A Categorical Perspective on Symmetry and Equivalence

Neil Dewar, Benjamin Eva

September 5, 2017

In recent years, a number of authors have employed the resources of category theory to shed new light on the notion of 'theoretical equivalence'. Here, we connect this work to the debate concerning the interpretational significance of symmetries in physics. Specifically, we put forward a novel criterion to be satisfied by any prospective 'category of models' of a scientific theory. This criterion is motivated primarily by the idea that the semantic representation of a theory should not include any redundant structure, and that redundant structure can be identified by representing the symmetries of a theory by auto-equivalences on the category of models of that theory.

1 Introduction: categories and equivalence

In recent years, a rich literature has developed around the application of the formal resources of category theory to problems arising in the philosophy of science in general and the philosophy of physics in particular. Importantly for our purposes, it has been persuasively argued (see e.g. Halvorson (2012, 2016), Halvorson and Tsementzis (2015), Rosenstock et al. (2015), Weatherall (2016a,b,c)) that

> ...[F]amiliar scientific theories (e.g. Hamiltonian mechanics, special and general relativity, quantum mechanics, quantum field theory) can themselves be fruitfully described as categories. (Halvorson and Tsementzis, 2015, pp. 1–2)

Now, the question of how one should go about achieving such a categorical representation of a given theory is controversial in at least two ways. Firstly, there is the long running debate concerning whether scientific theories are best characterised in semantic or syntactic terms.[1] In the categorical setting, this problem manifests itself in the question of whether the category theoretic representation of a theory should encode the semantic or syntactic structure of that theory. The syntactic approach has been defended and developed by Halvorson and Barrett (see e.g. Halvorson (2012, 2016), **?**), who argue that the 'syntactic category' of a theory is better able to provide a complete specification of the mathematical content of that theory than its semantic counterpart. In contrast, the semantic approach to providing a categorical representation of scientific theories has not yet been given any systematic defence. But this has not stopped authors from invoking purely semantic categorical representations of theories in order to settle contested questions of theoretical equivalence (see e.g. Weatherall (2016a,c)). A salient factor here is that while it seems that we have some intuitive grasp of roughly what the category of models of a scientific theory should look like, it is generally far more difficult to identify a suitable syntactic category for a non-trivial scientific theory. Thus, from a purely pragmatic perspective, the semantic approach has the important virtue that it can in practice be applied to real scientific theories and subsequently yield novel judgments regarding their philosophical interpretation.

But provided one agrees that, in practice, it is more useful to represent a theory by a category that encodes its semantic structure, another controversy arises. Specifically, how do we determine exactly which category is *the* 'semantic category' relative to which questions of equivalence and philosophical interpretation should be settled? Alternatively, how do we identify 'the category of models' of a given theory? In this paper, we aim to make a first contribution towards answering that question. In particular, we will posit a novel criterion to be satisfied by any prospective semantic categorical representation of a theory. This criterion will be motivated primarily by considerations regarding the way in which the symmetries of a physical theory should be represented in category-theoretic terms. We will go on to show how the criterion allows us to settle previously irresolvable questions

[1] For a critical overview of the debate between the semantic and syntactic views of scientific theories, see e.g. Lutz (2015).

concerning the theoretical equivalence of real physical theories.

The structure of the paper is as follows. In section 2 we present a brief introduction to some ways in which one can use category theory to gain perspicuous and philosophically useful representations of the content of scientific theories, focusing particularly on the formalisation of theoretical equivalence. In section 3 we rehearse some relevant details of the philosophical debate on symmetries of physical theories. In section 4 we motivate a general category-theoretic representation of the symmetries of a given theory, and then go on to apply this representation to obtain a novel criterion to be satisified by any categorical representation of a given theory. Intuitively, this criterion requires that the representation should not include any redundant structure. In section 5, we go on to apply the criterion to a contested question of theoretical equivalence in the philosophy of physics. Section 6 concludes.

In what follows, we assume a basic familiarity with category theory up to the level of e.g. chapter 2 of Mac Lane (1978).

2 Semantic Categories and Theoretical Equivalence

The so called 'Semantic view' of scientific theories, defended by e.g. Suppe (1989) and van Fraassen (1980), holds that the characteristic content of a scientific theory is given by a specification of the set of models of that theory. According to the semantic view, the formal language in which a theory is articulated does not play a definitive role in distinguishing that theory from its rivals. Rather, a theory's content is determined entirely by what we take to be the models of the theory. It is not our purpose here to contribute to the debate on whether scientific theories are best characterised in semantic or syntactic terms. We assume throughout that something like the semantic view can be used to isolate the characteristic content of theories, and focus our attention on the way in which we can use semantic representations of scientific theories to gain new insights into issues of philosophical interpretation.

A useful question to ask at this point is 'what, according to the semantic view, does it mean for two scientific theories T_1 and T_2 to be theoretically equivalent?'. Assuming the semantic view, the theories T_1 and T_2 can be faithfully represented by the correspnding sets of models M_1 and M_2, respectively. Clearly, it is not instructive to say that T_1 and T_2 are equivalent

if and only if $M_1 = M_2$, since we want to be able to say that theories which are not obviously identical may nevertheless turn out to be equivalent. This possibility is ruled out if we require that equivalent theories always have the same semantic representation.

A more interesting possibility is to use a model-theoretic variant of a syntactic criterion of theoretical equivalence, such as that offered by Glymour (1977). Specifically, Glymour proposes that T_1 and T_2 should be classed as equivalent if and only if they (i) are empirically equivalent, and (ii) are mutually inter-translatable. Focusing for now on condition (ii), which unlike condition (i) is a purely formal property that can be articulated without reference to any additional interpretive constraints, we need to spell out exactly what it means for T_1 and T_2 to be mutually inter-translatable. Assuming that we are working with first order theories, Glymour invokes the notion of 'definitional equivalence' (the interested reader can see e.g. Glymour (1977), Weatherall (2016a) for details) and notes that if two first order theories T_1 and T_2 are definitionally equivalent, this will induce a bijection between the sets of models M_1 and M_2. Thus, when Glymour attempts to formulate a semantic variant of the criterion of definitional equivalence, he stipulates that T_1 and T_2 can be equivalent theories *only if* there exists a translation which induces bijection between M_1 and M_2, i.e. for T_1 and T_2 to be equivalent, it is a necessary condition that the translation put the models of the two theories in one-to-one correspondence. This provides us with a useful criterion for theoretical equivalence that can easily be applied to the semantic representations of scientific theories. However, it turns out that this criterion gives surprising and unintuitive results when applied to real theories.

Weatherall (2016a) considers two alternative formulations of classical electromagnetism on Minkowski spacetime, which are standardly considered to be equivalent. On the first formulation, EM_1, a model of the theory is given by a specification of a Faraday tensor (antisymmetric rank-2 tensor) F_{ab} and four-current (four-vector field) J^a satisfying Maxwell's equations on Minkowski spacetime M,

$$\nabla_{[a} F_{bc]} = 0 \tag{1a}$$

$$\nabla_a F^{ab} = J^b \tag{1b}$$

(where the Minkowski metric has been used to raise indices). Intuitively,

F_{ab} represents the electromagnetic field on spacetime and J^a represents the distribution and motion of charged matter on spacetime. On the second formulation, EM_2, a model of the theory is given by a specification of a 4-vector potential (1-form) A_a and four-current J^a, satisfying the differential equation

$$\nabla_a \nabla^a A^b - \nabla^b \nabla_a A^a = J^b \tag{2}$$

where ∇ is the Levi-Civita connection on M and (again) we have used the Minkowski metric to raise indices. We may take the empirical content of these theories to be given by the classes of dynamically allowed trajectories, for all possible charge/mass ratios: for a given charge/mass ratio $q \in \mathbb{R}$, a curve $\gamma : I \to M$ (where I is an interval of $\mathbb{R}$) is dynamically allowed in EM_1 just in case it satisfies

$$\dot{\gamma}^n \nabla_n \dot{\gamma}^a = q F^a{}_b \dot{\gamma}^b \tag{3}$$

and is dynamically allowed in EM_2 just in case it satisfies

$$\dot{\gamma}^n \nabla_n \dot{\gamma}^a = q(\nabla^a A_b)\dot{\gamma}^b \tag{4}$$

Weatherall notes that given a 4-vector potential A_a satisfying equation (2), it is always possible to define a corresponding Faraday tensor

$$F_{ab} = \nabla_a A_b \tag{5}$$

satisfying equations (1); moreover, this Faraday tensor will pick out the same dynamically allowed trajectories as A_a. Conversely, given a Faraday potential satisfying (1), there will always exist a corresponding vector potential satisfying (2) and picking out the same dynamically allowed trajectories.[2] However, this correspondence is not bijective. The 4-vector potential uniquely specifies the corresponding Faraday tensor, but a given Faraday tensor will be compatible with many possible 4-potentials: for any 4-potential A_a, the 4-potential A'_a obtained by applying a *gauge transformation*

$$A_a \mapsto A'_a = A_a + \nabla_a \lambda \tag{6}$$

[2]Note that Weatherall (2016a) takes the Faraday tensor itself to constitute the empirical content of the theories, rather than the class of dynamically allowed trajectories. This amounts to the same thing, given that equations (3) and (4) ensure that the Faraday tensor uniquely determines the allowed trajectories, and that the Faraday tensor can be uniquely reconstructed if one knows the dynamically allowed trajectories (see (Misner et al., 1973, Box 3.1)).

where λ is any smooth scalar field on M, will define the same Faraday tensor.

