Sizing up gauge

John Dougherty

August 7, 2019

Abstract

Gauge theories feature a distinction between "large" and "small" gauge transformations. Small gauge transformations are said to relate two models representing the same physical state, while large gauge transformations are said to relate distinct physical states of affairs that agree on some or all observables—i.e., to be physical symmetries. This size distinction has real physical consequences, but its conceptual status is obscure. Indeed, it seems to be a hypothesis independent of the rest of quantum field theory, though various plausibility arguments for it do exist. In this paper I offer a conceptual justification of the size distinction on the basis of gauge invariance: in a theory with nontrivial gauge transformations, imposing boundary conditions in a gauge-invariant way produces the size distinction. The landscape of gauge theories is therefore richer than it is often taken to be. Two theories can both deny that gauge transformations are physical symmetries without agreeing on all of the gauge-theoretic facts.

1 Introduction

It is widely believed that the standard mathematical formalism of electromagnetism and its generalizations is redundant. This formalism distinguishes a certain class of mathematical transformations with the label "gauge transformation". If two models of the theory are related by a gauge transformation then the physical situations they represent have all the same observable features. More than this, the two models have the same modal profile. Indeed, they have the same "physical profile", in the sense that any differences between the two models are of no physical significance. These facts have led philosophers and physicists alike to conclude that gauge transformations reflect "redundant degrees of freedom" (Pokorski, 2000, 141) and "unphysical symmetry which merely relates different representations of the same physical state or history [and which] could in principle be eliminated" (Struyve, 2011, 226). This interpretation of gauge transformations is so well-entrenched that the term "gauge" is now applied generally to any transformation that sends each model of some theory to another model of the same physical state of affairs (Henneaux and Teitelboim, 1994, 16).

In this paper I confine myself to the specific case of electromagnetism and its generalization to the weak and strong forces and will use the term "gauge" in a more restricted sense.

Contemporary particle physicists seem to disagree with this interpretation of gauge transformations. For in the Standard Model of particle physics it is assumed that some gauge transformations—in particular, the gauge transformations that are nontrivial at infinity—are physical symmetries relating models of distinct states of affairs. That is, the Standard Model features an distinction between two "sizes" of gauge transformations. "Small" gauge transformations, which are the identity at infinity, relate two models of the same physical state of affairs. "Large" gauge transformations, which are nontrivial at infinity, are real physical symmetries. So if two models are related by a large gauge transformation then they represent distinct physical states of affairs. The Standard Model therefore countenances more physical possibilities than a theory in which gauge transformations are interpreted uniformly.

Richard Healey (2007, 2010) has argued that this distinction between different sizes of gauge transformation is conceptually unjustified, that it has generated a theoretical pseudo-problem, and that adopting an alternative interpretation of the theory dissolves this problem. Among the Standard Model's open questions is the "strong CP problem": why does the strong force appear to be symmetric under the combination of charge conjugation and parity inversion? The most popular answer to this question is a dynamical solution proposed by Peccei and Quinn (1977) that hypothesizes a yet-unobserved particle. Experimental searches for this particle are increasingly popular in low-energy particle physics (Irastorza and Redondo, 2018). But if all gauge transformations large and small should be interpreted as purely mathematical features of the theory then this is a wild goose chase. The strong force can only violate CP symmetry if large gauge transformations are real physical symmetries, so if we reject this size distinction then the strong CP problem dissolves. In particular, Healey argues, interpreting the theory in terms of properties attached to spacetime curves shows that we should reject the size distinction and explain away the strong CP problem.

In what follows I argue that the size distinction is justified if we take the structure of gauge transformations to be physically significant. This significance flows from the generally accepted principle that "no physical consequence of a gauge theory can depend on a choice of gauge" (Healey, 2007, xvi). On a strong reading of this principle, the presence or absence of the size distinction in a theory is determined by the presence or absence of nontrivial equivalences in that theory—the transformations that give a precise meaning to the notion of a "choice of gauge". If gauge transformations are "a purely formal feature of a gauge theory's representational framework with no physical significance" (Healey, 2007, xvi) then the ability to choose a gauge is also a formal feature of a theory. In a reformulation of the theory without gauge transformations all constructions are trivially invariant under a choice of gauge, so the gauge transformations in the theory we do have are all on the same trivial footing. In this case there is no way for the size distinction to come

about. Alternatively, if we take gauge transformations to be equivalences then the meaning of "choosing a gauge" changes, and constructions involving the configuration space have to be modified. In particular, in the configuration space for the asymptotically vanishing sector of the theory an equivalence must respect the asymptotically trivial behavior. From this we recover a precise characterization of the size distinction and its justified variants.

In light of this, I argue that Healey's disagreement with the Standard Model is best understood as a choice between two theories of the strong force, rather than as a disagreement over the proper interpretation of a single theory. That is, Healey and the particle physicists are not disagreeing over what a single theory says the strong force is like, they are disagreeing over two candidate theories of the strong force. In particular, they are advancing theories with different configuration spaces: one with trivial equivalence structure and one in which the equivalences are given by gauge transformations. One expression of this structure is the presence or absence of the size distinction. On a "more intrinsic" theory of the kind Healey envisions, which "would not even mention gauge" (2007, 185), the equivalence structure is trivial and there is no size distinction; in the Standard Model there is. Another expression is in the disagreement over the strong CP problem. On a theory without the size distinction CP-violating strong interactions aren't possible; in the Standard Model they are. Yet another expression is in the empirical content of a theory. On a theory without the size distinction the masses of light particles can't be predicted correctly; in the Standard Model they can ('t Hooft, 1986). I argue elsewhere that this makes Healey's envisioned theory empirically inadequate. In this paper, however, I confine myself to the theoretical features of the size distinction, since this is the target of Healey's critique. Confining ourselves to these considerations, we should recognize the equivalence structure of a theory as one of the theory's posits and send the theory to the lab for testing instead of trying to settle the disagreement on theoretical grounds.

