Sophistication about Symmetries
Neil Dewar

ABSTRACT

Suppose that one thinks that certain symmetries of a theory reveal 'surplus structure'. What would a formalism without that surplus structure look like? The conventional answer is that it would be a reduced theory: a theory that traffics only in structures invariant under the relevant symmetry. In this article, I argue that there is a neglected alternative: one can work with a sophisticated version of the theory, in which the symmetries act as isomorphisms.

1 Introduction

It is often claimed that the symmetries of a theory reveal 'surplus structure': structure that, in some sense, the theory could do without.[1] For example, the boost symmetry of Newtonian mechanics indicates the superfluousness of absolute velocities; the gauge symmetry of electromagnetism reveals the superfluousness of absolute potentials; and so on and so forth. Moreover, it is widely held that if this is the case, then some modification of one's theory is appropriate, so as to make explicit what structure is not surplus (for example, the replacement of Newtonian by Galilean spacetime, in response to the boost symmetry of Newtonian mechanics).[2] In this article, I compare and contrast two ways of making such a modification. The first is to replace the theory by

---

[1] See, for example, (Redhead [1975]), to which the phrase 'surplus structure' is due; or the essays in (Brading and Castellani [2003]) and references therein.
[2] See (Møller-Nielsen [2017]), and references therein.

(what I shall call) a reduced theory: a theory that deals only in quantities that are invariant under the relevant symmetry. The second is to replace the theory by (what I shall call) a sophisticated theory: a theory in which models related by a symmetry are isomorphic.

In the next section, I set up some necessary apparatus, by defining what symmetries are of interest to us in this article: namely, symmetries of first-order relational theories, and internal symmetries of local field theories. In Section 3, I outline the use of reduction to expunge surplus structure from a theory, and suggest that it is somewhat problematic as a general strategy— even though it is standardly assumed to be the ne plus ultra of ways to enact the lessons of symmetries. In Section 4, I outline an alternative way to enact those lessons, which I call sophistication. In very rough terms (to be made precise later), the idea is that whereas a reduced theory converts a class of symmetry-related models into a single model, sophistication converts a class of symmetry-related models into a class of isomorphic models. Finally, in Section 5, I discuss the senses in which sophistication and reduction are or are not alternatives, and how the original theory relates to its reduced and sophisticated versions: the key result here is that in typical cases, the categories of models for a sophisticated theory and a reduced theory are equivalent (and that neither is equivalent to the category of models of the original theory). Section 6 concludes.

2 Symmetry

Here, I outline the kinds of symmetries that will be the topic of this article. I consider symmetries for two kinds of theories: for theories formulated in terms of first-order relational model theory, and for theories formulated as local field theories.

Here is what I mean by a theory formulated in terms of relational first-order model theory.[3] In this context, the basic notion is that of a signature: a set Σ of monadic and polyadic predicates. Given a signature Σ, one can define the set Form(Σ) of well-formed Σ-formulae, using the standard compositional rules of predicate logic. The set of Σ-sentences is the set of closed Σ-formulae (formulae with no free variables). Notation for the logical vocabulary will be standard, including free variables $x, x_1, x_2, \ldots$ (equipped with this ordering).

The semantics for a language with signature Σ is given by Σ-pictures.[4] A Σ-picture M consists of a set $|M|$ (the domain of M), equipped with a

---

[3] Notation and concepts mostly follows (Hodges [1997]).
[4] This terminology is non-standard. The more standard term is a Σ-structure: I have changed the terminology in order to avoid confusion between informal use of 'structure' and its use as a term of art.

function $\cdot^M$ with domain Σ. For each n-ary predicate $\Pi \in \Sigma$, Π^M is a set of n-tuples with members drawn from $|M|$, that is, $\Pi^M \subseteq |M|^n$. A Σ-picture M determines the truth or falsity of elements of Form(Σ), relative to a variable-assignment v for M, via the standard recursive clauses. If M makes a formula ϕ true relative to v, we write $M \vDash_v \phi$; if ϕ is a sentence, then the variable-assignment no longer matters, and we write simply $M \vDash \phi$.

A theory T in the signature Σ (for short, Σ-theory) is a set of Σ-sentences.[5] A Σ-picture M is said to be a model of T if it satisfies each member of T; we denote the class of all models of T by Mod(T). Finally, T entails a Σ-sentence ϕ just in case $M \vDash \phi$ for every $M \in$ Mod(T); this will be denoted by $T \vdash \phi$.[6] For example, consider the theory T_H of handedness. Letting $\Sigma_H = \{L, R\}$, T_H is the theory consisting of the following sentences:

$$\forall x(Lx \lor Rx), \tag{1a}$$

$$\forall x \neg (Lx \land Rx). \tag{1b}$$

Think of this as a (very simple) theory about worlds in which there is nothing but gloves: everything is either left-handed or right-handed, but nothing is both.

For theories formulated in terms of relational first-order model theory, the notion of symmetry were are interested in here is as follows: a symmetry is a translational equivalence between a theory and itself.[7,8] First, define a dictionary map from Σ_1 to Σ_2 to be any function $\mathfrak{D} : \Sigma_1 \to$ Form(Σ_2) such that for any m-ary predicate-symbol Π, $\mathfrak{D}\Pi$ is a formula with precisely the m variables $x_1, \ldots, x_m$ free. Intuitively, we can think of $\mathfrak{D}$ as a foreign-language dictionary, assigning each definiendum (primitive symbol of Σ_1) to a definiens (formula of Σ_2). A dictionary map gives us a means of converting any Σ_1-formula into a Σ_2-formula, through a process of substitution: given a Σ_1-formula ϕ, simply replace any atomic subformula $\Pi y_1, \ldots, y_m$ occurring in ϕ by $(\mathfrak{D}\Pi)(y_1/x_1, \ldots, y_m/x_m)$.[9] Let us denote the result of applying such a

---

5   In accordance with standard practice in model theory, I don't require theories to be deductively closed.

6   The notion of entailment defined here is semantic rather than syntactic, and would therefore more standardly be denoted $\vDash$ rather than $\vdash$; the problem is that the former symbol is rather too similar to the symbol for satisfaction ($\vDash$ compared to $\vDash$). Given that we will only apply this symbol in the context of first-order theories, the completeness theorem means we need not get too worried about this misuse of notation.