Relative to the translation (5), EM_2 has strictly more models than EM_1. This is because EM_2, unlike EM_1, distinguishes between potentials that are related by gauge transformations. Applying Glymour's criterion that the models of equivalent theories should be put in bijective correspondence by the appropriate translation, we reach the verdict that EM_1 and EM_2 cannot be equivalent. But this seems wrong. For, as Weatherall argues,

> On their standard interpretation, models related by a gauge transformation are *physically equivalent*, in the sense that they have the capacity to represent precisely the same physical situations. Thus, EM_2 does *not* distinguish situations that EM_1 cannot. And indeed, it seems to me that there is a clear and robust sense in which two theories should be considered as equivalent if, on their standard interpretations, they differ only with regard to features that, by the lights of the theories themselves, have no physical content. (Weatherall, 2016a, p. 1079)

So Glymour's equivalence criterion cannot be correct. EM_1 and EM_2 are standardly considered to be equivalent in the sense that they allow for exactly the same physical possibilities, and yet their equivalence is ruled out by Glymour's criterion in light of the fact that the translation (5) does not induce a bijective correspondence between their models. Weatherall notes that this problem does not arise if we replace EM_2 by the theory EM_2' whose models are given not by 4-vector potentials, but rather by equivalence classes of physically equivalent 4-vector potentials. The models of EM_1 and EM_2' *are* in bijective correspondence and so do indeed satisfy Glymour's criterion. But of course, having the same number of models is not by itself sufficient to guarantee theoretical equivalence, even if one thinks it is necessary. We need some additional criteria to ground our judgments of theoretical equivalence. Glymour proposes that we should consider T_1 and T_2 to be equivalent if and only if for every model m_1 of T_1 there exists a unique model m_2 of T_2 that (i) has the same empirical content as m_1, and (ii) is such that the geometrical objects associated with m_2 are uniquely and covariantly definable in terms of the elements of m_1 and vice-versa. We've seen that EM_1 and EM_2' satisfy the first of these conditions.

However, as Weatherall notes, the notion of 'covariant definability' being employed here is left imprecise by Glymour, and it is difficult to motivate any privileged interpretation in the context of the theories of modern physics.

> The theories we are working with here are not first order theories—or at least, we have not specified a first order 'language of electromagnetism' or 'language of geometrized Newtonian gravitation', nor have we presented any axioms of any sort or relied only on proofs in a first order system. Indeed, it is not clear that satisfactory first order theories of electromagnetism or geometrized Newtonian gravitation exist, and more importantly, the question of whether such theories *do* exist does not seem to arise in practice. This suggests that there are, at least in principle, different notions of definability available for these theories, none of which has been made precise in the present context. So it is hard to know how to proceed. (Weatherall, 2016a, p. 1080)

Thus, even if one were to accept that a bijective correspondence of models is a necessary condition for theoretical equivalence, it is difficult to see how this condition could be extended to a full definition that can be of serious use to advocates of the semantic view. In light of these considerations, Weatherall goes on to consider alternative semantic definitions of equivalence. In order to formulate these alternative definitions, we need to enrich the representation of scientific theories given by the standard version of the semantic view.

According to the standard semantic view, a theory is represented by the set of intended models. This representation can be made richer by considering, rather than the bare *set* of models, the category whose objects are models of the theory and whose morphisms are structure-preserving homomorphisms between those models: for instance, in the case of first-order logic, whose morphisms are elementary embeddings between models.[3] This representation includes a great deal of mathematical structure that is neglected by the standard set-theoretic articulation of the semantic view. Importantly, we know that if two first order theories T, T' are definitionally

[3]For an alternative approach to category-theoretic representations of scientific theories, see Nguyen et al. (2017).

equivalent, then their categories of models (henceforth 'semantic categories') are isomorphic.[4] This suggests that, when using semantic representations of theories, we should treat theories T_1, T_2 as equivalent just in case there exists an isomorphism of their semantic categories that preserves empirical content. Applying this new definition of equivalence to the case of classical electromagnetism, Weatherall defines the semantic categories of the theories EM_1, EM_2 and EM_2' as follows,

- $\mathbf{EM}_1$ is the category whose objects are models of EM_1 (i.e. Faraday tensors) and whose arrows are isometries of Minkowski spacetime that preserve the Faraday tensor.
- $\mathbf{EM}_2$ is the category whose objects are models of EM_2 (i.e. gauge potentials) and whose arrows are isometries of Minkowski spacetime that preserve the vector potentials.
- $\mathbf{EM}_2'$ is the category whose objects are models of EM_2' (i.e. equivalence classes of gauge potentials) and whose arrows are isometries of Minkowski spacetime that preserve the vector potentials.

Weatherall proves that $\mathbf{EM}_2'$ and $\mathbf{EM}_1$ are isomorphic as categories, which means that the new definition of theoretical equivalence gives the right answer in this case. However, the proposed translation (5) is not an isomorphism since it does not put the objects of $\mathbf{EM}_2$ and $\mathbf{EM}_1$ into bijective correspondence. And as we've already seen, this seems to be the incorrect verdict, since EM_1 and EM_2 allow for exactly the same physical possibilities. Thus, the ammended definition of theoretical equivalence is also unsatisfactory. We need a weaker condition that allows for the equivalence of EM_1 and EM_2. In order to formulate Weatherall's final definition of theoretical equivalence, we need to recall the notion of an 'equivalence' of categories.

Definition 2.1. *An equivalence of two categories $\mathcal{C}$ and $\mathcal{D}$ is a functor $F : \mathcal{C} \to \mathcal{D}$ such that there exists an "almost inverse" functor $G : \mathcal{D} \to \mathcal{C}$: that is, a functor such that $G \circ F$ is naturally isomorphic to $Id_{\mathcal{C}}$ and $F \circ G$ is naturally isomorphic to $Id_{\mathcal{D}}$. It is helpful to note that F is an equivalence if and only if*

[4] We say that categories $\mathcal{C}, \mathcal{D}$ are 'isomorphic' if there exist functors $F : \mathcal{C} \to \mathcal{D}$, $G : \mathcal{D} \to \mathcal{C}$ such that $G \circ F = \mathrm{Id}_{\mathcal{C}}$ and $F \circ G = \mathrm{Id}_{\mathcal{D}}$. This implies that both the objects and morphisms of $\mathcal{C}, \mathcal{D}$ are in bijective correspondence.

it is essentially surjective (for every object X in $\mathcal{D}$, there is some object A in $\mathcal{C}$ such that $F(A)$ is isomorphic to X), and full and faithful (for any objects A and B of $\mathcal{C}$, the induced map $f \in \mathrm{Hom}(A, B) \mapsto F(f) \in \mathrm{Hom}(F(A), F(B))$ is bijective).

Equivalence of categories is strictly weaker than isomorphism. Crucially, it does not require that $\mathcal{C}$ and $\mathcal{D}$ have the same number of objects, although it does require that they have the same number of isomorphism classes of objects (or that their 'skeletons' are equinumerous). As Weatherall puts it, 'Equivalent categories may be thought of as categories that are isomorphic "up to object isomorphism"' (Weatherall, 2016b, p. 15). Armed with this notion, Weatherall posits the following definition of theoretical equivalence,

Two theories are theoretically equivalent if and only if their exists an equivalence of their semantic categories that preserves empirical content.