The most common theoretical justification for the size distinction, reviewed in Section 2, aims to show that the size distinction is required for the consistency of a particular formulation of electromagnetism and its generalizations. I argue that this justification isn't successful on its own terms. Even if it were, these aren't terms that Healey should accept, since they presuppose the falsity of his view. In Section 3 I offer a precise sense in which Healey's theory and the Standard Model can be said to disagree over the gauge structure of the strong force. This gauge structure is part of the theory's configuration space, and therefore that two theories disagreeing about gauge structure have different configuration spaces. Section 4 shows that if we take gauge transformations to be equivalences the size distinction can be justified, and if we follow Healey in eliminating all talk of gauge then the size distinction cannot arise. We can therefore justify the size distinction without presupposing that Healey's view is false. Healey is right to say that the size distinction, and therefore the strong CP problem, disappears in his theory of the strong force. But this disappearance is

unrelated to the fact that his theory is formulated in terms of properties attached to curves in spacetime; it is only in virtue of the fact that his theory has trivial equivalence structure (Section 5). The disappearance of the strong CP problem is a virtue, but—as I argue elsewhere—it is won at the cost of empirical inadequacy.

The following discussion keeps a narrow focus on Healey's disagreement with the Standard Model, but the point is broader. I take the following discussion to generalize in the same way that the term "gauge transformation" has been generalized. The disagreement between Healey and the Standard Model is an instance of a more general disagreement, with Healey playing the role of the philosophical orthodoxy and Gordon Belot (2018) on the side of the physicists. Both sides agree that we ought to treat many other mathematical transformation in some theory in the same way we treat gauge transformations in electromagnetism and its generalizations.[1] This includes, *inter alia*, global spacetime shifts in Newtonian gravitation theory, Poincaré transformations in special relativistic theories, diffeomorphisms in general relativity. The structure of these transformations can have physical consequences in the same way as gauge transformations in the narrow sense. In particular, imposing boundary conditions in a gauge-invariant way reproduces the size distinction in spacetime theories whose importance Belot has stressed.

2 An unsuccessful argument

Healey's argument against the size distinction has a negative and a positive component: the arguments in favor of the size distinction are insufficient, and a commitment to widely-accepted invariance principles makes the size distinction unjustified. This section addresses the negative component, and the next three deal with the positive component. In the physics literature is often said of the size distinction that "[w]hile some plausible arguments can be given in support of this hypothesis... in the end we must recognize it as an assumption" (Jackiw, 1980, 665). In this section I review the most popular argument of this kind and the best response available on a view like Healey's. I agree with Healey that the argument presented in this section is unsuccessful. Indeed, any argument of this kind, which tries to arrive at the size distinction without postulating it or something stronger, is doomed to fail.

Let us fix notation. In the rest of this paper we will be concerned with Yang–Mills theories on four-dimensional Minkowski space $\mathbb{M}^4$ and its open subsets. Each Yang–Mills theory has a structure group G, a Lie group with Lie algebra $\mathfrak{g}$.[2] We are particularly interested in the case of $G = SU(3)$, since this is the

[1] Belot (2018, fn. 27) says that "some see something like an emerging consensus among philosophers of physics around" the view that all isomorphisms ought to be treated as gauge transformations and eliminated, citing Baker (2010), Greaves and Wallace (2014), and Teh (2016) as observers of this consensus. Belot disagrees with this consensus, arguing that it identifies too many distinct states of affairs. But he agrees that gauge transformations and spacetime symmetries ought to be treated uniformly and provides counterexamples of both kinds.

[2] For the most part my notation follows Baez and Muniain (1994). I will assume that the Lie algebra of G is a direct sum of commuting compact simple and $\mathfrak{u}(1)$ subalgebras so that the Lagrangian is positive definite (Weinberg, 1995, §15.2). This assumption implies that G is a matrix group—i.e., that it has a faithful, finite-dimensional representation—so we assume without

theory of the strong force, but this case is helpfully contrasted with the case $G = U(1)$ of electromagnetism; I will assume that all Yang–Mills theories should be treated uniformly. A field configuration in a region U can be specified by a $\mathfrak{g}$-valued one-form A on U. We fix, once and for all, some set of coordinates on $\mathbb{M}^4$, so that a $\mathfrak{g}$-valued one form on U is equivalently a quadruple of smooth functions $A_\mu : U \to \mathfrak{g}$, corresponding to the one-form $A_\mu \, dx^\mu$. For any $\mathfrak{g}$-valued one-forms A_μ and A'_μ on U, a gauge transformation from the former to the latter is a smooth function $h : U \to G$ satisfying

$$A'_\mu = hA_\mu h^{-1} + h\partial_\mu h^{-1}$$

To each configuration A_μ is associated a field strength tensor $F_{\mu\nu}$, defined by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$$

where $[-,-]$ is the Lie bracket of $\mathfrak{g}$. We adopt the convention that each field configuration and its associated field strength tensor have the same decorations. So, for example, if A_μ and A'_μ are two $\mathfrak{g}$-valued one-forms on a region U then they have field strengths $F_{\mu\nu}$ and $F'_{\mu\nu}$, respectively.

The argument of this section appeals to the Dirac–Bergmann analysis of equivalence, which is set in the constrained Hamiltonian framework.[3] This argument applies to both abelian and nonabelian Yang–Mills theories; for simplicity we will consider the case of electromagnetism and take the gauge group to be $U(1)$. Taking π^μ to be the canonical momentum of A_μ, the canonical Hamiltonian density of the theory is

$$H = -\frac{1}{2}\pi_i\pi^i - \frac{1}{4}F_{ij}F^{ij} + A_0\partial_i\pi^i$$

with one primary and one secondary first-class constraint:

$$\pi^0 = 0 \qquad\qquad \partial_i\pi^i = 0 \qquad\qquad \text{(on shell)}$$

On shell we have $\pi^\mu = -F^{0\mu} = E^\mu$, meaning that π^i is proportional to the electric field. So, on shell, the second constraint implies Gauss's law, $\partial_i E^i = 0$. Call the first constraint the momentum constraint and the second the Gauss constraint.

According to the Dirac–Bergmann analysis, first-class constraints generate equivalences in the following

loss of generality that G is a group of matrices.