7   Which is not to say that there are not other interesting notions of symmetry in model theory. In particular, most uses of the term 'symmetry' in model theory use the term to refer to automorphisms of models (that is, isomorphisms from a model to itself). That notion will not be our focus here.

8   The notion of a translational equivalence is taken from (Barrett and Halvorson [unpublished]); it should be noted that—as is shown there—translational equivalence is, modulo trivial relabellings of predicates, equivalent to definitional equivalence.

9   Here, $\psi(y/x)$ denotes the result of uniformly substituting y for x everywhere in $\psi(x)$.

substitution to ϕ as $\mathfrak{D}\phi$. For the sake of brevity, I will write $\mathfrak{D} : \Sigma_1 \to \Sigma_2$ to indicate that $\mathfrak{D}$ is a dictionary map from Σ_1 to Σ_2.

Now suppose that we have two theories T_1 and T_2, in signatures Σ_1 and Σ_2, respectively. Then we say that a dictionary map $\mathfrak{D} : \Sigma_1 \to \Sigma_2$ is a translation of T_1 into T_2 if, for every ϕ such that $T_1 \vdash \phi$, $T_2 \vdash \mathfrak{D}\phi$: that is, if $\mathfrak{D}$ converts all consequences of T_1 into consequences of T_2. In such a case, we will write $\mathfrak{D} : T_1 \to T_2$. Then, theories T_1 and T_2 are translationally equivalent if there are translations $\mathfrak{D} : T_1 \to T_2$ and $\mathfrak{D}' : T_2 \to T_1$ such that, for any Σ_1-formula $\phi(x_1, \ldots, x_m)$ and any Σ_2-formula $\psi(x_1, \ldots, x_n)$,[10]

$$T_1 \vdash \forall x_1, \ldots, \forall x_m(\phi(x_1, \ldots, x_m) \leftrightarrow \mathfrak{D}'\mathfrak{D}\phi(x_1, \ldots, x_m)), \tag{2a}$$

$$T_2 \vdash \forall x_1, \ldots, \forall x_n(\psi(x_1, \ldots, x_n) \leftrightarrow \mathfrak{D}\mathfrak{D}'\psi(x_1, \ldots, x_n)). \tag{2b}$$

That is, T_1 and T_2 are translationally equivalent if there are translations between them that are 'almost inverse': the compositions of the two translations need not take every formula back to itself, but must take it to a formula that is equivalent (modulo T_1 or T_2, as appropriate). We will refer to a pair, $(\mathfrak{D}, \mathfrak{D}')$, satisfying Equation (2) as a translational equivalence between T_1 and T_2.[11]

A symmetry is then simply a translational equivalence for the special case $T_1 = T_2$.[12] Of course, for any theory, the trivial translational equivalence (Id, Id) is a symmetry. But many theories have non-trivial such symmetries. For example, in the theory T_H, consider the dictionary map $\mathfrak{E}$ such that

$$\mathfrak{E}(L) = Rx, \tag{3a}$$

$$\mathfrak{E}(R) = Lx. \tag{3b}$$

It is easy to see that $\mathfrak{E}$ is a translational equivalence between T_H and itself, with $\mathfrak{E}$ as its own inverse.

Here is what I mean by a theory formulated as a local field theory. The role of a signature is played by a set Ψ of q field-variables $\psi_1, \ldots, \psi_q$, and a set X of n base-variables $x^1, \ldots, x^n$. The role of Σ-pictures is played by Ψ-fields, where a Ψ-field is a map from $\mathbb{R}^n$ to $\mathbb{R}^q$.[13] We will use the field-variables as

---

[10] It is crucial that $\mathfrak{D}$ and $\mathfrak{D}'$ be translations. For instance, suppose that Σ_1 and Σ_2 are a pair of signatures such that $\mathfrak{D}$ is a one-to-one arity-preserving bijection between them (or rather, is the dictionary map corresponding to such a bijection), and that $\mathfrak{D}'$ is the inverse (strictly, is the dictionary map corresponding to the inverse). Then the conditions below will be satisfied with respect to any pair of theories T_1 and T_2; but $\mathfrak{D}$ and $\mathfrak{D}'$ will not, in general, be translations.

[11] cf. Glymour's work on definitional equivalence as a necessary condition for theoretical equivalence (for example, Glymour [1970]).

[12] The general case is also of interest to philosophy of physics. In particular, one promising way of understanding duality is in terms of intertranslatability; see (Huggett [2017]; Rickles [2017]). I will not discuss here whether and to what extent the analysis of this article might apply also to the case of dualities.

[13] As such, I am mostly only considering local field theories where the base space is simply connected (though see the discussion of $\tilde{T}_A$ in Section 3).

coordinates for the copy of $\mathbb{R}^q$, that is, the range of the Ψ-fields, and the base-variables as coordinates for the copy of $\mathbb{R}^n$, that is, the domain of the Ψ-fields. The role of Σ-sentences is played by differential equations, constructed by the application of standard differential operators (associated to the base variables in X) to the field variables in Ψ. A Ψ-field is a solution of a differential equation just in case it satisfies the equation at every point of $\mathbb{R}^n$. A Ψ-theory T is just a set of differential equations constructed from Ψ in the manner described. A Ψ-field is a model of T if it is a solution to every member of T; we will denote the class of all models of T by $\mathrm{Mod}(T)$.

For example, consider the theory, T_P, of instantaneous electrostatics in terms of potentials. The field variables of this theory are ρ and ϕ (so $q = 2$), and the base variables are x^1, x^2, and x^3 (so $n = 3$). This theory has one equation:

$$\nabla^2 \phi = 4\pi\rho, \tag{4}$$

where $\nabla = \left(\partial/\partial x^1, \partial/\partial x^2, \partial/\partial x^3 \right)$.

