formal/mathematical machinery is necessary to cancel out the effects of the local gauge transformations. Securing gauge invariance is then, it would seem, a mere formal manoeuvre of no physical content. This analogy is perhaps not unexpected since, as already mentioned, general covariance is taken to be the first example of appeal to a local gauge symmetry principle. In the next section, I consider some possible avenues toward addressing this objection and any corresponding tension.

4.2 How gauge principles got their oomph – a Just So Story?

One way to respond to the above tension is simply to deny that there is one. We might argue that we can accept both the redundancy and the profundity aspects of gauge symmetry without worrying about any resulting tension. One might take the historical, heuristic success to be where the only real physical oomph of gauge invariance lies. As discussed above, the role of local symmetry requirements as an attractive selection criterion in arriving at (as it turned out, successful) physical theories should not to be understated. Relatedly, there are the many important affiliations between local gauge invariance and other important, even fundamental, features of physical theory (e.g. renormalizability and asymptotic freedom). Moreover, we might, as is often done, ascribe substantial pragmatic and aesthetic value to the ‘simplicity’ of the gauge paradigm, in that relatively few inputs are required to specify full theories. The fact that all non-gravitational interactions fit into the gauge framework then lends this simplicity to a large part of fundamental physics. This ‘minimalist view’ is not incompatible with taking gauge symmetry as having no direct physical counterpart – as concerning mere formal redundancy. The physical import of these principles is not tied to their (or, rather, the associated transformations’) having direct physical correlates. And, as already discussed, one can reasonably argue that this view is all that the relevant history and physics support.

Perhaps, though, we take it that there is, or that there must be, or that we simply want more than this. One might hold out hope of locating some sort of interesting physical counterpart to the featuring of local gauge invariance, perhaps even a sense

67 Weinberg (1993) discusses the appeal of the structural rigidity of gauge theories.

68 How precisely and to what extent the gauge paradigm provides a true unification of the various interactions has been the topic of some philosophical discussion. See Maudlin (1996) and Morrison (2000).
in which gauge invariance can, contra Wigner, properly be said to be a symmetry. Some have claimed, for example, that gauge transformations can be or, even, must be viewed actively lest the associated forces (interactions) be relegated to fictitious forces (Rosen, 1990). The intuition here is clear enough. It is that physically meaningful symmetries, in agreement with Wigner, must have some sort of physical realization or physical counterpart if they are to be associated with physically meaningful invariance principles.

One possibility already mentioned in the context of assessing the gauge argument was seeing the demand of a local (rather than a global) gauge invariance as accompanying a sort of a ‘locality’ requirement. As I said, though, making precise the connection between the global/local distinction as it figures in classifying gauge transformations and as it does in the spacetime sense is non-trivial in the first place. The fields and transformations ostensibly live in configuration space and need to be ‘brought down’ to spacetime.69 Also, the local vs. global distinction for gauge transformations as commonly employed is itself, in general, dependent on a choice of gauge. This stems from the fact that in order to construe the fields (and transformations) as living in spacetime, one must fix a gauge. This is made clear in treatments of GFTs in the principal fibre bundle framework, where gauge transformations are bundle automorphisms and the typical global/local distinction for gauge transformations is seen to be of little use (Bleecker, 1981).

In any event, it would seem that such a local-gauge-as-locality argument, if it can be mounted, must rely on some sort of active reading of the local gauge transformations. The force of the argument, presumably, derives from the untenability of actually performing the ‘same’ transformation everywhere at the same time (i.e. a global transformation), this violating some notion of locality. But if, as the received view holds, the transformations are taken to be merely passive coordinate relabellings – even ‘ultra-passive’ in the sense that there is no change of physical reference frame but only a change in mathematical description – then it seems that global transformations cannot possibly pose any threat to locality. This is because, by stipulation, there is nothing physical that gets changed under the global transformation. Thus, I do not see how one can mount any argument for local gauge symmetry in the name of locality if one ascribes to the received view of gauge symmetry/invariance.

Perhaps we could take it that the physical significance of the requirement of local gauge invariance lies specifically in that the ‘required’ gauge potential(s) be dynamical. The common definition of what it is to be dynamical in the context of field theory is for the field in question to appear in the action and for it to be varied in deriving equations of motion, including its own. Weinberg, for

69 See Wald and Lee (1990) for further discussion.
example, argues that the real force of the demand of local gauge invariance is carried by the specific requirement that the gauge potential (or the metric connection for gravity) be dynamical in this way (Weinberg, 1996, p. 1). Similarly, Mack (1981) cashes out his construal of the demand of local gauge invariance, which he presents as a principle of no information at a distance (Naheinformationprinzip), in terms of its requiring specifically a dynamical connection/gauge potential to compare (generalized) phases at different spacetime points.

Locating any potential ‘oomph’ of gauge principles in the dynamical nature of the associated gauge fields in this way is consistent with what I indicated above in discussing the gauge argument. There we saw that the physically loaded step in the argument was the investing of the gauge potential with physical life through the addition to the Lagrangian of a kinetic term describing the free electromagnetic field and the subsequent variation of the gauge potential. Recall, though, that this ‘dynamization’ was in large part effected by hand, rather than necessitated by the local invariance requirement. Thus, any argument to the effect that the dynamical gauge field is associated with a, therefore, physical local invariance requirement faces challenges. As I discussed, one can reasonably argue that the local gauge symmetry is rather a by-product or accompaniment of the specific dynamical interaction field(s) in question.