Since categorical equivalence is strictly weaker than isomorphism, this definition still gives the desired result that EM_1 and EM_2' are theoretically equivalent. Furthermore, Weatherall shows that, given some slight modifications of the categorical representation of EM_2, this definition allows us to respect the strong intuition that EM_1 and EM_2 should be counted as equivalent theories. Specifically, let $\overline{\mathbf{EM}}_2$ be the category whose object are models of EM_2 and whose morphisms are pairs consisting of a gauge transformation and an isometry of M preserving the gauge-transformed vector potential: we will specify such a morphism by a pair (ψ, λ), where $\psi : M \to M$ is an isometry and λ is a scalar field (which specifies a gauge transformation as in (6). Then it turns out that there does indeed exist a categorical equivalence between $\overline{\mathbf{EM}}_2$ and $\mathbf{EM}_1$ that preserves empirical content (in the sense of taking models to models with the same dynamically allowed trajectories for any charge/mass ratio). So according to the new definition of theoretical equivalence as empirical-content-preserving equivalence of semantic categories, EM_1 and EM_2 are equivalent theories, as desired.

With Weatherall, we take (empirical-content-preserving) equivalence of semantic categories to represent the best available semantic formalisation of the notion of theoretical equivalence. However, it is important to note that in order for this formalisation to be fruitfully applied, there needs to be a

clear methodology for determining the appropriate semantic category of a given theory. In order to achieve the desired verdict that EM_1 and EM_2 are theoretically equivalent, we needed to make a choice about which semantic categorical representation of EM_2 is the right one to use. If we simply use the category $\mathbf{EM}_2$ whose morphisms are vector-potential preserving isometries, then the theories EM_1 and EM_2 will *not* be classed as equivalent. But if we use the more sophisticated category $\overline{\mathbf{EM}}_2$ described above, we get the verdict that EM_1 and EM_2 are indeed equivalent. This is an instance of a general phenomenon. It is often far from obvious which categories provide suitable semantic representations of the theory under consideration. In the present case, we can at least appeal to the fact that the semantic representation $\overline{\mathbf{EM}}_2$ allows us to achieve the intuitively correct equivalence judgments, while the alternative representation $\mathbf{EM}_2$ fails to do so. But if we hope to use the semantic representation of theories to settle contested questions of theoretical equivalence, we clearly cannot rely on strong intuitions about what the 'right' equivalence judgments are. We need some clear independent criteria that guide us in choosing the correct semantic representation of a theory. One of our primary aims in this paper is to forward a first criterion of this type, which can be used independently of our intuitions about what the theoretical equivalence judgments should be.

3 Symmetries and Equivalence

The above discussion is important for another reason: besides illustrating the application of category theory to questions of equivalence, it also sheds light on the relationship between equivalence and *symmetries*. In the example of electromagnetism, a crucial role is played by the fact that gauge-related models are regarded as physically equivalent; in turn, this equivalence (which is between models rather than theories) is often thought to be undergirded by the fact that a gauge transformation is a *symmetry* of electromagnetism. In this section, we elaborate upon what is here meant by a symmetry, and on its relationship to physical equivalence between models.

In the first instance, a symmetry of a physical theory is a special kind of *transformation*: specifically, one which preserves the equations of motion, or (equivalently) which maps solutions to solutions. For example, in the case of electromagnetism, a gauge transformation (6) is a symmetry: the equation

(2) is invariant under the substitution of A_a with A'_a; equivalently, if a given four-potential A_a satisfies (2), then so does the transformed four-potential A'_a (and if A_a does not satisfy (2), then neither does A'_a).

Philosophically, much of the interest in the notion of symmetry stems from the role they play in the so-called *symmetry-to-reality* inference.[5] Roughly speaking, this inference may be stated as follows:[6]

1. Physical theory T admits of a certain symmetry, σ.
2. Feature X in T is not invariant under σ.

C. Therefore, T ought not to be interpreted in such a way that X represents a real physical feature of the world; σ-related solutions of T ought not to be interpreted as representing distinct possibilities.

For instance, Newtonian mechanics admits boosts as a symmetry, and so (says the symmetry-to-reality inference) it should not be interpreted as involving a commitment to absolute space; electromagnetism admits gauge transformations of the potential as a symmetry, and so (says the inference) should not be interpreted as involving a commitment to absolute potentials; and so on and so forth.[7]

The question then naturally arises: can we characterise symmetries in general terms, in such a way as to underwrite the symmetry-to-reality inference? It turns out that doing so proves a little trickier than one might have thought. One proposal can be immediately ruled out: that a symmetry is any bijection on the set of solutions to a theory.[8] On this proposal, any pair of solutions would be related by some symmetry or other, and so—if we also accepted the

[5] See (Dasgupta, 2016).

[6] This way of stating the symmetry-to-reality inference differs from that given by (Dasgupta, 2016) in two ways. First, it speaks of *theories* where Dasgupta speaks of *laws*; this is because we prefer to regard the symmetry-to-reality inference as a constraint on the best way to interpret theories, i.e., on the best way to work out what laws a given theory should be taken to express. Second (and relatedly), Dasgupta's version of the symmetry-to-reality inference is applied specifically to cases where the laws in question are the "complete laws of motion governing our world." We impose no such restriction, since we also wish to understand the application of this inference to cases where we do not take the laws in question to be true or complete: for example, the use of the symmetry-to-reality inference in the context of Newtonian mechanics (or, for that matter, in the context of any of our current scientific theories).

[7] Note that one need not follow the symmetry-to-reality inference quite this far; on the so-called "motivational" view of symmetries, the presence of a symmetry is merely a motivation to seek a version of the theory which identifies σ-related solutions, rather than a warrant to interpret the theory in the way described here. For further discussion of this distinction, see Møller-Nielsen (2016); for a defence of the "interpretational" view taken here, see Dewar (2015).

[8] This is what Belot (2013) calls the "Fruitless Definition" of symmetry.

symmetry-to-reality inference—we would conclude that any theory describes just one possible world, and hence that any physical feature which is not identical in every model is unreal.

Some authors[9] have proposed that the right way to characterise symmetries is in *epistemic* terms: that is, as bijections which map solutions to *epistemically indiscernible* solutions. Although we think that an epistemic component is necessary (as our proposal in section 4 will make clear), we do not think that it is sufficient by itself; it fails to track the widespread feeling that part of what is distinctive about symmetry transformations is that they preserve (a significant component of) the formal structure of a theory.

Moreover, in specific contexts, we can often find such formal criteria for symmetries. For example, suppose that our theory is presented as a set of differential equations, governing maps from one differential manifold T to another differential manifold X.[10] Then one typically[11] limits attention to local transformations of the space of dependent and independent variables, i.e., to diffeomorphisms of the space $T \times X$. Given any such diffeomorphism g and any smooth function $f : T \to X$, the transformed function $g[f]$ (if it exists) is the smooth function whose graph is that obtained by acting on the graph of f by g. A symmetry is then a diffeomorphism $g : T \times X \to T \times X$ such that for any $f : T \to X$, whenever $g[f]$ is defined, $g[f]$ is a solution to the differential equations if and only if f is. By limiting attention to only those bijections on solutions which can be "generated" from local transformations in this way, we are able to rule out the most egregious examples of "symmetries", mapping solutions to arbitrarily physically different solutions.

Moreover, this perspective lets us go a little deeper into the dual nature of symmetries mentioned above (as transformations that preserve the equations of motion, and as transformations that map solutions to solutions): this duality corresponds to the distinction between the so-called "passive" and "active" conceptions of transformations. To illustrate this, consider the example of a free inertial particle moving in two spatial dimensions, as expressed in Cartesian coordinates: so, in the notation above, we let $T \cong \mathbb{R}$ and $X \cong \mathbb{R}^2$. Then, we can state the theory at hand as consisting of the set

[9] e.g. Ismael and van Fraassen (2003), Dasgupta (2016).

[10] Where both T and X may carry further structure beyond their differential structure (perhaps they are affine spaces, or even copies of $\mathbb{R}^k$); we just mean that T and X have *at least* differential structure.