[3]The *locus classicus* for the constrained Hamiltonian formulation is Henneaux and Teitelboim (1994); see their exercise 19.4 for their treatment of Yang–Mills theory. Earman (2003) gives a philosophical overview of the constrained Hamiltonian formalism and argues for the importance of the Dirac–Bergmann analysis of gauge described in the next paragraph. I follow Pitts's (2014) conventions for the Hamiltonian theory, accounting for the fact that I am treating A_μ and $F_{\mu\nu}$ as $\mathfrak{u}(1)$-valued.

sense.[4] For any smooth function f on phase space the Poisson bracket induces a vector field $\{-, f\}$, and the flow of this vector field is a continuous family of maps from phase space to itself. According to the Dirac–Bergmann analysis, if f is a first-class constraint then each of these maps is an equivalence: it sends every point of phase space to another point representing the same physical state of affairs. Smearing the constraint by an arbitrary weight merely reparametrizes the flow, and so by the same token smeared first-class constraints must also generate equivalences.

We might run an argument for the size distinction in this framework by showing that any gauge transformation generated by the Gauss constraint must be trivial at spatial infinity.[5] The Gauss constraint is a first-class constraint. Smearing it with a spacetime function ϵ, we have the following Poisson bracket

$$\begin{aligned}\delta A_\mu(t,x) &= \left\{ A_\mu(t,x), \int d^3y\, \epsilon(t,y)\, \frac{\partial}{\partial y^i}\pi^i(t,y) \right\} \\ &= -\eta^i{}_\mu \frac{\partial}{\partial x_\mu}\epsilon(t,x) + \eta^i{}_\mu \int d^3z \int d^3y\, \frac{\partial}{\partial y^i}(\epsilon(t,y)\,\delta(x-z)\,\delta(y-z))\end{aligned}$$

with $\eta_{\mu\nu}$ the Minkowski metric. This expression is generally ill-defined, because it involves the product of two delta functions. We can rectify this by imposing some conditions on ϵ so that the singular support of $\epsilon(t,y)\,\delta(x-z)$ is disjoint from the singular support of $\delta(y-z)$ and their pointwise multiplication is therefore well-defined. Since the integral is a total divergence it suffices to demand that ϵ fall off sufficiently quickly at spatial infinity. This means that any transformation generated by the smeared Gauss constraint will be trivial at infinity. Since first-class constraints generate equivalences, any equivalence will be trivial at infinity. Hence we have the size distinction.

The case of electromagnetism demonstrates a more general problem with trying to formulate a criterion of equivalence within some general framework. Maxwell's electromagnetism, formulated in terms of an electromagnetic potential, is a perfectly respectable theory. Within this theory we can take every gauge transformation to be an equivalence. If it turns out that in the constrained Hamiltonian framework only small gauge transformations are equivalences then our version of Maxwell's electromagnetism just can't be formulated in the constrained Hamiltonian framework. So much the worse for that framework. If a formal criterion says that large gauge transformations can't be equivalences then it can't apply to Maxwell's electromagnetism, and that makes it a bad criterion. And the same remarks hold, *mutatis mutandis*, for nonabelian Yang–Mills theories. A general analysis of equivalence has to allow for a theory in which all gauge transformations, large or small, can be equivalences.

[4] The claim originates with Dirac (1964), and is regularly repeated; for example, Henneaux and Teitelboim (1994, §1.2) claim that any first-class primary constraint generates an equivalence, and that outside of exotic cases so does any first-class secondary constraint. See Pitts (2014) for many other examples.

[5] The basic idea of this argument is usually credited to Balachandran (1994, §4); for philosophical discussions see Struyve (2011, §6) and Teh (2016, §4.1).

The Dirac–Bergmann argument might make us think that a theory featuring the size distinction is interesting or its consequences worth investigating. But it doesn't show that an interpretation of gauge transformations like Healey's is inconsistent or inadequate or that it is already part of Yang–Mills theory. More generally, any analysis of equivalence has to accommodate the fact that we can take Maxwell electromagnetism to be a theory in which large gauge transformations are equivalences. Any general account of equivalence needs to recognize that it is up to me what the equivalences in my theory are. You can tell me that my theory is inconsistent or empirically inadequate or bad in some other way, but you can't tell me my equivalences.

3 Gauge fixing

Healey's positive argument against the size distinction is based on his "unifying principle" that "no physical consequence of a gauge theory can depend on a choice of gauge" (2007, xvi). I propose to take this principle as a point of departure. If we are to draw any precise conclusions from the claim that large gauge transformations are merely symmetries and not equivalences, we need some positive account of equivalence. The Dirac–Bergmann analysis of equivalences in terms of first-class constraints in Section 2 gives the kind of positive account required, but it's insufficiently expressive: it doesn't permit large gauge transformations to be equivalences, thereby ruling out Healey's view by assumption. In this section I argue that the equivalence structure of a theory makes a difference to what it means to "choose a gauge" and therefore to the demand that the physical consequences of a gauge theory be independent of such a choice. We can use this difference to articulate Healey's disagreement with the Standard Model and to justify the size distinction in the latter.

Calculations in Yang–Mills theory routinely make use of the freedom afforded by gauge transformations. If two configurations are related by a gauge transformation then they correspond to the same physical state of affairs. We are thus free to choose from among many representations of any given physical state of affairs, and we can exploit this freedom of choice for practical purposes. Of course, nothing of any physical consequence can depend on this choice. In particular, any physically significant constructions involving the configuration space must also be invariant under a choice of representative. So, for example, a choice of representative should make no difference to the space of solutions to the Yang–Mills equation, or the space of configurations that vanish in some region, or any other such space. A choice of equivalences tells us which field configurations are interchangeable.

This freedom of choice is used more systematically when "choosing a gauge". This allows us to make assumptions about mathematical features of a gauge potential without making assumptions about its physical features. For example, we can assume without loss of generality that the timelike component of every configuration of the Yang–Mills field vanishes. We can do this because for any configuration A_μ the

configuration

$$A'_\mu = A_\mu + h\partial_\mu h^{-1} \qquad\qquad h(t,x) = \mathcal{P}\exp\left(\int_0^t ds\, A_0(s,x)\right)$$

satisfies $A'_0 = 0$ and is related to A_μ by the gauge transformation h on the right, where $\mathcal{P}\exp$ is the path-ordered exponential. So if we ever encounter a configuration whose timelike component doesn't vanish we can apply a gauge transformation to obtain a configuration whose timelike component does vanish. And anything physically significant is invariant under gauge transformations, so applying this gauge transformation makes no physical difference. But we can't demand that every component of the configuration vanish, because not every configuration is equivalent to the vanishing potential. A gauge condition is only legitimate if every configuration is equivalent to one that satisfies it.