In considering further the ascription of any physical content to gauge invariance, one might seek inspiration from discussions of the physical content and place of general covariance in the foundations of GTR. Why might one do this? I have already mentioned that many take GTR to be a (in fact, the first) gauge field theory, in some sense. More generally, some take it that GTR and gauge field theories are, at the very least, ‘constructed’ in roughly analogous ways according to similar principles of local invariance. Whether there are more disanalogies than analogies here, either in the respective historical contexts or according to current understanding, is debatable.70

Certainly, one appeal to seeing some such analogy here is that it might bring a certain unity or economy of fundamental principles to physical theories which, arguably, are already close in formalism.71 A further promise of this analogy lies in that many have argued that, in the case of GTR, there are avenues available towards ascribing to general covariance a non-trivial physical content, thereby addressing the Kretschmannian objection.72 The hope might be that we could provide some sort of analogous rejoinder to a generalized Kretschmannian objection facing the gauge principle in non-gravitational gauge theories.

70 See note 26, above.
71 Moriyasu (1978), Mack (1981), and Lyre (2001) speak of ‘gauge principles of equivalence’ – essentially, the ability to chose a gauge (cf. coordinate system) such that the gauge potential (cf. metric connection) vanishes in some neighbourhood (generally, it vanishes only at a point). Mack (1981) further discusses gauge invariance as a generalized principle of general covariance.
72 See the references in note 9, above. See also Rovelli (1991; 1997).
Taking such a tack is somewhat complicated by the fact that there is no universally accepted view concerning what physical content, if any, is to be ascribed to the demand of general covariance. Typically, physical content is ascribed to the (active) diffeomorphism invariance of GTR rather than to the (passive) general coordinate invariance of the theory. Though the two types of transformations can be identified with one another mathematically, symmetry under the respective transformations can have different physical significances. Most immediately, active diffeomorphism invariance precludes absolute spacetime objects, such as, for example, the non-dynamical Minkowski metric of STR. As we discussed above, the metric is dynamical in GTR, encoding, besides the metrical properties of spacetime, also the features of the gravitational field itself.

It is clear that, in its historical context, coming to terms with this active diffeomorphism invariance of GTR engendered a major revision in our physical conception of spacetime: it was not a fixed, non-dynamical background arena as was previously thought. Relatedly, spacetime reference systems are, themselves, thoroughly dynamical, part of the physics, and spacetime points are individuated only via the dynamical, physical fields (including, importantly, the metric). These profound realizations about the nature of spacetime (and/or gravitation) go hand-in-hand with identifying and understanding a certain redundancy in the theory: the redundancy that is washed away by the active diffeomorphism invariance characteristic of GTR. Ridding the theory of this redundancy, through physically identifying models related by active diffeomorphisms, is crucial to securing a deterministic theory.73 And this identification, at least, in historical context, enjoins a revamped notion of spacetime. Such invariance is non-trivial physically: the profound realization concerns the physical consequences of this redundancy (and corresponding (active) invariance) for what spacetime (and spacetime reference systems) are not, namely absolute, or non-dynamical. In this way, redundancy and profundity go hand-in-hand in the case of diffeomorphism invariance and its role (particularly, its historical role) in the formulation of GTR.

There are clearly many historical disanalogies between the two cases, and attempting to construct some sort of analogous account about the figuring of gauge invariance in gauge field theories is strained. Even if one can (contra Wigner) in some way sensibly speak of invariance under active gauge transformations – say, for example, as automorphisms of a reified bundle space – it is not clear that this invariance could necessarily be parleyed into anything interesting physically, as in the case of diffeomorphism invariance and GTR. Though not part of the historical discussion, one possible avenue is to view the physical content of local gauge

73 This is the central issue in the well-known ‘hole argument’. This identification is often denoted, separately, as ‘Leibniz equivalence’. See Earman and Norton (1987).
Continuous symmetries

symmetry principles in a manner similar to the way some view the physical content of diffeomorphism invariance of GTR: as underwriting the non-existence of certain absolute (non-dynamical) elements in the theory.

In assessing the physical content of gauge symmetry principles, any analogy with the ‘conceptual foundations’ of GTR (or of spacetime theories generally) is perhaps more trouble than it is worth. In addition to general risks faced in attempting to provide such ‘principled’ stories about theories in the first place, I take it that, in the end, any such hypothetical analogy must be navigated very carefully, if for no other reason than the entirely different historical and theoretical contexts of general relativity and of gauge theories respectively. In the end, how far any such analogy can (or should) be taken, even forgetting the obvious historical disanalogies, is not immediately clear. The promise of such an analogy, though, is presumably that it may yield insight into seeing both profundity and redundancy in the figuring of gauge invariance in successful physical theory, thereby alleviating a certain nagging tension.

References