[11] e.g. in Olver (1986).

of differential equations

$$\ddot{x} = 0 \tag{7a}$$

$$\ddot{y} = 0 \tag{7b}$$

Now, consider the transformation to polar coordinates (r, θ), where

$$x = r\cos\theta \tag{8a}$$

$$y = r\sin\theta \tag{8b}$$

(We act on T with the identity.) Given any function $f : T \to X$, let $\widetilde{f}$ be the transformed function that arises by applying (8) to f in the manner described above.[12]

On the passive conception, the transformation (8) represents a change in the *description*. That is, it is a change in the representational conventions governing the mathematics, such that (as a matter of stipulation) $\widetilde{f}$ represents whatever particle-history was previously represented by f. Now, if such a change is to be consistent, then it ought to be accompanied by a change in the differential equations being used. After all, we don't think that a change of representation ought to make any difference to the physics! So it had better be the case that the same physical histories get picked out as before; given that those histories are now represented by different functions, we want to change the equations so as to pick out that new set of functions. If the old set of differential equations was Δ, then we want a new set $\widetilde{\Delta}$ such that $\widetilde{f}$ satisfies $\widetilde{\Delta}$ iff f satisfies Δ: one can easily show that such a $\widetilde{\Delta}$ is given by

$$\ddot{r} - r\dot{\theta}^2 = 0 \tag{9a}$$

$$r\ddot{\theta} + 2\dot{r}\dot{\theta} = 0 \tag{9b}$$

On the active conception, the transformation (8) is just a transformation of the variables, which transforms functions from T to X into other functions from T to X. In particular, the representational conventions are held invariant across the application of this change. As such, there is no concomitant change to the differential equations: and so, in general, a function $f : T \to X$

[12]Note that in this case, the function $\widetilde{f}$ will always exist, since (8) is (in the terminology introduced below) fibre-preserving.

and its transformed counterpart $\widetilde{f} : T \to X$ will represent different particle trajectories.

These two conceptions yield slightly different ways of thinking about symmetries, and why they might be interesting. On the passive conception, we can characterise symmetry transformations as those transformations which do *not* require any alteration to the differential equations being used. For, if f satisfies Δ iff $\widetilde{f}$ satisfies Δ, then f satisfies Δ iff f satisfies $\widetilde{\Delta}$: and hence, $\widetilde{\Delta} = \Delta$. On the active conception, symmetry transformations are simply those transformations under which the space of solutions is invariant. In other words, both conceptions agree that in a symmetry transformation, both the (mathematical) theory in play, and the dynamical status of any model, remain unchanged. However, on the passive conception, the latter condition comes "for free" (since it holds for *any* transformation), and so the distinctive thing about symmetry transformations in particular is their satisfaction of the former condition. On the active conception, by contrast, it is the former condition that is upheld across all transformations, and so the distinctive nature of a symmetry transformation is that it also meets the latter condition.

In turn, this lets us be a little clearer about the conclusion of the symmetry-to-reality inference. The conclusion of that inference is that a pair of symmetry-related models, considered as models of the same theory and without changing our representational conventions, should not be interpreted as representing distinct possibilities. This makes clear that it is not just the trivial observation that we can change how the mathematics is to represent the physics; rather, it is a claim about the proper way to extract physical content from a theory. In many cases, such a claim can yield quite striking results.

There is a further relationship between symmetries and equivalence, which we are now in a position to appreciate: symmetries such as the above can be thought of as equivalences (or translations) from a theory to itself. We can use the same example to illustrate this. The pair of theories (7) and (9) is a paradigmatic example of a pair of equivalent theories: we have constructed them so that are (interpretable as) two alternative representations of the same physical laws. More formally, we can recognise this by observing that the translation (8) translates the theory (7) to the theory (9); and that the

inverse translation

$$r = (x^2 + y^2)^{1/2} \tag{10a}$$

$$\theta = \tan^{-1}\left(\frac{y}{x}\right) \tag{10b}$$

translates the theory (9) to (7).

However, now suppose that we had used a different transformation: say, a Galilean transformation to coordinates (x', y') where

$$x' = x + ut + a \tag{11a}$$

$$y' = y + vt + b \tag{11b}$$

In this case, of course, we find that the "transformed" version of the theory (7) is just (7) again: exactly the observation which, on the passive conception of transformations being applied here, betokens the status of (11) as a symmetry. Thus, the transformation (11) may be regarded as a translation from the theory (7) *to itself*. This is suggestive of a justification for the symmetry-to-reality inference: it is natural to take solutions related by a translation (e.g. the solutions $(x = t, y = t)$ and $(r = t\sqrt{2}t, \theta = \pi/4)$) to represent the same physical possibility; the symmetry-to-reality inference is just the specific application of this in the case where the source and target of the translation are the same theory. We see, therefore, that the characterisation of symmetries in terms of transformations of variables yields valuable insights into their philosophical use.

However, there remain some reasons to be unsatisfied. For one thing, this proposal is rather specific to a particular way of presenting physical theories. If the theory is characterised, for example, as governing sections of a fibre bundle rather than functions between a pair of differential manifolds, then the above is clearly not appropriate, and would need to be adapted (in that context, the requirement is typically that the map on sections be "generated" from bundle automorphisms). And even in one context, there are various different kinds of transformation that one might want to consider as the "generating" transformations. For instance, in the context of differential equations governing maps between manifolds T and X, one often wants to limit attention to "fibre-preserving" diffeomorphisms: those in which the action on the independent variables does not depend on the dependent vari-

ables (i.e., which may be specified by a pair of diffeomorphisms $\tau : T \to T$ and $\xi : T \times X \to X$, as $g : (t, x) \mapsto (\tau(t), \xi(t, x))$. Note that if g is fibre-preserving, then for any smooth function $f : T \to X$, the function $g[f]$ is well-defined.

Thus, one might seek some way of characterising symmetries which can be applied to a wide variety of theories, presented using many different formalisms. As already discussed, the framework of category theory offers such a unifying framework: there are reasons for thinking that many theories, from diverse formalisms, can be given a perspicuous categorical representation. In the next section, we propose a criterion for characterising symmetries in a categorical context.

4 Symmetry, Equivalence and Representation

Let's summarise our conclusions so far. In section 2 we argued that theoretical equivalence should be formalised by the existence of an empirical-content-preserving categorical equivalence between the semantic categories of the theories under consideration. In section 3 we showed that symmetries of a physical theory T can be thought of as theoretical equivalences of T with itself. Together, these positions motivate the following thesis:

T1: Let T be a physical theory with associated semantic category $\mathcal{C}$. Then the symmetries of T can be identified with empirical-content-preserving auto-equivalences of $\mathcal{C}$.[13]

In order to illustrate and motivate T1, recall again the electromagnetic theories EM_1 (electromagnetism with fields), EM_2 (electromagnetism with potentials), EM_2' (electromagnetism with equivalence classes of gauge-related potentials), and the prospective semantic representations $\mathbf{EM}_1$, $\mathbf{EM}_2$, $\overline{\mathbf{EM}}_2$, $\mathbf{EM}_2'$. If **T1** is to hold, then any symmetry of classical electromagnetism had better be representable as an empirical-content-preserving auto-equivalence of the semantic categories $\mathbf{EM}_1$, $\mathbf{EM}_2$, $\overline{\mathbf{EM}}_2$ and $\mathbf{EM}_2'$. The paradigmatic symmetries of electromagnetism are given by gauge transformations. If we take the models of electromagnetism to be given by vector potentials satisfying the relevant differential equations (as in EM_2, for example), then we

[13] By 'auto-equivalence of $\mathcal{C}$', we mean a categorical equivalence $F : \mathcal{C} \to \mathcal{C}$.

know that models obtained by gauge transformations are physically equivalent. Similarly, if we take models of electromagnetism to be given by a set of vector potentials closed under gauge transformations, then the action of a gauge transformation can then be seen as the image of the set under that transformation, which will clearly be identical to the original set. In both cases, gauge transformations take models to physically equivalent models. The following proposition shows that the symmetries of electromagnetism given by gauge transformations can indeed be faithfully represented as empirical-content-preserving auto-equivalences of the semantic categories of the proposed formulation (all absent proofs in Appendix).