Choosing a gauge can restrict the form of permissible gauge transformations, though it cannot eliminate them entirely. Every configuration A_μ is equivalent to some configuration A'_μ with $A'_0 = 0$, but this A'_μ isn't unique. For any smooth function $h : \mathbb{M}^4 \to G$ with no time dependence the configuration

$$A''_\mu = A'_\mu + h\partial_\mu h^{-1}$$

will also satisfy $A''_0 = 0$, and it's equivalent to A'_μ and thus also A_μ. Choosing a gauge is the physicist's idiom for the mathematician's phrase "without loss of generality". We can assume that $A_0 = 0$ because, up to gauge equivalence, this is true of every gauge potential. This assumption does not eliminate any facts about gauge potentials; in particular, it does not eliminate any gauge transformations between them.

Identifying interchangeable configurations is only part of what equivalences are for. Anything physically significant about a gauge theory must be invariant under a choice of gauge. In particular, no physically significant construction involving the configuration space can distinguish between it and any of its gauge-fixed subspaces. For example, consider the construction that sends the configuration space of electromagnetism to its space of vacuum solutions. One model for this solution space is the space of $\mathfrak{u}(1)$-valued one-forms A_μ satisfying the vacuum Maxwell equation:

$$\partial_\mu A_\nu - \partial_\nu A_\mu = 0$$

But, if we like, we could instead start with the gauge-fixed subspace of configurations A_μ that satisfy $A_0 = 0$, the Weyl gauge condition. The space of configurations satisfying both the vacuum Maxwell equation and the Weyl gauge condition is a subspace of the space of solutions to the vacuum Maxwell equation. And it's a strict subspace, since there are configurations that satisfy the vacuum Maxwell equation and not the Weyl gauge condition. So there's a sense in which the space of solutions to the vacuum Maxwell equation depends

on a choice of gauge: it is smaller if we impose the Weyl gauge condition than if we don't. But there's a more important sense in which imposing the Weyl gauge condition makes no difference. The space of solutions to the vacuum Maxwell equation that satisfy the Weyl gauge condition is a gauge-fixed subspace of the space of solutions to the vacuum Maxwell equation. Any solution is gauge-equivalent to some solution that also satisfies the Weyl gauge condition. Since there is no physically significant difference between a space and any of its gauge-fixed subspaces, imposing the Weyl gauge condition makes no physically significant difference.

The process of imposing a gauge condition and forming a gauge-fixed subspace is really about equivalences, not gauge transformations. A gauge-fixing condition is legitimated by the fact that the mathematical differences between equivalent configurations are of no physical significance. For example, if large gauge transformations were not equivalences then we could not impose the Weyl gauge condition. For every potential A_μ there is a gauge transformation g sending A_μ to a potential A'_μ such that $A'_0 = 0$, as noted above. But this h will be large, in general, since A_0 could be nontrivial at infinity. Therefore, most potentials will not be interchangeable with a potential satisfying the Weyl gauge conditions. So the process of choosing a gauge involves gauge transformations insofar as they are equivalences; it is unrelated to the fact that they happen to be given by smooth Lie-group-valued functions on spacetime. It would perhaps be better to have a name for the process that did not involve the word "gauge". However, given the entrenchment of the terminology and lack of better options (my best proposal is "taking a deformation retract"), I will continue to call this process "choosing a gauge".

I will stipulate that nothing of any physical significance can distinguish between a space and any of its gauge-fixed subspaces. That is, I will take Healey's principle that "no physical consequence of a gauge theory can depend on a choice of gauge" (2007, xvi) to fix the meaning of "equivalence", through the role of equivalence in the process of choosing a gauge. This principle subsumes the principle that two equivalent configurations are physically indistinguishable, because any quantity or construction that distinguishes between two equivalent configurations A_μ and A'_μ will also distinguish gauge-fixed subspaces in which only A_μ appears from gauge-fixed subspaces in which only A'_μ appears. But it says more than this, because it also places restrictions on constructions involving configuration spaces. So, for example, it requires that spaces of solutions be invariant under a choice of gauge in the sense illustrated above using electromagnetism and the Weyl gauge: the construction of the space of solutions should not distinguish between the configuration space and any of its gauge-fixed subspaces.[6]

[6] Officially the stipulation is the following. The configuration space of a Yang–Mills theory is naturally a groupoid in which objects are configurations and arrows are gauge transformations. For any groupoid X, say that a groupoid X' of X is a "gauge-fixed subgroupoid" if the inclusion functor $X' \to X$ is full and faithful. Say that a map of groupoids $f : X \to Y$ is an "equivalence" if forcing every inclusion functor of every gauge-fixed subspace to be an isomorphism would also force f to be an isomorphism—i.e., if f becomes an isomorphism in the localization of the category of groupoids at the inclusions of gauge-fixed subgroupoids. This terminology is compatible with the usual meaning of "equivalence" of groupoids. A construction involving groupoids may only be physically significant if it is a functor that preserves equivalences. The discussion of this section and the

Equivalences between configurations thus have an "internal" and an "external" role. Within a particular configuration space a choice of equivalences is a claim about which mathematical configurations represent the same physical state of affairs. When comparing two configuration spaces the equivalences within each determine whether one is a gauge-fixed subspace of the other. These roles for equivalence are motivated by the way that gauge transformations are used in Yang–Mills theory, but now that we have explicitly identified and abstracted them there is a substantive question about whether gauge transformations are equivalences in this sense. In the next two sections I argue that we can understand Healey's disagreement with the Standard Model as a disagreement over the equivalences in a Yang–Mills theory.