Alternatively, consider the theory T_A of electromagnetism in terms of potentials. The field variables of this theory are A_μ and J^μ, and the base variables are x^μ, with $0 \leq \mu \leq 3$ (so $q = 8$ and $n = 4$). This theory has four equations:

$$\partial_\mu(\partial^\mu A^\nu - \partial^\nu A^\mu) = J^\nu, \tag{5}$$

with $0 \leq \nu \leq 3$. In the above, the Einstein summation convention is used, and raised indices are raised by application of the (inverse) Minkowski matrix $g^{\mu\nu}$ (for example, $A^\mu := g^{\mu\nu} A_\nu$), where

$$g^{\mu\nu} = \begin{cases} 1 & \text{if } \mu = \nu = 0 \\ -1 & \text{if } \mu = \nu = 1, 2, 3 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

For such a theory, the notion of symmetry we shall consider is this: a smooth diffeomorphism $\eta : \mathbb{R}^n \times \mathbb{R}^q \rightarrow \mathbb{R}^q$ that maps solutions to solutions. (Thus, we are only considering so-called internal symmetries.) To spell this out, it is helpful to think of η in terms of its curried form, that is, as a map of type $\mathbb{R}^n \rightarrow (\mathbb{R}^q \rightarrow \mathbb{R}^q)$: that is, η assigns, to every point $p \in \mathbb{R}^n$, a map $\eta_p : \mathbb{R}^q \rightarrow \mathbb{R}^q$.[14] Any such map η naturally induces a transformation of any Ψ-field into another Ψ-field: if the original field has the value $\psi_i(p)$ at point $p \in \mathbb{R}^n$ (for $1 \leq i \leq q$), then the new field has value $\eta_p(\psi_i(p))$ at p. We then say that η is a symmetry of T if η induces a bijection on the space $\mathrm{Mod}(T)$ of solutions of T.[15]

---

[14] If we regard the total space $\mathbb{R}^q \times \mathbb{R}^n$ as a (trivial) fibre bundle, then these transformations are precisely the vertical bundle automorphisms. (Automorphisms, that is, of the bundle structure alone—not of any of the further structure of $\mathbb{R}^n \times \mathbb{R}^q$. See the next footnote.)

[15] Note that we have put very few restrictions on what structure η is to preserve: it need not be (pointwise) linear, nor respect the decomposition of $\mathbb{R}^q$ into clusters of field variables, and so on.

For example, in the theory T_P, let k be some real number, and let η be the map such that

$$\eta(\mathbf{x}, \phi, \rho) = (\phi + k, \rho). \tag{7}$$

One can quickly verify that Equation (7) transforms solutions of Equation (4) into other solutions. For another example, in the theory T_A, let $\lambda : \mathbb{R}^4 \to \mathbb{R}$ be any smooth scalar function, and let η be the map such that

$$\eta(x^\mu, A_\mu, J^\mu) = \left(A_\mu + \partial_\mu \lambda(x^\mu), J^\mu\right). \tag{8}$$

Note that unlike Equation (7), this η depends non-trivially upon the base-variables. Again, one can verify (although a little less straightforwardly) that this transforms all and only solutions of Equation (5) into other solutions.

At first, it might seem a little opaque how these two notions of symmetry relate to one another. In fact, however, there is good reason to think that they are expressions of the same basic idea. To see this, observe first that a dictionary map $\mathfrak{D} : \Sigma \to \Sigma$ could be thought of as a map from the 'value-space' of a first-order theory to itself: if we regard predicates (both simple and complex) as indicating ways for n-tuples to be, and think of field-values as indicating ways for points of a base space to be, then we can see how a dictionary map $\mathfrak{D}$ and a vertical bundle automorphism η do the same kind of thing.[16,17] Furthermore, just as η transforms Ψ-fields into other Ψ-fields, so $\mathfrak{D}$ transforms Σ-pictures into other Σ-pictures. In fact, any dictionary map $\mathfrak{D} : \Sigma_1 \to \Sigma_2$ naturally induces a dual map $\mathfrak{D}^*$ from Σ_2-pictures to Σ_1-pictures. Given any Σ_2-picture M, $\mathfrak{D}^*M$ is the Σ_1-picture given by

$$|\mathfrak{D}^*M| = |M|, \tag{9a}$$

$$\Pi^{\mathfrak{D}^*M} = (\mathfrak{D}\Pi)^M, \tag{9b}$$

where the extension in M of an arbitrary formula ϕ, denoted ϕ^M in Equation (9b),

The reason to be liberal is just that we are in the game of using symmetries to figure out what structure in the theory is important (that is, as we shall see, invariant under symmetry). So as far as possible, we should avoid antecedently privileging structures by making them the kinds of things that symmetries must leave invariant by definition.

[16] Though this interpretation of field-values is not mandatory. Indeed, in some cases it seems rather inappropriate: the possible values of a mass density function at a given spacetime point, for instance, seem most perspicuously described as properties of the matter that could be located at that spacetime point, rather than properties of the point itself. I borrow this observation from Weatherall ([unpublished]). One could either seek a different way of expressing the analogy, or just take the analogy to hold only for those fields that are best thought of as expressing properties over spacetime points.

[17] Though I'm glossing over an important disanalogy. The points of the base space, unlike the members of the domain, form a structured set. It is for this reason that I confine my attention here to internal rather than external symmetries: it is rather more opaque what the model-theoretic analogue of an external symmetry might be.

is defined as the set of those n-tuples from $|M|$ that satisfy ϕ in M (where the order of the n-tuples is fixed by the ordering of the free variables in ϕ).

Finally, we have the following important result from model theory[18]:

Proposition 1: Suppose that we have translations $\mathfrak{D} : T_1 \rightarrow T_2$ and $\mathfrak{D}' : T_2 \rightarrow T_1$. Then $\mathfrak{D}$ and $\mathfrak{D}'$ implement a translational equivalence between T_1 and T_2 if and only if $\mathfrak{D}^*$ is a bijection $\mathrm{Mod}(T_2) \rightarrow \mathrm{Mod}(T_1)$, with $(\mathfrak{D}')^*$ as its inverse.

See the appendix for proof.