Proposition 1. *Let λ be any smooth scalar field on M. Let $F_2 : \boldsymbol{EM}_2 \to \boldsymbol{EM}_2$ be defined as follows:*

- *For any object (M, A_a, J^a) in $\boldsymbol{EM}_2$,*

$$F_2(M, A_a, J^a) = (M, A_a + \nabla_a \lambda, J^a) \tag{12}$$

- *For any morphism $f : X \to Y$ given by an isometry $\psi : M \to M$, $F_2(f)$ is the morphism from $F_2(X)$ to $F_2(Y)$ given by that same isometry ψ.*

Let $\overline{F_2} : \overline{\boldsymbol{EM}_2} \to \overline{\boldsymbol{EM}_2}$ be defined as follows:

- *For any object (M, A_a, J^a) in $\overline{\boldsymbol{EM}_2}$,*

$$\overline{F_2}(M, A_a, J^a) = (M, A_a + \nabla_a \lambda, J^a) \tag{13}$$

- *For any morphism $f : X \to Y$ given by an isometry $\psi : M \to M$ and a scalar field χ on M, $\overline{F_2}(f)$ is the morphism from $\overline{F_2}(X)$ to $\overline{F_2}(Y)$ given by that same isometry ψ and that same scalar field χ.*

Let $F_2' : \boldsymbol{EM}_2' \to \boldsymbol{EM}_2'$ be defined as follows:

- *For any object $[(M, A_a, J^a)]$ in $\boldsymbol{EM}_2'$,*

$$F_2'[(M, A_a, J^a) = [(M, A_a + \nabla_a \lambda, J^a)] \tag{14}$$

- *For any morphism $f : X \to Y$ given by an isometry $\psi : M \to M$, $F_2'(f)$ is the morphism from $F_2'(X)$ to $F_2'(Y)$ given by that same isometry ψ.*

Then F_2, $\overline{F_2}$ and F_2' are empirical-content-preserving auto-equivalences of

$\boldsymbol{EM}_2$, $\overline{\boldsymbol{EM}_2}$ *and* $\boldsymbol{EM}_2'$, *respectively.*[14]

So at least in this case, it is clear that symmetries of a theory can indeed be adequately formalised as empirical content preserving auto-equivalences of the relevant semantic category. Indeed, the preceding example motivates the idea that a basic criterion for choosing a semantic categorical representation of a theory is that the symmetries of that theory be formalisable as content preserving auto-equivalences of the semantic category. However, this by itself is not yet of much help to us. Recall that we are interested in spelling out a general criterion in virtue of which the category $\mathbf{EM}_2$ does not provide a suitable representation of classical electromagnetism. And the criterion we've just described fails to do so. Proposition 1 tells us that gauge transformations do indeed define content preserving auto-equivalences on $\mathbf{EM}_2$, so we need something more than the requirement that symmetries be representable as auto-equivalences of the semantic category. The following observation will be useful in this regard.

Proposition 2. *Let* F_2, $\overline{F_2}$, F_2' *be as in proposition 1. Then* $F_1 = id_{\boldsymbol{EM}_1}$, $F_2' = id_{\boldsymbol{EM}_2'}$, $\overline{F_2}$ *is naturally isomorphic to the identity on* $\overline{\boldsymbol{EM}_2}$, *but* F_2 *is not naturally isomorphic to the identity on* $\boldsymbol{EM}_2$.

Recall that given two functors $F, G : \mathcal{C} \to \mathcal{D}$, a *natural transformation* $\eta : F \to G$ is a family $\{\eta_X : FX \to GX\}_X$ of morphisms in $\mathcal{D}$, indexed by the objects X of $\mathcal{C}$, such that the following diagram commutes for any morphism $f : X \to Y$ of $\mathcal{C}$: A *natural isomorphism* is just a natural transformation η

$$\begin{array}{ccc} FX & \xrightarrow{Ff} & FY \\ \downarrow{\scriptstyle \eta_X} & & \downarrow{\scriptstyle \eta_Y} \\ GX & \xrightarrow{Gf} & GY \end{array}$$

where for every object X of $\mathcal{C}$, η_X is an isomorphism (in $\mathcal{D}$).

So proposition 2 draws a significant distinction between $\mathbf{EM}_2$ on the one hand, and the three categorically equivalent representations $\mathbf{EM}_1$, $\overline{\mathbf{EM}_2}$ and $\mathbf{EM_2}'$ on the other. Specifically, the representations of gauge transformations given by the latter three categories are all auto-equivalences which are *naturally isomorphic to the identity functor*. This means that from the perspective of these three categories, gauge transformations fail to effect any

[14] Note that the 'gauge functor' F_1 is trivially just the identity on $\mathbf{EM}_1$.

discernible change on the space of possible models of the theory. In contrast, the representation of gauge transformations given by $\mathbf{EM}_2$ is not naturally isomorphic to the identity, which means that gauge transformations do induce a significant transformation on the space of possible models given by $\mathbf{EM}_2$.

Intuitively, a functor that is naturally isomorphic to the identity is one which, from the perspective of the relevant category, leaves everything essentially unchanged. Since symmetries are often described as transformations that preserve physical content and so affect no discernible change to the relevant physical system, it seems natural to require that their representations as auto-equivalences of the semantic category be naturally isomorphic to the identity. And this is a criterion that we can use to diagnose the representational deficiency of $\mathbf{EM}_2$. For, the representation of gauge symmetries as auto-equivalences on $\mathbf{EM}_2$ is not naturally isomorphic to the identity, while the corresponding representations in the categories $\mathbf{EM}_1$, $\overline{\mathbf{EM}_2}$ and $\mathbf{EM_2}'$ do have this property.

Assuming T1 (that we should identify symmetries with content preserving auto-equivalences of the semantic category), together with the claim that symmetries should not affect any discernible change on the space of models of a theory, we are left with the following criterion for prospective semantic representations of physical theories.

Sophistication 1 (S1): In order for a category $\mathcal{C}$ to provide a suitable semantic representation of a physical theory T, it is necessary that $\mathcal{C}$ should be *sophisticated*, i.e. it should not allow for any empirical-content-preserving auto-equivalences that are not naturally isomorphic to the identity.

The requirement that semantic categories always be sophisticated in the sense defined above is a general formal criterion that refers only to intrinsic properties of prospective representations of theories. Note that S1 does not require us to invoke any desired verdicts regarding theoretical equivalence when determining whether or not a potential semantic representation $\mathcal{C}$ is suitable. S1 talks only about the mathematical structure of $\mathcal{C}$, and provides a general criterion that immediately narrows down the class of prospective categorical representations of physical theories quite radically. Proposition 2 guarantees that $\mathbf{EM}_2$ could never provide an adequate representation of

classical electromagnetism, regardless of the fact that it gives the wrong judgments regarding theoretical equivalence. For, $\mathbf{EM}_2$ is not sophisticated and so admits of symmetries which induce discernible change on the space of models. In contrast, it seems that $\mathbf{EM}_1$, $\overline{\mathbf{EM}_2}$ and $\mathbf{EM}_2'$ are sophisticated in the relevant sense and so may be capable of providing a satisfactory representation of the content of classical electromagnetism.

Thus, S1 appears to provide a plausible formal criterion to be satisfied by any prospective semantic representations of physical theories. Crucially, it allows us to disallow the representation given by $\mathbf{EM}_2$ without needing to refer to intuitions regarding the equivalence of different formulations of electromagnetism. Given that part of the purpose of providing a semantic categorical representation of theories is to settle contentious questions of categorical equivalence, it would be hopelessly circular to rely on intuitions concerning theoretical equivalence to determine which representations are adequate. S1 dispenses with this reliance on intuitions and provides a general formal criterion with respect to which $\mathbf{EM}_2$ is fundamentally unsuitable for representing the content of physical theories.

At this point, one might be tempted to object that the criterion S1 is too strong a restriction to place on the possible representations of physical theories. And indeed, there are some initially plausible weakenings of S1 that also warrant closer inspection. Consider first the following definition.

Definition 4.1. *Call a category $\mathcal{C}$ 'semi-sophisticated' if every content preserving auto equivalence of $\mathcal{C}$ sends all objects to isomorphic objects. Call $\mathcal{C}$ 'nearly sophisticated' if every content preserving auto-equivalence sends all objects to isomorphic objects and all arrows to isomorphic arrows.*[15]

Let S3 be the criterion according to which a semantic category should always be semi-sophisticated and let S2 be the criterion according to which a semantic category should always be nearly sophisticated. The following result shows that S1, S2 and S3 give substantially different verdicts on which categories are suitable for providing semantic representations of physical theories.

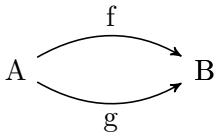
Proposition 3. *S1 is strictly stronger than S2 and S2 is strictly stronger*

[15] We say that two arrows f, g are 'isomorphic' if there exist isomorphisms $a : dom(f) \to dom(g)$, $b : cod(f) \to cod(g)$ such that $g \circ a = b \circ f$. This is exactly the notion of isomorphism at play in $\mathcal{C}$'s 'arrow category'.

than S3.