4 The size distinction from gauge invariance

If we take the equivalences to coincide with the gauge transformations in the configuration space of Yang–Mills theory over Minkowski space then the size distinction obtains in the sector of the theory containing potentials that vanish at infinity. Since quantum field theory computations are almost always conducted in this sector, the size distinction is almost always in effect. The size distinction in this sector is a consequence of the principle, made precise in the previous section, that any construction must not distinguish between a space and its gauge-fixed subspaces. If we take gauge transformations to be equivalences then the size distinction follows, and if we suppose the equivalence structure to be trivial then there is no size distinction.

For the sake of simplicity I will adopt an oversimplified notion of "asymptotically trivial". Fix, once and for all, some compact subset K of $\mathbb{M}^4$. Rather than requiring the configuration to be trivial at infinity, we will require the configuration to be trivial outside of K. The size distinction will arise in this sector for the same reason that it arises in the asymptotically trivial sector of the theory, but requiring triviality on the submanifold $\overline{K} = \mathbb{M}^4 \setminus K$ instead of at the point at infinity simplifies the geometry of the question.

The space of configurations that vanish outside K is physically significant. It is of physical interest for various reasons. For example, in classical Yang–Mills theory it's often assumed that the configuration is compactly supported when varying the classical action, as in Noether's theorem or in the derivation of the Yang–Mills equation of motion (e.g., Schwartz, 2014, 31). In the quantum theory the space of compactly-supported configurations is the space over which the path integral is performed and the space whose algebra of observables is represented (Weinberg, 1995, Ch. 23, Strocchi, 2013, §8.2). The details of this space can therefore make a difference to the classical and quantum theories.

Since the space of K-supported configurations is physically significant, it must not distinguish between a space and any of its gauge-fixed subspaces. The naive approach to forming this space violates this condition.

next can therefore be naturally situated in categorical treatments of gauge theories such as those of Dougherty (2017), Nguyen et al. (2018), and Weatherall (2016).

Naively, the space of configurations supported in K is just the space in which an element is a configuration A_μ such that $A_\mu = 0$ outside K and in which an equivalence is a gauge transformation. But this can't be a physically meaningful space, because it depends on a choice of gauge. For example, suppose that we fix a gauge in which A_0 is some nonzero constant v. Then no configurations satisfy $A_\mu = 0$ outside K, since they all satisfy $A_0 = v$ everywhere. Applying the naive subspace construction gives the empty space when applied to this gauge-fixed subspace and gives a nonempty space when applied to the unfixed subspace, so it's not invariant under gauge fixing.

Clearly something is defective about this construction of the K-supported sector; we should have a principled way of fixing it. This failure isn't surprising, perhaps, since the construction involves the equivalence-violating condition $A_\mu = 0$. But we needn't object to every violation of equivalence, since some of these are just part of choosing a gauge. If we want to fix the naive construction, rather than just abandoning it for some other construction entirely, we need to identify just where the equivalence violation happens in the naive construction. And we should want to fix it. In this case it's not hard to come up with other equivalence-respecting constructions that we might plausibly take to formalize something like "the space of K-supported fields". But the failure of naive constructions involving spaces is a generic phenomenon, so we need a principled way of dealing with this kind of breakdown. Furthermore, equivalence violations are not always so blatant as the one in the previous paragraph, so we should like a method for ensuring that equivalences are respected.

Replacing the equivalence-violating condition $A_\mu = 0$ with a condition that respects equivalence may not fix the naive construction; this is an illustration of the more subtle ways that a construction might distinguish between a space and its gauge-fixed subspaces. Rather than taking the physical subspace to be the subspace consisting of configurations such that $A_\mu = 0$ on the nose outside K, you might think that the physical subspace consists of those configurations such that A_μ is gauge equivalent to 0 outside K. That is, you might think that the physical subspace consists of those configurations such that there is a smooth function $h : \overline{K} \to G$ such that

$$0 = hA_\mu h^{-1} + h\partial_\mu h^{-1}$$

at every point of $\overline{K}$. Unlike the naive subspace construction, this more sophisticated construction is invariant under gauge fixings of the space of configurations on $\mathbb{M}^4$, and if we take the equivalence structure to be trivial we are done: this construction is invariant under all gauge-fixed subspaces.

However, if we take the theory's equivalences to be the gauge transformations then this construction still depends on features of the condition "$A_\mu = 0$ outside K" that aren't invariant under equivalence. This condition makes reference to a particular configuration on $\overline{K}$, the vanishing configuration. For this

construction to be independent of the choice of representative for this configuration, an equivalence in the region outside K must induce an equivalence in the space of configurations that are trivial outside K. That is, the construction of the physical subspace must send gauge transformations on $\overline{K}$ to gauge transformations in the physical subspace of K-supported configurations. But a smooth function $h : \overline{K} \to G$ doesn't induce a gauge transformation in this attempt at a physical subspace construction. For this we would need some determinate way to extend the function h to all of $\mathbb{M}^4$, and there isn't one: h may not be extendible to $\mathbb{M}^4$, and when it is the extension won't be unique. If gauge transformations and equivalences coincide, then for the subspace construction to be independent of the choice of 0 in the expression "$A_\mu = 0$ outside K" it must be invariant under gauge transformations of this choice, but the construction proposed in this paragraph is not.

In order to make the physical subspace construction invariant under gauge fixing when there are nontrivial equivalences we must keep track of the witness to the configuration's vanishing outside of K. The corrected physical subspace of configurations that smoothly vanish outside K is the space of pairs (A_μ, h), where A_μ is a $\mathfrak{g}$-valued one-form on $\mathbb{M}^4$ and $h : \overline{K} \to G$ is a smooth function such that

$$0 = hA_\mu h^{-1} + h\partial_\mu h^{-1}$$

outside K. An equivalence in this space sending the pair (A_μ, h) to the pair (A'_μ, h') is a gauge transformation $k : \mathbb{M}^4 \to G$ sending A_μ to A'_μ such that $h'k = h$ outside K. This construction commutes with gauge fixing: imposing a gauge condition on the space of $\mathfrak{g}$-valued one-forms and then forming the space of K-supported configurations in this way gives the same result as forming the space of K-supported configurations and then imposing the gauge condition. More generally, replacing the space of configurations on $\mathbb{M}^4$ or the space of configurations in $\overline{K}$ with any equivalent space leaves the results of the construction invariant up to equivalence.[7]