Thus, in the special case where $T_1 = T_2$, the demand that a dictionary map $\mathfrak{D}$ be a translational equivalence is the same as demanding that it (or rather, its induced map $\mathfrak{D}^*$) act as a bijection on $\mathrm{Mod}(T)$. This parallels the characterization of symmetries in local field theories as those vertical bundle automorphisms that take solutions to solutions. Hopefully, this is enough to suggest that we are indeed dealing with a reasonably unified concept here. To some extent, the remainder of this article should serve as a further defence of the claim that they are analogous: as we shall see, the same issues show up in the one case as in the other.

With this set-up complete, let us turn to the main task. In this article, I will suppose that, at least under certain circumstances and for certain theories, the following claim is true:

For a theory containing symmetries, we should not interpret that theory in such a way that the symmetry-related models (that is, models related by a map induced by a symmetry) represent distinct ways for the world to be.[19]

What those circumstances might be (indeed, whether there are such circumstances) is contentious, as is the question of why symmetries, under those circumstances, warrant such interpretational circumspection.[20] As I have set things up here, there is good reason to think that this lesson should not be applied to all symmetries in the sense defined above. For instance, the transformation

$$A_\mu \mapsto A_\mu + B_\mu, \tag{10}$$

---

[18] This result is standard (although the statement of it has been tweaked to mesh with the above definition of translational equivalence); see, for example, (de Bouvère [1965], Theorem 2) or (Hodges [1997], p. 54). The proof is provided for convenience (especially since the standard proofs are attuned to the statement in terms of definitional rather than translational equivalence).

[19] Note that the 'should' here is with respect to norms of theory choice, not norms of historical fidelity. The claim is not supposed to be that, when working on Newton's thought, one should not interpret the Principia as committed to absolute space. I thank a referee for pressing me on this point.

[20] The literature addressing these questions is very large; see, for example, (Saunders [2003a]; Brading and Castellani [2003]; Baker [2010]; Caulton [2015]; Dasgupta [2016]) and references therein.

$$J^{\mu} \mapsto J^{\mu} + \partial_{\nu}(\partial^{\nu} B^{\mu} - \partial^{\mu} B^{\nu}), \tag{11}$$

is a symmetry of T_A, but it is best interpreted as enacting a real physical change.[21] (My own view is that the issue here is not the definition of symmetries, but rather the fact that we are dealing with partial or incomplete toy theories. The transformation in Equation (10), for example, will not be a symmetry once appropriate continuity equations for the charged matter are included.) However, let us set this issue aside, and focus instead on the question of what we should do next. That is, suppose that we do indeed have a theory that contains symmetries, and that we have become convinced—for whatever reason—that the above applies to (some of) those symmetries. What should our next move be?

In the next section, I consider one popular account of what the next move should be. This account says that we should seek a reduced theory: a theory that deals only in quantities that are invariant under the relevant symmetry. After explicating this account, I offer some reasons to think that this is not the best way of implementing the above lesson. In Section 4, I consider an alternative way of implementing the lesson above: that of leaving the syntactic sentences of the theory alone, but seeking instead a different semantics for interpreting it (what I will call a sophisticated semantics). Section 5 discusses how the results of applying these two strategies compare to one another.

This all presumes, of course, that there should be a next move at all. Why not take some more quietist attitude, and remark merely that the symmetry-related models are to be understood as representing the same way for the world to be? I don't have any particular beef with such an attitude; indeed, as we shall see later, providing a sophisticated semantics may be thought of as a way of taking this attitude in a mathematically careful way. But it would be a mistake to presume, from the off, that such an attitude can be taken without causing problems. After all, models related by a symmetry can appear very different: the electromagnetic potential will take a very different form after the application of a gauge transformation, for example! So it behoves us to underwrite this attitude, by appropriate mathematical analysis, and show that it is of good conceptual standing.

3 Reduction

In many discussions about the proper way to implement the above interpretational principle for symmetries, it is taken for granted that what we seek is a theory that is the result of a reduction by the relevant symmetry. In very general terms, the idea is that we (i) identify some collection of invariants of the original theory; (ii) specify a theory in terms of those invariants; and

---

[21] I thank an anonymous referee for this example.

(iii) show that the new theory captures all the symmetry-invariant content of the old theory. Before getting more specific, it will be best to introduce examples.

First, consider the case of the handedness theory T_H. The invariant that we use to specify our reduced theory is the congruence relation: that is, we introduce a relation C that is defined by

$$\forall x \forall y (Cxy \leftrightarrow ((Lx \wedge Ly) \vee (Rx \wedge Ry))). \tag{12}$$

Informally, congruence is just the relationship that holds between two objects if and only if they have the same handedness. Let us use θ_C as a shorthand for the formula in Equation (12). If we supplement T_H by this definition, then we get its definitional extension $T_H^+ := T_H \cup \{\theta_C\}$, in signature $\Sigma_H^+ = \{L, R, C\}$. The first observation is that agreement on the congruence relation suffices for agreement on all invariant content, in the following sense: if M and N are two models of T_H^+, such that $|M| = |N|$ and $C^M = C^N$, then either $M = N$, or else $M = \mathfrak{C}^* N$—where $\mathfrak{C}$ is the 'handedness swap' dictionary map introduced earlier, by Equation (3).

Now consider the theory T_C, in signature $\Sigma_C := \{C\}$, comprised by the following axioms:

$$\forall x Cxx, \tag{13a}$$

$$\forall x \forall y (Cxy \rightarrow Cyx), \tag{13b}$$

$$\forall x \forall y \forall z ((Cxy \wedge Cyz) \rightarrow Cxz), \tag{13c}$$

$$\forall x \forall y \forall z ((\neg Cxy \wedge \neg Cyz) \rightarrow Cxz). \tag{13d}$$

Informally, this theory states that C is an equivalence relation, with at most two equivalence classes. Models of T_C closely correspond to models of T_H^+ (and hence, to models of T_H). On the one hand, for any model M of T_H^+, its reduct $M|_{\Sigma_C}$ is a model of T_C. Indeed, suppose that $M \vDash T_H^+$; then M satisfies Equations (1) and (12); but Equation (13) of T_C are simply a consequence of those sentences, and so M must make Equation (13) true as well; since these refer only to C, it follows that $M|_{\Sigma_C} \vDash T_C$. On the other hand, for any model N of T_C, there is a Σ_H^+-expansion N^+ of N (that is, a Σ_H^+-picture N^+ such that $N^+|_{\Sigma_C} = N$) that is a model of T_H^+. Indeed, if N is a model of T_C, then it is clear from Equations (13a–13c) that C^N is an equivalence relation over $|N|$, and from Equation (13d) that it partitions the domain into at most two equivalence classes. So just let L^{N+} be one of these equivalence classes, and let R^{N^+} be the other (if such there be). It is then obvious that N^+ satisfies Equations (1) and (12), that is, that $N^+ \vDash T_H^+$.