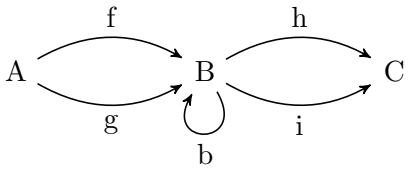
Proof. To see that S1 implies S2, just observe that if η is a natural isomorphism between F and the identity, then for any object X, the isomorphism η_X relates X to FX (so F takes objects to isomorphic objects) in such a way that for any morphism $f : X \to Y$, $F(f) \circ \eta_X = \eta_Y \circ f$ (so F takes arrows to isomorphic arrows). That S2 implies S3 is immediate from the definitions.

To illustrate the converse, consider the following category $\mathcal{C}$ with two objects and two distinct non-identity morphisms, which satisfies S3 but not S2 or S1.



Let $F : \mathcal{C} \to \mathcal{C}$ be the functor that switches f and g ($F(f) = g$, $F(g) = f$) but acts as the identity elsewhere. It is easy to see that F is an auto-equivalence of $\mathcal{C}$ that sends objects to isomorphic objects. However, it is also easy to see that F is not naturally isomorphic to the identity, since there are no natural automorphisms on A and B, and F does not send morphisms to isomorphic morphisms (for the same reason). Furthermore, it is easy to see that, apart from the identity, F is the only auto-equivalence of $\mathcal{C}$. So (assuming for now that all auto-equivalences are content preserving) $\mathcal{C}$ is semi-sophisticated, but not nearly or fully sophisticated. So according to S1 and S2, $\mathcal{C}$ is not capable of faithfully representing the content of physical theories.

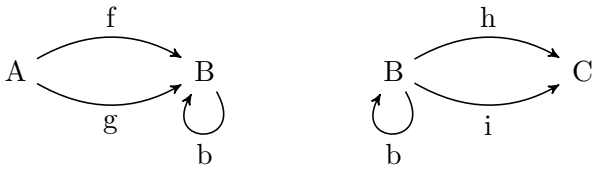
Similarly, consider the category $\mathcal{C}$ depicted below, which satisfies S3 and S2, but not S1, where we stipulate that $b \circ b = id_B$, $b \circ f = g$, $b \circ g = f$, $h \circ b = i$, $i \circ b = h$.



Let $F : \mathcal{C} \to \mathcal{C}$ be the functor which acts as the identity on $\mathcal{C}$ except that $F(h) = i$, $F(i) = h$. It is easy to see that F is an auto-equivalence that sends objects to isomorphic objects and arrows to isomorphic arrows, but is

not naturally isomorphic to the identity. Furthermore, it is also easy to see that all other auto-equivalence on $\mathcal{C}$ are naturally isomorphic to the identity, and so (assuming again that all auto-equivalences are content preserving) $\mathcal{C}$ is both semi and nearly sophisticated, but not fully sophisticated. Thus, according to our proposal, categories of this form are also fundamentally incapable of representing the content of physical theories. □

There is a sense in which one can think of an auto-equivalence that sends objects to isomorphic objects and arrows to isomorphic arrows as being 'locally isomorphic to the identity'. Such a functor behaves like the identity on local regions of the category, but fails to preserve the global structure of the category in a sufficiently robust way. Thus, categories which are nearly but not fully sophisticated can be thought of as having 'excess global structure'. In the previous example, the relevant auto-equivalence *is* naturally isomorphic to the identity when restricted to either of the two subcategories below.



It is only when we put these subcategories together that the global structure plays a role and the auto-equivalence can be distinguished from the identity.

In general, it is well known that there exist many extremely rich categories for which every auto-equivalence is naturally isomorphic to the identity. Two particularly striking examples are the category of sets and functions and the category of groups and group homomorphisms.[16] The fact that categories of this size and complexity would always satisfy S1 should provide reassurance that the proposed criterion will not limit our ability to represent theories with rich mathematical structure.

Now, once we accept that the proposed criterion S1 is not too strong a restriction on the permitted representations of physical theories, the inverse

[16]For a proof that the category of groups has this property, see p. 31 of Freyd (1964). The fact that the category of sets is guaranteed to satisfy S1 is particularly important, since it shows that topos theoretic theoretic representations of theories of the type discussed by e.g. Eva (2016), Döring and Isham (2011), Heunen et al. (2009) are not generally prohibited.

question also naturally arises. Specifically, how do we choose between multiple possible semantic representations, all of which satisfy S1? In the interests of space, however, we will not address that question here. To be clear, we do not believe that S1 gives the full story regarding which categories have the capacity to represent the content of physical theories. We simply posit S1 as a first necessary condition which, it turns out, is sufficient to settle some existing controversies about the semantic representations of theories. It may well be that there are other general conditions that we should impose on potential categorical representations of theories, but we restrict ourselves here to talking only about S1–S3.

5 Newtonian gravitation

We are now ready to study the implications of S1 for disputed questions of theoretical equivalence. Weatherall (2016a) addresses the question of whether Newtonian gravitation (NG) is theoretically equivalent to geometrized Newtonian gravitation (GNG, otherwise known as 'Newton-Cartan gravitation'). Briefly,[17] models of NG are of the form (L, ∇, ρ, ϕ), where L is a Leibnizian spacetime (assumed to be topologically $\mathbb{R}^4$ and spatially flat),[18] ∇ is a flat connection compatible with L,[19] and ρ and ϕ are scalar fields on L (representing the mass density and gravitational potential respectively); models are required to satisfy *Poisson's equation*,[20]

$$\nabla_a \nabla^a \phi = 4\pi\rho \tag{15}$$

Models of GNG are of the form $(L, \tilde{\nabla}, \rho)$ where L is again a (spatially flat and topologically $\mathbb{R}^4$) Leibnizian spacetime, $\tilde{\nabla}$ is a (not necessarily flat) connection compatible with L, and ρ is again a scalar field on L; these models

[17] See Malament (2012) for a full account.

[18] A Leibnizian spacetime may be regarded as a bundle of three-dimensional Euclidean spaces over a one-dimensional oriented Euclidean base space: so it carries a standard of absolute simultaneity, a directed temporal metric (usually expressed via a closed one-form t_a), and a spatial metric (usually written h^{ab}), but no further structure. See (Earman, 1989, Chapter 2) or (Malament, 2012, Chapter 4) for more details.

[19] "Compatible" in the sense that $\nabla_a t_{bc} = 0$ and $\nabla_a h^{bc} = 0$.

[20] Here, $\nabla^a := h^{ab}\nabla_b$.

are required to satisfy

$$R_{bc} = 4\pi\rho t_b t_c \tag{16a}$$

$$R^a{}_{b}{}^c{}_d = R^c{}_d{}^a{}_b \tag{16b}$$

$$R^{ab}{}_{cd} = 0 \tag{16c}$$

where $R^a{}_{bcd}$ is the Riemann curvature tensor of $\tilde{\nabla}$.

The empirical content of these theories is given by the trajectories of test bodies that are subject to no non-gravitational influences. Given a model (L, ∇, ρ, ϕ) of NG, therefore, its empirical content is the set Γ of curves $\gamma : I \to L$ (for I some interval in $\mathbb{R}$) such that

$$\dot{\gamma}^n \nabla_n \dot{\gamma}^a = -\nabla^a \phi \tag{17}$$

For a model $(L, \tilde{\nabla}, \rho)$ of GNG, its empirical content is given by the set $\tilde{\Gamma}$ of curves γ satisfying

$$\dot{\gamma}^n \tilde{\nabla}_n \dot{\gamma}^a = 0 \tag{18}$$

Weatherall proposes that we represent GNG by the category **GNG**, defined as follows:

- Objects are models of GNG $(L, \tilde{\nabla}, \rho)$ (with all such models using the same Leibnizian spacetime L)
- A morphism $f : (L, \tilde{\nabla}, \rho) \to (L, \tilde{\nabla}', \rho')$ is given by a diffeomorphism $d : L \to L$ such that $\nabla' = d_* \nabla$ and $\rho' = d_* \rho$.[21]

This certainly seems to provide the most natural representation of the theory.