The size distinction therefore follows from taking the equivalences to be the gauge transformations. It is the requirement that a gauge transformation is only an equivalence in the space of K-supported configurations if it preserves the boundary data in the pair (A_μ, h). Consider two configurations (A_μ, h) and (A'_μ, h) in the space of K-supported configurations whose second entries are equal. An equivalence sending (A_μ, h) to (A'_μ, h) is a gauge transformation $k : \mathbb{M}^4 \to G$ from A_μ to A'_μ such that k is the identity outside K. In other words, if we restrict attention to a family of K-supported configurations whose second entries are identical

[7]There is a precise sense in which this is the best approximation to the naive construction that does not distinguish between a groupoid and its gauge-fixed subgroupoids. Equipping the category of groupoids with equivalences, as defined in Footnote 6, gives a homotopical category. The naive space of K-supported configurations is the fiber over 0 of the functor sending a configuration on $\mathbb{M}^4$ to a configuration on $\overline{K}$. The construction in this paragraph is the right derived functor of the fiber—i.e., the homotopy fiber or mapping path space (cf. Riehl, 2014, Ex. 6.5.2). My official view, described informally in the body, is that the space of K-supported configurations is the homotopy fiber of this restriction functor over 0. More generally, informally-described constructions like "the space of K-supported configurations" should be formalized by derived functors of the naive formalization.

then a smooth function $k : \mathbb{M}^4 \to G$ is an equivalence only if it is the identity outside K. It follows that if there is a gauge-fixed subspace of the space of K-supported configurations in which the second entry of every pair is identical then we may ignore it and take the equivalences to be gauge transformations that are the identity outside of K. Of course, any gauge transformation is still an equivalence of the first entries, as long as we allow the boundary data to vary, but when we restrict attention to two configurations for which the boundary data is fixed it looks like some gauge transformations are not equivalences.

Everything works out quite differently if we suppose that there is at most one equivalence between any two configurations. If equivalence is a yes-or-no affair, and two configurations are equivalent whenever there is some gauge transformation sending one to the other, then the size distinction doesn't come about. On this theory, a configuration A_μ has support K just in case there is some smooth function $h : K \to G$ such that

$$0 = hA_\mu h^{-1} + h\partial_\mu h^{-1}$$

Because configurations have no nontrivial equivalences the condition "$A_\mu = 0$ outside K" doesn't either. So the space of K-supported configurations in this theory is just the space of $\mathfrak{g}$-valued one-forms A_μ that are gauge equivalent to 0 on K, and an equivalence between two such configurations is a gauge transformation over $\mathbb{M}^4$. In other words, nothing like the size distinction arises.

The size distinction is a consequence of the stipulation that equivalences between Yang–Mills configurations are given by gauge transformations. More carefully: when we restrict attention to configurations that vanish outside of some compact set K, a gauge transformation is only an equivalence if it also preserves the boundary data that witnesses the configuration's vanishing outside of K. This boundary data is required by the demand that no construction distinguish between a configuration space and any of its gauge-fixed subspaces. If we assume instead that there is no interesting equivalence structure—i.e., if we suppose that there is at most one equivalence between any two configurations—then there is no boundary data to preserve, so every gauge transformation is an equivalence. Either way, two configurations in the full theory over $\mathbb{M}^4$ are equivalent just in case they are related by a gauge transformation. Healey's disagreement with the Standard Model concerns the space of configurations whose support is in K and the equivalences in this space.

5 Holonomy is a red herring

In light of the previous section, we should understand Healey's disagreement with the Standard Model to be a disagreement over the equivalence structure of the theory: on Healey's view there are no nontrivial equivalences in the configuration space, while in the Standard Model the equivalences are precisely the

gauge transformations. This reproduces the disagreement over the size distinction, and it takes seriously Healey's claim that a more intrinsic formulation of Yang–Mills theory "would not even mention gauge, and so the issue of its gauge symmetry would not arise" (2007, 185). I take it that on the kind of theory Healey intends there is one mathematical representative for each gauge-equivalence class of potentials, and gauge transformations are eliminated. It wouldn't really be right to say that such a theory is gauge-invariant—the notion of "gauge transformation" would just be absent. In such a theory there would be no equivalences to speak of, and they certainly wouldn't be the gauge transformations, so the size distinction would not arise in the compactly-supported sector.

One problem with this reading is that Healey takes the elimination of the size distinction and subsequent dissolution of the strong CP problem to be an argument for a curve-theoretic formulation of Yang–Mills theory, and the last two sections do not mention curves. However, as I argue in the rest of this section, it's only in virtue of having a trivial equivalence structure that a curve-based theory might avoid the size distinction. Formulating a theory in terms of curves does not by itself eliminate the size distinction; for example, Wu and Yang's theory of nonintegrable phase factors is a curve-based theory that features the size distinction.

Healey's curve-based formulation of Yang–Mills theory is inspired in part by Wu and Yang's claim that the gauge-invariant content of a $\mathfrak{u}(1)$-valued one-form A_μ is entirely captured by the integral of A_μ along curves in spacetime. They claim that the electromagnetic state of the world is exactly captured by the "phase factor"

$$\exp\left(\int_\gamma A\right)$$

assigned to each loop γ in spacetime. They argue for this conclusion by showing that the phase factor lies between the other usual representations of the electromagnetic field. The Aharonov–Bohm effect shows us that the field strength tensor $F_{\mu\nu}$ doesn't tell us all of the electromagnetic facts: even if $F_{\mu\nu}$ vanishes in some region there can be nontrivial electromagnetic phenomena. And equivalent but mathematically distinct $\mathfrak{u}(1)$-valued one-forms A_μ and A'_μ represent the same physical state of affairs, so there is a sense in which they "overdescribe" the electromagnetic facts. Two different phase factors can correspond to the same field strength tensor $F_{\mu\nu}$, and any two $\mathfrak{u}(1)$-valued one-forms A_μ and A'_μ give the same phase factor.