Thus, there is a natural sense in which T_C captures the 'invariant part' of T_H. On the one hand, any models of T_H that agree with respect to all the structure

invariant under $\mathfrak{C}^*$ will correspond to a single model of T_C; and on the other, every model of T_C corresponds to some (indeed, more than one) model of T_H.

Second, consider the case of electrostatics. This time, the chosen invariant is the electric field $\mathbf{E}$, defined by

$$\mathbf{E} := \nabla \varphi. \tag{14}$$

Again, the first thing we want is some kind of indication that the electric field suffices to capture all the invariant content of the electrostatic theory. So, let T_P^+ be the definitional extension of T_P by Equation (14), and suppose that M and N are two models of T_P^+, such that $\mathbf{E}^M = \mathbf{E}^N$. Then by elementary integration, their potentials agree to within a symmetry transformation—that is, for some constant k,

$$\varphi^N = \varphi^M + k. \tag{15}$$

So now consider the following theory, T_E. The field-variables of T_E are ρ and E_i (where $1 \leq i \leq 3$); I will use vector notation, $\mathbf{E}$, for the latter. The base-variables are the same as T_P. The equations of the theory are

$$\nabla \times \mathbf{E} = 0, \tag{16a}$$

$$\nabla \cdot \mathbf{E} = 4\pi\rho. \tag{16b}$$

Again in analogy to the handedness case, we have the following pair of observations about how the models of T_E relate to those of T_P. First, for any model M of T_P^+, the electric field $\mathbf{E}^M$ satisfies Equation (16). This is obvious just from plugging the definition Equation (14) into Equation (16). Second, for any model N of T_E, there is a model N^+ of T_P^+ such that $\mathbf{E}^{N^+} = \mathbf{E}^N$. This is also standard: an irrotational vector field over a simply connected base space admits some scalar field of which it is the gradient.

Finally, consider the case of electromagnetism. The invariant we use here is the electromagnetic field:

$$F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{17}$$

Let T_A^+ be the result of supplementing T_A with the definition in Equation (17). Once again, we observe first that the electromagnetic field determines all gauge-invariant quantities. That is, for any models $\left(A_\mu, F_{\mu\nu}\right)$ and $\left(A'_\mu, F'_{\mu\nu}\right)$ of T_A^+, if $F_{\mu\nu} = F'_{\mu\nu}$ then for some scalar function λ, $A'_\mu = A_\mu + \partial_\mu\lambda$. This is, again, a standard result.

Now consider the theory T_F. The field-variables of T_F are J^μ and $F_{\mu\nu}$, where $0 \leq \mu, \nu \leq 3$ (so $q = 20$), whilst the base variables are the same as those of T_A. The equations are

$$F_{\mu\nu} = -F_{\nu\mu}, \tag{18a}$$

$$\partial_{[\mu}F_{\nu\rho]} = 0, \tag{18b}$$

$$\partial_{\mu}F^{\mu\nu} = J^{\nu}, \tag{18c}$$

where, again, indices (all of which range from 0 to 3) are raised using the Minkowski matrix (and the square bracket [...] indicates anti-symmetrization). Then, once more, we find a certain kind of alignment between the models of T_F and the models of T_A^+. That is, for any model M of T_A^+, the field $F_{\mu\nu}^M$ is a solution of Equation (18); and for any model N of T_F, there is a model N^+ of T_A^+ such that $F_{\mu\nu}^{N^+} = F_{\mu\nu}^N$.

These examples make fairly clear what is meant by a reduced theory; let us now offer a general definition. Suppose that T is the target theory, admitting some group G of symmetries (and let us denote the action of $g \in G$ on models by $M \mapsto g^*M$). Say that a collection Q of symmetry-invariant quantities/predicates (in T, or in some definitional extension T^+) is complete if agreement on Q guarantees agreement to within G—that is, if it is the case that for any models M and N of $T^{(+)}$, if $q^M = q^N$ for every $q \in Q$, then for some $g \in G$, $N = g^*M$. A reduction of T to Q is a theory T', of (complete) signature Q, such that

  (i)   for any model M of T', there exists some model N of $T^{(+)}$ such that for every $q \in Q$, $q^M = q^N$; and

  (ii) for any model M of $T^{(+)}$, the reduct of M to Q is a model of T'.

I'll refer to the pair of conditions (i) and (ii) as the Goldilocks conditions for symmetry reduction: they state that the class of models of the reduced theory must be neither too big nor too small.

Many discussions of symmetry assume, implicitly or explicitly, that changing one's theory to incorporate the lessons of a symmetry—to get rid of the 'surplus structure' the symmetry reveals—means moving from the original theory to a reduced theory. It is worth pointing out, however, that there are problems with making reduction the gold standard for expunging surplus structure. First, it is highly non-trivial to find such a reduced theory—or even to demonstrate with confidence that such a theory could exist. All the examples above were chosen as cases where we know how to specify the reduced theory. But doing so required that we could both find a complete set of invariant quantities Q, and then provide a theory in terms of Q whose class of models meets the Goldilocks conditions. Note that these tasks are somewhat in tension. Plausibly, the set of all invariant quantities will always be complete.[22] But the more invariant quantities one wants to use in Q, the

---

[22] Note that proving this will not be entirely straightforward: one could imagine certain global obstructions (for example, topological issues) that might yield a pair of models agreeing on all invariants, yet lying on different symmetry orbits.

harder it is going to be to construct a finitely or recursively axiomatizable theory out of them (satisfying both the Goldilocks conditions).[23]