However, the situation is more complicated with NG, for which there seem to be two distinct initially plausible semantic representations. Clearly, objects should be models of the theory, i.e. flat classical spacetimes equipped with gravitational potentials and mass densities (L, ∇, ϕ, ρ), satisfying Poisson's equation (15). And it is natural to take the morphisms to be those maps which preserve all this structure. So let **NG** be the category defined as follows:

- Objects are models of NG (L, ∇, ϕ, ρ) (with all such models using the same Leibnizian spacetime L)

[21] Following Weatherall, given two manifolds $\mathcal{M}$ and $\mathcal{N}$ and a diffeomorphism $k : \mathcal{M} \to \mathcal{N}$, let $k_* \nabla$ be the unique derivative operator on $\mathcal{N}$ such that for any tensor field α, $(k_* \nabla_a)\alpha = k_*(\nabla_a \alpha)$.

- A morphism $f : (L, \nabla, \phi, \rho) \to (L, \nabla', \phi', \rho')$ is given by a diffeomorphism $d : L \to L$ such that $\nabla' = d_* \nabla$, $\phi' = d_* \phi$, and $\rho' = d_* \rho$.

However, like EM_2, NG carries a certain kind of gauge symmetry. Specifically, given any model (L, ∇, ρ, ϕ) of NG, a joint transformation of the connection and the gravitational potential of the form[22]

$$\nabla \mapsto \nabla' = (\nabla, (\nabla^a \psi) t_b t_c) \tag{19a}$$

$$\phi \mapsto \phi' = \phi + \psi \tag{19b}$$

where ψ is any scalar field on L such that $\nabla^a \nabla^b \psi = 0$, will map solutions to solutions (equivalently, is such that Poisson's equation is invariant under this transformation). Moreover, empirical content is preserved, in the sense that the dynamically allowed trajectories of the two models coincide: i.e., for any curve γ,

$$\dot{\gamma}^n \nabla_n \dot{\gamma}^a = -\nabla^a \phi \text{ iff } \dot{\gamma}^n \nabla'_n \dot{\gamma}^a = -\nabla'^a \phi' \tag{20}$$

According to our thesis **T1**, this transformation should correspond to an (empirical-content-preserving) autoequivalence of **NG**. This is indeed the case, as the following proposition verifies.

Proposition 4. *Let ψ be any smooth scalar field on L such that with respect to any compatible connection ∇, $\nabla^a \nabla^b \psi = 0$.*[23] *Let $K : \boldsymbol{NG} \to \boldsymbol{NG}$ be defined as follows:*

- *For any object (L, ∇, ϕ, ρ) in **NG**,*

$$K_1(L, \nabla, \phi, \rho) = (L, (\nabla, (\nabla^a \psi) t_b t_c), \phi + \psi, \rho) \tag{21}$$

- *For any morphism $f : X \to Y$ given by a diffeomorphism $d : L \to L$, $K_1(f)$ is the morphism from $K_1(X)$ to $K_1(Y)$ given by that same diffeomorphism d.*

*Then: K is an autoequivalence of **NG** which preserves empirical content.*

However, the presence of this symmetry also problematises the choice of **NG** as the appropriate semantic category for representing NG: the functor

[22] The notation for the transformed potential follows (Malament, 2012, §7).

[23] This definition is well-formed: one can show that for any compatible connections ∇ and ∇', for any tensor field α which is spacelike in its contravariant indices and timelike in its covariant indices, $\nabla'^a \alpha = \nabla^a \alpha$.

K_1 is not sophisticated, i.e., is not naturally isomorphic to the identity (indeed, it is not even semi-sophisticated, since in general an object X of **NG** will not be isomorphic to $K(X)$). This motivates the definition of the following category $\overline{\mathbf{NG}}$, as providing a better semantic representation of the content of NG:

- Objects are models of NG (L, ∇, ϕ, ρ) (with all such models using the same Leibnizian spacetime L)
- A morphism $f : (L, \nabla, \phi, \rho) \to (L, \nabla', \phi', \rho')$ is given by a pair (d, θ), where θ is a smooth scalar field such that $\nabla^a \nabla^b \theta = 0$, and d is a diffeomorphism $d : L \to L$ such that $\nabla' = d_*(\nabla, (\nabla^a \theta) t_b t_c)$, $\phi' = d_*(\phi + \theta)$, and $\rho' = d_* \rho$.

The question then is which of **NG** and $\overline{\mathbf{NG}}$ provides the better semantic representation of the content of NG. On this question, Weatherall writes

> The first option better reflects how physicists have traditionally thought of Newtonian gravitation. On the other hand, this option appears to distinguish between models that are not empirically distinguishable, even in principle. Moreover, there are physical systems for which option 1 leads to problems, such as cosmological models with homogeneous and isotropic matter distributions, where option 1 generates contradictions that option 2 avoids. These latter arguments strike me as compelling, and I tend to think that option 2 is preferable. But I will not argue further for this thesis, and for the purposes of the present paper, I will remain agnostic about these options. (Weatherall, 2016a, p. 1085)

While there appear to be some significant theoretical considerations that count in favour of $\overline{\mathbf{NG}}$, there is no extant general formal criterion that we can employ to determine which of the representations to prefer when evaluating questions of theoretical equivalence and philosophical interpretation. And this is problematic. For, Weatherall proves that while there is no categorical equivalence between **NG** and **GNG**, there is such an equivalence between $\overline{\mathbf{NG}}$ and **GNG**. Applying Weatherall's definition of theoretical equivalence, we see that the answer to the question of whether NG is theoretically equivalent to GNG depends crucially on which categorical representation of NG

we choose to employ. Until now, there has been no general criterion to help us resolve problems of this type. Happily, S1 is up to the task. Specifically, we have the following result.

Proposition 5. *Let ψ be any smooth scalar field on L such that with respect to any compatible connection ∇, $\nabla^a\nabla^b\psi = 0$. Let $\overline{K} : \overline{\boldsymbol{NG}} \to \overline{\boldsymbol{NG}}$ be defined as follows:*

- *For any object (L, ∇, ϕ, ρ) in $\overline{\mathbf{NG}}$,*

$$K_2(L, \nabla, \phi, \rho) = (L, (\nabla, (\nabla^a\psi)t_b t_c), \phi + \psi, \rho) \tag{22}$$

- *For any morphism $f : X \to Y$ given by a diffeomorphism $d : L \to L$ and scalar field θ, $K_2(f)$ is the morphism from $K_2(X)$ to $K_2(Y)$ given by that same diffeomorphism d and that same scalar field θ.*

Then: $\overline{K}$ is an autoequivalence of $\overline{\mathbf{NG}}$ which preserves empirical content, and which is naturally isomorphic to the identity.

By S1, propositions 4 and 5 guarantee that **NG** is not capable of faithfully representing the content of Newtonian gravitation. This leaves $\overline{\mathbf{NG}}$ as the only remaining candidate representation. Thus, the categorical equivalence of $\overline{\mathbf{NG}}$ with **GNG** is sufficient to warrant the conclusion that NG and GNG are indeed theoretically equivalent. This provides a strong demonstration of the philosophical fecundity of S1. The conclusion that GNG and NG are theoretically equivalent was unavailable in the absence of any general restrictions on categorical semantic representations of theories. By forwarding a criterion of this type, we are able to settle previously irresolvable questions of theoretical equivalence in a principled way.

6 Conclusion

In summary, then, we have forwarded a novel criterion to be satisfied by any prospective semantic representations of physical theories. We've seen that this criterion allows us to decisively settle open questions regarding theoretical equivalence in a way that doesn't rely on our prior intuitions regarding what the answers to those questions should be. We've also defended a general formal definition of the symmetries of a physical theory that can be applied as soon as the semantic representation of the theory has been

specified. In future work, we hope to identify further constraints on what counts as an admissible semantic representation of a physical theory.

Appendix

Proof of Proposition 1: For F_2, the first part of the proof is a matter of showing that F is indeed a functor. So, first, observe that for any X in $\mathbf{EM}_2$, Id_X is given by the isometry Id_M; so $F(\mathrm{Id}_X)$ is (by the definition above) the morphism from $F(X)$ to $F(X)$ given by Id_M, i.e., $\mathrm{Id}_{F(X)}$. Second, for any morphisms $f : X \to Y$ given by ψ_f and $g : Y \to Z$ given by ψ_g, the morphism $g \circ f$ is that given by $\psi_g \circ \psi_f$; consequently, $F(g \circ f) = F(g) \circ F(f)$.

Second, we need to show that F_2 is an auto-equivalence. So consider the functor H_2, defined as follows:

- For any object (M, A_a, J^a) in $\mathbf{EM}_2$,

$$H_2(M, A_a, J^a) = (M, A_a - \nabla_a \lambda, J^a) \tag{23}$$

- For any morphism $f : X \to Y$ given by an isometry $\psi : M \to M$, $H_2(f)$ is the morphism from $H_2(X)$ to $H_2(Y)$ given by that same isometry ψ.