However, Wu and Yang do not study phase factors assigned to loops. They instead study a theory in which group elements are assigned to open paths in spacetime and which has nontrivial equivalence structure. They are motivated to study the theory with nontrivial equivalence structure by the fact that the loop phase factor

> is less easy to use (especially when one makes generalizations to non-Abelian groups) as a

> fundamental concept than the concept of a phase factor for any path... provided that an arbitrary gauge transformation... does not change the prediction of the outcome of any physical measurements. ... [W]e shall call [this] a nonintegrable (i.e., path-dependent) phase factor. (Wu and Yang, 1975, 3846)

This passage has a false implicature. The nonintegrable phase factor is easier to use than the phase factor because the two things give different theories. The nonintegrable phase factor gives a theory that features nontrivial equivalences—and thus the size distinction—and whose configuration space is isomorphic to the space of $\mathfrak{g}$-valued one-forms. By restricting to integrable phase factors assigned to loops one can eliminate equivalences and hence the size distinction. The nonintegrable phase factor carries significantly more information than the loop phase factor.

To be more precise about the differences, consider how you might define phase factors for a Yang–Mills theory without making reference to a Lie-algebra-valued one-form (Schreiber and Waldorf, 2009). For the nonintegrable case, we can take a configuration of Yang–Mills theory with gauge group G in Minkowski spacetime to be an assignment of elements of G to paths in $\mathbb{M}^4$. Call this assignment a "nonintegrable phase factor map"; formally, it is a smooth function H that takes a path in $\mathbb{M}^4$ as an argument and gives an element of G as output, subject to certain compatibility conditions. An equivalence sending a nonintegrable phase factor map H to the nonintegrable phase factor map H' is a smooth function $h : \mathbb{M}^4 \to G$ such that for any path γ from x to y we have

$$H'(\gamma) = h(y)\, H(\gamma)\, h^{-1}(y)$$

As Wu and Yang stipulate, any physical measurements must be invariant under such a transformation. Any $\mathfrak{g}$-valued one-form A on $\mathbb{M}^4$ therefore defines a nonintegrable phase factor map

$$H_A(\gamma) = \mathcal{P}\exp\left(\int_\gamma A\right)$$

We don't require every nonintegrable phase factor map to be defined by a $\mathfrak{g}$-valued one-form, of course; the point of introducing phase factor maps is to show that talk of phase factors does not depend on talk of $\mathfrak{g}$-valued one-forms. Nevertheless, it's true that that every nonintegrable phase factor map comes from a unique $\mathfrak{g}$-valued one-form, and that equivalences between $\mathfrak{g}$-valued one-forms are in bijection with equivalences between the nonintegrable phase factor maps they induce (Schreiber and Waldorf, 2009, Prop. 4.7).

Nonintegrable phase factor maps can be nontrivially equivalent, so they exhibit the same kinds of behavior as $\mathfrak{g}$-valued one-forms. In particular, they have some mathematical features that lack physical significance, and they support gauge fixing. They also feature the size distinction. The mechanism is exactly the same

as in Section 4. The space of G-valued phase factor maps that vanish outside of the compact region K of $\mathbb{M}^4$ is the space of pairs (H, h), where H is a phase factor map sending paths in $\mathbb{M}^4$ to elements of G and $h : \overline{K} \to G$ satisfies

$$1 = h(y)\, H(\gamma)\, h(x)^{-1}$$

for all paths γ from x to y in $\overline{K}$. An equivalence sending a pair (H, h) to a pair (H', h') is a smooth function $k : \mathbb{M}^4 \to G$ such that

$$H'(\gamma) = h(y)\, H(\gamma)\, h^{-1}(x) \qquad\qquad h'k = h$$

for all paths γ from x to y. So, as in the case of $\mathfrak{g}$-valued one-forms, a smooth function $k : \mathbb{M}^4 \to G$ is only an equivalence in the space of phase factor maps vanishing outside of K if it respects the boundary conditions. Since loop phase factor maps have no nontrivial equivalences the boundary conditions impose no constraints.

Healey's loop theory does not feature the size distinction. The main difference between this theory and one formulated in terms of nonintegrable phase factor maps is the lack of nontrivial equivalences equivalences. Two loop phase factor maps are equivalent if and only if they assign the same complex number to any loop. That is, two phase factor maps are equivalent just in case they are equal. This is what makes loop phase factor maps so attractive for a project like Healey's: there are no nontrivial equivalences between phase factor maps, and so equality of phase factor maps as mathematical objects reflects equality of physical configurations. And because there are no equivalences between loop phase factor maps, anything that you do with them will respect equivalences. This means that the space of K-supported loop phase factor maps is just the space of loop phase factor maps that assign the identity to every loop outside of K. There is no difference between large and small gauge transformations on this theory. Indeed, as Healey says, a theory formulated in terms of loop phase factor maps "would neither be, nor fail to be gauge symmetric. It could represent no gauge properties" (2007, 185).

Healey argues that the strong CP problem dissolves if you adopt a curve-based formulation of Yang–Mills theory, because "when a theory is formulated in a loop/path representation, all states and variables are automatically invariant under both 'small' and 'large' gauge transformations" (Healey, 2007, 198). The condition doing the work in this argument is the fact that there are really no gauge transformations in the loop representation, so there can be no distinctions between different kinds. In Wu and Yang's path representation there are gauge transformations, and taking these to be equivalences reproduces the size distinction. So Healey's disagreement with the Standard Model when it comes to the size distinction isn't a disagreement about curves or fields—at least not in the first place—but about the absence or presence of nontrivial equivalence structure. A theory without equivalence structure will not feature the size distinction, and a theory in which gauge transformations are equivalences will. Consequently, a theory without equivalence

structure will dissolve the strong CP problem that a theory with equivalence structure faces.

6 Conclusion

I draw two conclusions from the above discussion. First, the distinction often made between large and small gauge transformations can be justified on commonly-accepted grounds. It is widely accepted that nothing of any physical significance can hang on a choice of gauge. On a strong reading of this principle, this means that any physically significant construction involving the configuration space of a gauge theory must respect equivalence, as defined in Section 3. If we take gauge transformations to be equivalences then the naive construction of the space of K-supported configurations violates this principle. Minimally correcting this construction gives a space of K-supported configurations in which an element is a configuration along with a particular gauge transformation sending this configuration to the vanishing potential outside of K and in which an equivalence must respect the boundary data. This justification also applies in other theories with nontrivial equivalence structure, thereby justifying many of the equivalence judgements that Belot has argued should be accounted for.