As an illustration of these perils, consider the theory $\tilde{T}_A$. The equations of this theory are precisely the same as those of T_A: the only difference is that models of this theory are now taken to be maps of the form $U \to \mathbb{R}^{20}$, where U is permitted to be any open subset of $\mathbb{R}^4$. So, in particular, models of this theory include cases where the base space is topologically non-trivial. It is now no longer the case that the set $\{F_{\mu\nu}\}$ comprises a complete set of quantities: there are gauge-invariant quantities that are not determined by fixing the value of $F_{\mu\nu}$ everywhere. To take the best-known example, define the holonomy of a loop γ to be

$$h(\gamma) = \exp\left(\oint_\gamma A_\mu dx^\mu \right). \tag{19}$$

It is straightforward to verify that holonomies are gauge-invariant. Yet if U is not simply connected, the value of $F_{\mu\nu}$ everywhere in U underdetermines the values of the holonomies: two models of $\tilde{T}_A$ (both with base space U) might agree on the former, yet disagree on the latter.[24] Of course, this does not mean that there can be no reduced theory of $\tilde{T}_A$. It certainly doesn't mean that there is no complete set of invariant quantities for $\tilde{T}_A$: in fact, it can be shown that the set of all holonomies comprises just such a complete set.[25] However, it remains very much an open question whether one can give some closed-form set of equations for holonomies, such that the solutions of those equations satisfy the Goldilocks conditions (relative to the definitional extension of T_A by Equation (19)).[26]

The second problem with insisting that one must provide a reduced theory is that, even if such a theory can be found, that theory may seem to have explanatory deficits relative to the original theory. For the reduced theory treats the invariant quantities Q as primitives; this means that if some $q \in Q$ obeys some non-trivial condition as a result of its definition (in the unreduced theory), it must be asserted to obey that condition (in the reduced theory) as a simple posit. Let us consider some examples of this phenomenon.

---

[23] The rider 'finitely or recursively axiomatizable' is necessary to rule out theories consisting simply of all the logical consequences of T expressible in terms of Q. (I thank Teruji Thomas for drawing this to my attention.) Of course, in the context of first-order theories, Craig's theorem prevents this from being a serious restriction; but in richer formalisms (such as local field theory) the rider has bite.

[24] This fact is the essential kernel of the Aharonov–Bohm effect (Aharonov and Bohm [1959]); for further details, see (Healey [2007]).

[25] At least, this is true when—as here—we presuppose a fixed coordination on the base space. In more general contexts, settling issues about the reducibility of gauge potentials to holonomies turns upon subtleties concerning the base points of the holonomy maps; see (Rosenstock and Weatherall [2016]). (My thanks to a referee for bringing this to my attention.)

[26] See (Loll [1994]) for discussion.

For the handedness theory, note that the reduced theory T_C includes axioms to the effect that C is an equivalence relation. No such axioms are needed in the theory T_H^+, since—in that theory—the definition of C, in Equation (12), entails that it is an equivalence relation. For example, the claim that C is symmetric becomes, when translated using Equation (12), the tautology:

$$\forall x \forall y (((Lx \wedge Ly) \vee (Rx \wedge Ry)) \rightarrow ((Ly \wedge Lx) \vee (Ry \wedge Rx))). \qquad (20)$$

Thus, there is some sense in which the fact that C is an equivalence relation is rendered 'more necessary' in T_H^+ than in T_C: it is, we might say, an analytic or definitional necessity rather than a 'mere' law.[27] How seriously one should take this difference will be a matter of taste. But it is a *prima facie* reason for thinking that the passage to the reduced theory has not been entirely without cost—and so for wondering if there might be some way to remove the surplus structure without paying these dues.

In the case of electrostatics, one can see that Equation (16b) in T_E corresponds to Equation (4) of T_P. Equation (16a) is a new addition, however; again, the reason it is not needed in T_P is because, translated using Equation (14), it becomes the mathematical truth that

$$\nabla \times \nabla \varphi = 0. \qquad (21)$$

As such, the unreduced theory can be considered to contain an answer to the question 'why does the curl of the electric field vanish?'—namely, that the electric field is the gradient of a scalar field, and so (like all such gradients) has vanishing curl. Again, the explanatory claim here is contestable, but the intuition that definitions possess explanatory force is in accord with popular views about explanation in philosophy of science. The intuition could be made out in intensional terms (since mathematical truths such as Equation (21) contain a greater degree of modal force than physical truths such as Equation (16a)), or in hyperintensional terms (if one thinks that Equation (21) shows that the curl of **E** vanishes 'in virtue of' **E**'s being a gradient); but either way, at least some accounts of explanation will want to count its absence in T_F as a deficit.

This example also demonstrates, incidentally, that this phenomenon is part of what makes finding a reduced theory so hard. In trying to find the reduced

---

[27] That said, one could argue that T_H^+ also contains explanatory deficits relative to T_C. For instance, simply by virtue of being a reduced theory, T_C need not deal in any primitive claims about the behaviour of L and R. (My thanks to an anonymous referee for pressing this point.) But even if we grant this, the point in the text still stands. For the issue here is not whether T_C or T_H^+ have a better explanatory status in absolute terms; rather, it is merely that the passage to T_C involves some explanatory loss (whether or not that is compensated for by other gains). This is sufficient to motivate us to seek a way of expunging surplus structure without incurring that loss.

version of T_P, one might be encouraged by the observation that ϕ only ever appears in Equation (4) in the form $\nabla\phi$—which is a complete invariant. Even then, though, one still has work to do. It's not enough to merely substitute **E** for $\nabla\phi$ in Equation (4); one also has to add in further equations to recapture theorems such as Equation (21).

For electromagnetism, it is Equations (18a) and (18b) that have no counterpart in the unreduced theory T_A; for in that theory, they reduce to the mathematical trivialities that

$$\partial_\mu A_\nu - \partial_\nu A_\mu = -\left(\partial_\nu A_\mu - \partial_\mu A_\nu\right), \tag{22a}$$

$$\partial_{[\mu}\partial_\nu A_{\rho]} = 0. \tag{22b}$$

The list goes on. Any attempt to reduce $\tilde{T}_A$ to holonomies must stipulate that the holonomies obey various identities; relationalist theories of space must posit constraints amongst the spatial relations (for example, the triangle inequality) that merely follow from the definitions of those relations on substantivalist views; and so on and so forth (Arntzenius [2012], Chapter 6; Maudlin [2007], Chapter 3).