By the same reasoning as above, H_2 is indeed a functor; moreover, it is clear that $H_2 \circ F_2 = \mathrm{Id}_{\mathbf{EM}_2}$ and $F_2 \circ H_2 = \mathrm{Id}_{\mathbf{EM}_2}$, and so H_2 is an almost inverse (in fact, exactly inverse) functor to F_2.

Finally, we need to show that F_2 preserves empirical content. If we take the empirical content of the models to be either the Faraday tensor or the (charge/mass-indexed) dynamical trajectories, then this is immediate (since the Faraday tensor is preserved under gauge transformations).

The proof that $\overline{F_2}$ is an autoequivalence of $\mathbf{EM}_2$ proceeds along very similar lines as that just given.

Finally, for F_2', we need only to observe that $[(M, A_a + \nabla_a \lambda, J^a)] = [(M, A_a, J^a)]$; consequently, $F_2' = \mathrm{Id}_{\mathbf{EM}_2'}$. So it follows immediately that F_2' is a functor, and that it is an autoequivalence of $\mathbf{EM}_2'$. □

Proof of Proposition 2. That F_2 is not naturally isomorphic to $\mathrm{Id}_{\mathbf{EM}_2}$ is immediate: in general, an object X of $\mathbf{EM}_2$ and its image $F_2(X)$ are not isomorphic, and so no natural isomorphism can be specified between $\mathrm{Id}_{\mathbf{EM}_2}$ and F_2.

For any object X of $\overline{\mathbf{EM}}_2$, however, let η_X be the morphism specified by (Id_M, λ); note that η_X is an isomorphism. Then for every $f : X \to Y$, η_X and η_Y are such that the diagram

$$\begin{array}{ccc} X & \xrightarrow{\quad f \quad} & Y \\ \updownarrow{\scriptstyle \eta_X} & & \updownarrow{\scriptstyle \eta_Y} \\ \overline{F_2}(X) & \xrightarrow{\overline{F_2}(f)} & \overline{F_2}(Y) \end{array}$$

commutes; in other words, the family η is a natural isomorphism from $\mathrm{Id}_{\overline{\mathbf{EM}_2}}$ to $\overline{F_2}$.

Finally, the fact that $F_2' = \mathrm{Id}_{\mathbf{EM}_2'}$ (and hence is trivially naturally isomorphic to $\mathrm{Id}_{\mathbf{EM}_2'}$) was noted in the proof of Proposition 1. □

Proof of Proposition 4. The proof that K is a functor is precisely analogous to that given in the proof of Proposition 1. To show that K is an equivalence, consider the functor K', defined as follows:

- For any object (L, ∇, ϕ, ρ) in $\mathbf{NG}$,

$$K'(L, \nabla, \phi, \rho) = (L, (\nabla, -(\nabla^a \psi) t_b t_c), \phi - \psi, \rho) \tag{24}$$

- For any morphism $f : X \to Y$ given by a diffeomorphism $d : L \to L$, $K'(f)$ is the morphism from $K'(X)$ to $K'(Y)$ given by that same diffeomorphism d.

It is straightforward to show that (in general) if $\nabla' = (\nabla, C^a_{bc})$ and $\nabla'' = (\nabla', C'^a_{bc})$, then $\nabla'' = (\nabla, C^a_{bc} + C'^a_{bc})$. It follows that $K' \circ K = \mathrm{Id}_{\mathbf{NG}} = K \circ K'$, and hence that K is an autoequivalence of $\mathbf{NG}$.

Finally, that K preserves empirical content follows from equation (20). □

Proof of Proposition 5. First, the fact that $\overline{K}$ is an autoequivalence follows by considering the functor $\overline{K}'$. $\overline{K}'$ acts on objects in the same manner as K'; and for any morphism $f : X \to Y$ given by a pair (ψ, d), $\overline{K}'(f)$ is the morphism from $\overline{K}'(X)$ to $\overline{K}'(Y)$ given by that same pair. As above, it follows that $\overline{K}' \circ \overline{K} = \mathrm{Id}_{\overline{\mathbf{NG}}} = \overline{K} \circ \overline{K}'$, so $\overline{K}$ is an autoequivalence of $\overline{\mathbf{NG}}$. And, as above, the fact that $\overline{K}$ preserves empirical content follows from equation (20).

Second, for any object X of $\overline{\mathbf{NG}}$, let η_X be the morphism specified by (Id_L, ψ); note that η_X is an isomorphism. Then for every $f : X \to Y$, η_X and η_Y are such that the diagram

$$\begin{array}{ccc} X & \xrightarrow{\quad f \quad} & Y \\ \updownarrow{\scriptstyle \eta_X} & & \updownarrow{\scriptstyle \eta_Y} \\ \overline{K}(X) & \xrightarrow{\overline{K}(f)} & \overline{K}(Y) \end{array}$$

commutes; in other words, the family η is a natural isomorphism from $\mathrm{Id}_{\overline{\mathbf{NG}}}$ to $\overline{K}$. Thus, $\overline{K}$ is naturally isomorphic to the identity.

□

References

Belot, G. (2013). Symmetry and equivalence. In Batterman, R. W., editor, *The Oxford Handbook of Philosophy of Physics*. Oxford University Press, New York.

Dasgupta, S. (2016). Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science*, 67(3):837–878.

Dewar, N. (2015). Symmetries and the Philosophy of Language. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52:317–327.

Döring, A. and Isham, C. (2011). "What is a Thing?": Topos Theory in the Foundations of Physics. In *New Structures for Physics*, pages 753–937. Springer.

Earman, J. (1989). *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time*. MIT Press, Cambridge, MA.

Eva, B. (2016). Topos Theoretic Quantum Realism. *The British Journal for the Philosophy of Science*, axv057.

Freyd, P. J. (1964). Abelian Categories. In *Abelian Categories: An Introduction to the Theory of Functors*, volume 1964. Harper and Row, New York.

Glymour, C. (1977). The Epistemology of Geometry. *Noûs*, 11(3):227–251.

Halvorson, H. (2012). What Scientific Theories Could Not Be. *Philosophy of Science*, 79(2):183–206.

Halvorson, H. (2016). Scientific Theories. In Humphreys, P., editor, *The Oxford Handbook of Philosophy of Science*, pages 585–608. Oxford University Press, Oxford.

Halvorson, H. and Tsementzis, D. (2015). Categories of Scientific Theories. In Landry, E., editor, *Categories for the Working Philosopher*. Oxford University Press, Oxford. Forthcoming; draft of July 29, 2015.

Heunen, C., Landsman, N. P., and Spitters, B. (2009). A Topos for Algebraic Quantum Theory. *Communications in Mathematical Physics*, 291(1):63–110.

Ismael, J. and van Fraassen, B. C. (2003). Symmetry as a guide to superfluous theoretical structure. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 371–392. Cambridge University Press, Cambridge.

Lutz, S. (2015). What Was the Syntax-Semantics Debate in the Philosophy of Science About? *Philosophy and Phenomenological Research*, 91(3). Early view version.

Mac Lane, S. (1978). *Categories for the Working Mathematician*. Springer, New York.

Malament, D. B. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*. University of Chicago Press, Chicago, IL.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation*. W.H. Freeman, San Francisco.

Møller-Nielsen, T. (2016). Invariance, Interpretation, and Motivation. *Philosophy of Science*. Forthcoming; draft of July 2016.

Nguyen, J., Teh, N., and Wells, L. (2017). Why surplus structure is not superfluous. page unpublished. Draft of June 17, 2017.

Olver, P. J. (1986). *Applications of Lie Groups to Differential Equations*. Springer-Verlag, New York, NY.

Rosenstock, S., Barrett, T. W., and Weatherall, J. O. (2015). On Einstein Algebras and Relativistic Spacetimes. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52, Part B:309–316.

Suppe, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. University of Illinois Press, Chicago.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford University Press, Oxford, UK.

Weatherall, J. O. (2016a). Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent? *Erkenntnis*, 81(5):1073–1091.

Weatherall, J. O. (2016b). Categories and the Foundations of Classical Field Theories. In Landry, E., editor, *Categories for the Working Philosopher*. Oxford University Press, Oxford. Forthcoming; draft of January 26, 2016.

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