Second, we ought to frame Healey's disagreement with the size distinction as a claim about the structure of the Yang–Mills configuration space. On Healey's view gauge transformations serve only to define an equivalence relation on potentials, while on my reconstruction of the Standard Model they are equivalences in a more robust sense. This is a faithful representation of Healey's view, which is explicitly intended to be an interpretation of Yang–Mills theory that eliminates all reference to the notion of "gauge". It also locates his successful avoidance of the strong CP problem in the right place: the curve-based character of Healey's view doesn't dissolve the strong CP problem, since the size distinction holds in Wu and Yang's nonintegrable phase factor formulation just as much as it does in the more traditional field-theoretic formulation. The strong CP problem needn't trouble Healey because his theory lacks any nontrivial equivalences, not because it is formulated in terms of curves. Finally, this framing of the disagreement makes it, in part, an empirical dispute. On the one hand, Healey's view correctly predicts an absence of CP violation in the strong sector. On the other, it cannot support any predictions that make use of the size distinction.

These conclusions generalize. As I indicated in the introduction, the consensus view among philosophers of physics is that we should try to eliminate all gauge transformations and their generalizations. Belot objects to this consensus, citing a bevy of cases along the lines of the size distinction in the Standard Model, where physicists treat generalized gauge transformations as redundancies except when they don't. For example, he notes that "while relativists *do* often speak as if solutions of general relativity are gauge equivalent if and only if isometric, they drop this way of speaking when asymptotic boundary conditions… are in view" (2018,

22). This behavior is unjustified on the consensus interpretation; this is Healey's disagreement with the Standard Model writ large. But it's also hard to account for this behavior if we take gauge transformations to be physical symmetries, since some of them are supposed to be redundancies on any view.

On the view I offer above there is something right about the consensus view, but Belot's reservations are well-founded. The apparent context-sensitivity of physicists' equivalence judgements can be justified by appealing to mechanisms like the one giving rise to the size distinction—i.e., a strong reading of gauge invariance. In particular, if we take diffeomorphisms to be equivalences in general relativity then it will also feature a size distinction when we impose boundary conditions. The consensus view is right to say that we should take these spacetime transformations to have the same status as gauge transformations, and it's right to say that models related by a gauge transformation represent the same physical state of affairs, at least in the most general context. But it does not follow from this that we should try to eradicate the equivalences; they're an important part of the theory.

References

Baez, J. and Muniain, J. P. (1994). *Gauge Fields, Knots and Gravity*. World Scientific.

Baker, D. J. (2010). Symmetry and the metaphysics of physics. *Philosophy Compass*, 5(12):1157–1166.

Balachandran, A. P. (1994). Gauge symmetries, topology, and quantisation. *AIP Conference Proceedings*, 317(1):1–81.

Belot, G. (2018). Fifty million Elvis fans can't be wrong. *Noûs*, 52:946–981.

Dirac, P. A. M. (1964). *Lectures on Quantum Mechanics*. Belfer Graduate School of Science, Yeshiva University.

Dougherty, J. (2017). Sameness and separability in gauge theories. *Philosophy of Science*, 84(5).

Earman, J. (2003). Tracking down gauge: an ode to the constrained Hamiltonian formalism. In Brading, K. and Castellani, E., editors, *Symmetries in Physics*, chapter 8, pages 140–162. Cambridge University Press.

Greaves, H. and Wallace, D. (2014). Empirical consequences of symmetries. *The British Journal for the Philosophy of Science*, 65(1):59–89.

Healey, R. (2007). *Gauging What's Real*. Oxford University Press.

Healey, R. (2010). Gauge symmetry and the theta-vacuum. In Suárez, M., Dorato, M., and Rédei, M., editors, *EPSA Philosophical Issues in the Sciences: Launch of the European Philosophy of Science Association*, pages 105–116. Springer.

Henneaux, M. and Teitelboim, C. (1994). *Quantization of Gauge Systems*. Princeton University Press.

Irastorza, I. G. and Redondo, J. (2018). New experimental approaches in the search for axion-like particles. *Progress in Particle and Nuclear Physics*, 102:89 – 159.

Jackiw, R. (1980). Introduction to the Yang–Mills quantum theory. *Reviews of Modern Physics*, 52(4):661–673.

Nguyen, J., Teh, N. J., and Wells, L. (2018). Why surplus structure is not superfluous. *The British Journal for the Philosophy of Science*, page axy026.

Peccei, R. D. and Quinn, H. R. (1977). Constraints imposed by CP conservation in the presence of pseudoparticles. *Phys. Rev. D*, 16:1791–1797.

Pitts, J. B. (2014). A first class constraint generates not a gauge transformation, but a bad physical change: The case of electromagnetism. *Annals of Physics*, 351:382–406.

Pokorski, S. (2000). *Gauge Field Theories*. Cambridge University Press, 2nd edition.

Riehl, E. (2014). *Categorical Homotopy Theory*. Cambridge University Press.

Schreiber, U. and Waldorf, K. (2009). Parallel transport and functors. *Journal of Homotopy and Related Structures*, 4:187–244.

Schwartz, M. D. (2014). *Quantum Field Theory and the Standard Model*. Cambridge University Press.

Strocchi, F. (2013). *An Introduction to Non-Perturbative Foundations of Quantum Field Theory*. Oxford University Press.

Struyve, W. (2011). Gauge invariant accounts of the Higgs mechanism. *Studies in History and Philosophy of Modern Physics*, 42(4):226–236.

't Hooft, G. (1986). How instantons solve the $U(1)$ problem. *Physics Reports*, 142(6):357–387.

Teh, N. J. (2016). Galileo's gauge: Understanding the empirical significance of gauge symmetry. *Philosophy of Science*, 83(1):93–118.

Weatherall, J. O. (2016). Understanding gauge. *Philosophy of Science*, 83(5):1039–1049.

Weinberg, S. (1995). *The Quantum Theory of Fields*. Cambridge University Press.

Wu, T. T. and Yang, C. N. (1975). Concept of nonintegrable phase factors and global formulation of gauge fields. *Physical Review D*, 12:3845–3857.