4 Sophistication

Is there an alternative, then? Is there some other way of taking on board the above interpretational principle, without seeking out a reduced theory? I suggest that there is. In a slogan, the idea is that we need not insist on finding a theory whose models are invariant under the application of the symmetry transformation, but can rest content with a theory whose models are isomorphic under that transformation. That is, if M and N are symmetry-related models of the unreduced theory, then they give rise to the same model of the reduced theory discussed in the previous section; the proposal is that we instead look for a theory such that M and N give rise to distinct but isomorphic models. Often, however, finding such a theory may mean leaving the syntax of the theory alone, but instead modifying the semantics. To see what I mean by this, let's consider some examples.

First, consider the handedness theory. I introduce the concept of a de-handed picture. A de-handed picture m comprises:

- A set $|m|$
- A two-element set χ^m
- A function $\epsilon^m : \chi^m \to \mathcal{P}(|m|)$

Intuitively, the set χ^m (of 'chiralities for m') takes over the role of the set $\{L, R\}$; the function ϵ (the 'extension function') assigns each chirality to a subset of the

domain.[28] The point of doing so comes in the introduction of a new definition of 'homomorphism' for such pictures: we take a homomorphism $h : m \rightarrow n$ to comprise a map $h_1 : |m| \rightarrow |n|$ and a bijection $h_2 : \chi^m \rightarrow \chi^n$, such that for either $q \in \chi^m$,

$$h_1[\epsilon^m(q)] = \epsilon^n(h_2(q)). \tag{23}$$

In other words, it is no longer the case that isomorphisms must preserve extensions: indeed, since each model m carries its own personal set of chiralities χ^m, there is no longer a well-defined notion of trans-model identity for extensions, and hence no longer a criterion for what it would be to preserve extensions. To compose a pair of such homomorphisms, simply compose the components.

We now stipulate how de-handed pictures are to determine truth-values for sentences of the first-order language over Σ_H. Since there is no model-transcendent set of chiralities, it will no longer be the case that a picture determines an unambiguous truth-value for every sentence of the handedness language. For a sentence like $\exists x L x$, for example, there is no privileged way to determine which of the two chirality-extensions in a de-handed picture ought to count as the extension of L. But this is as it should be, if we are really interested in doing away with the structure that is variant under the symmetry. Sentences that are not invariant under the symmetry are defective, if we do not take symmetry-variant structure seriously. Instead, truth in a de-handed picture m is (generally) relativized to a bijection $V : \{L, R\} \rightarrow \chi^m$. In a certain sense, it is as though the predicate-letters L and R are being treated as second-order variables (although they can only range over χ^m); we will therefore refer to the map V as a second-order variable-assignment. Relative to such an assignment V, and to a first-order variable-assignment v, the truth-values of atomic sentences in a model, m, are determined as follows:

$$m \vDash_{V,v} L x \text{ if and only if } v(x) \in \epsilon(V(L)),$$
$$m \vDash_{V,v} R x \text{ if and only if } v(x) \in \epsilon(V(R)). \tag{24}$$

The clauses for non-atomic sentences are unchanged. (These semantics could fruitfully be compared to either second-order semantics or supervaluationist semantics.) We then obtain the result below.

> Proposition 2: Suppose that ϕ is logically equivalent to $\mathfrak{E}\phi$, let m be a de-handed picture, and let v be a first-order variable-assignment for m. Then for any second-order variable-assignments V and V' for m,

---

[28] A more elegant semantics for this theory would be one in which each element of the domain is simply assigned one chirality or the other: that is, in which we use a function $\gamma^m : |m| \rightarrow \chi^m$ in place of ϵ^m. But that would build satisfaction of the axioms of T_H into the semantics themselves—whereas all I want to do here is construct a semantics that is sophisticated with respect to the symmetries of T_H.

$$m \vDash_{V,v} \phi \text{ if and only if } m \vDash_{V',v} \phi. \tag{25}$$

See appendix for proof.

As a consequence, the truth-value of any parity-invariant formula is unambiguously determined by a de-handed picture (together with a first-order variable-assignment). Note that all the members of T_H are (of course) logically equivalent to their 'swapped' versions. Hence, we can define the de-handed models of T_H as those de-handed pictures that make T_H true. We then obtain the following.

Proposition 3: Suppose that ϕ is equivalent modulo T_H to $\mathfrak{C}\phi$, let m be a de-handed model of T_H, and let v be a first-order variable-assignment for m.[29] Then for any second-order variable-assignments V and V' for m,

$$m \vDash_{V,v} \phi \text{ if and only if } m \vDash_{V',v} \phi. \tag{26}$$

See appendix for proof.

We can therefore take our new theory to be given by the same set of sentences T_H, but where the semantics for those sentences is that just outlined (that is, is done in terms of de-handed pictures, rather than handed pictures).

Next, consider the electrostatic theory. Again, we retain the same set of equations, but change what objects are used to semantically interpret those equations. Rather than taking ϕ to range over $\mathbb{R}$, we instead take it to range over Φ, where Φ is a one-dimensional, oriented, metric affine space (such a space could be defined as a set equipped with a free, transitive of $\mathbb{R}$ as an additive group). Φ has sufficient structure to enable $\partial\phi/\partial x^i$ (for $i = 1, 2, 3$) to be defined, for instance,

$$\frac{\partial\phi}{\partial x^1} := \lim_{\epsilon \to 0} \frac{\phi(x^1 + \epsilon, x^2, x^3) - \phi(x^1, x^2, x^3)}{\epsilon}, \tag{27}$$

and similarly for x^2 and x^3. So $\nabla\phi$, and hence $\nabla^2\phi$, can be defined as well. We can therefore continue to use Equation (4), interpreted as equations governing models of this kind rather than the original kind. The transformation in Equation (7) also still makes sense, but is now an automorphism of Φ. As a result, if two Φ-valued fields are related by the application of such a transformation, they are isomorphic to one another.[30] Moreover, note that it's not just that the symmetry

---

[29] Where, again, $\mathfrak{C}$ is the 'handedness swap' map in Equation (3).
[30] Given two functions $f : U \to V$ and $f' : U' \to V'$, the appropriate definition of morphism is as follows: a pair of morphisms $\alpha : U \to U'$ and $\beta : V \to V'$ such that $\beta \circ f = f' \circ \alpha$. An isomorphism is then an invertible morphism.

transformations of the form of Equation (7) are automorphisms of Φ: every automorphism of Φ is a transformation of the form of Equation (7).

Finally, consider the electromagnetic theory. This time, models of the theory are to be connections on a principal $U(1)$-bundle over $\mathbb{R}^4$.[31] Once more, we retain Equation (5), but now interpreted in a way that makes use only of the more minimalist structure available in the models: A_μ is now interpreted as the vector potential of the target connection relative to some arbitrarily chosen flat connection on the principal bundle. It is a standard exercise in gauge theory to show that any two such flat connections will be related by a gauge transformation (a vertical automorphism of the bundle).[32] Hence it doesn't matter which flat connection we choose as a reference point: if Equation (5) holds with respect to any one such connection, they hold with respect to all of them. And since gauge transformations are vertical automorphisms of the bundle, the action of Equation (8) on the target connection will yield a model isomorphic to the original. Finally, the extension to the theory $\tilde{T}_A$ is straightforward: we simply take models to be connections on principal $U(1)$-bundle over $U \subseteq \mathbb{R}^4$. Unlike the case of $\tilde{T}_F$, these models do indeed contain all the same gauge-invariant quantities as the unsophisticated models. In particular, a connection on a principal $U(1)$-bundles over a topologically non-trivial U fixes the values of all the holonomies over U.

Hopefully, these examples make clear enough what is intended; let us now seek a general characterization. Note that the proposal on the table—that we can do justice to a symmetry using isomorphism rather than invariance—is a generalization of the 'sophisticated substantivalist' method for dealing with spacetime symmetries.[33] With that in mind, let us refer to theories equipped with semantics of this sort as sophisticated (rather than reduced) theories. The striking thing about our sophisticated theories above was that we altered the semantical constructs with respect to which the theory's syntactic conditions were interpreted, rather than those syntactic conditions themselves ('interpret' here meaning merely assign truth-values to sentences, rather than anything more philosophically substantive). Given a theory T, subject to some group G of symmetries, let us use the term 'picture' for any object that (like a Σ-picture or a Ψ-field) can be used to systematically determine the truth-values of sentences in the language of T; to provide a semantics is just to specify both what

---

[31] See (Baez and Muniain [1994]; Healey [2007]; Weatherall [2016a]) for an introduction to the fibre bundle formalism.

[32] See, for example, (Baez and Muniain [1994], Section II.2).

[33] See (Pooley [2006], [2013]). Note, however, that sophisticated substantivalism is often recommended as an attitude to take towards spacetime theories whose diffeomorphism-related models are already isomorphic (as is the case, for example, with respect to modern formulations of general relativity), rather than (as is the case here) urging the introduction of 'new' isomorphisms. But insofar as I'm taking sophistication here to involve the claim that these isomorphic models represent the same possible world, the analogy is good enough for our purposes. I thank an anonymous referee for this observation.

kinds of things count as pictures, and exactly how it is that pictures assign truth-values to sentences of the language. In all the cases above, we moved from the original semantics (comprised of pictures upon which the symmetries of the theory do not act as isomorphisms) to a sophisticated semantics (comprised of pictures upon which the symmetries of the theory do act as isomorphisms). This new semantics assigns determinate truth-values to symmetry-invariant sentences; since the axioms of the theory are (by definition) so invariant, we may characterize some of the sophisticated pictures as models of the new theory. Finally, in all cases, the sophisticated pictures could be systematically obtained from the original ones, through the 'forgetting' of structure. So in all cases, we had an obvious (that is, 'natural' or 'canonical') map from the unsophisticated pictures to the sophisticated ones.

Thus, let us say that sophistication of a theory T (with symmetries G) is exactly this process: the provision of a new semantics, comprised of pictures that are

- related to those of the old semantics by some 'forgetful' map F;
- adequate to assign truth-values to the G-invariant sentences of T's language; and
- such that for any unsophisticated pictures M and M', $M' = g^*M$ for some $g \in G$ if and only if $F(M) \cong F(M')$.

However, this remains somewhat vague. Is there a way to precisify what is meant? Here is one way to do so. Rather than trying to define the objects of the new semantics 'internally', as mathematical structures of such-and-such a kind (paradigmatically, as sets equipped with certain relations or operations), we instead define them 'externally': as mathematical structures of a given kind, but with certain operations stipulated to be homomorphisms (even if they're not 'really' homomorphisms of the given kind). For example, one way to define vector spaces is to define them as sets equipped with operations of addition and scalar multiplication, obeying appropriate axioms. This is the internal method. The alternative is to define them as spaces of the form $\mathbb{R}^k$, with the further feature that linear transformations are declared to be homomorphisms—and in particular, that invertible linear transformations are isomorphisms. This is the external method. It would also be apposite to refer to the internal method as a 'synthetic' approach, and the external method as an 'analytic' approach, following the terminology of synthetic and analytic geometry. Alternatively, one could see the external method as following in the tradition of Klein's Erlangen programme for geometry, and the internal method as falling more under the Riemannian tradition.[34]

Hence, the proposal is that the pictures on the new semantics are simply what we obtain by taking the old objects, and declaring, by fiat, that the

[34] See (Wallace [forthcoming]) for a detailed defence of using the external method for defining spacetime geometry, and for an expansion on the connection to Klein and Riemann.

symmetry transformations are now going to 'count' as isomorphisms.[35] If we consider our examples above, we can see that—in fact—the method for introducing the new semantics was often very much in this vein. In the case of the handedness theory, an alternative way of introducing de-handed models would be to say that de-handed models are constructed from the same formal materials as handed models (that is, from sets equipped with extensions for L and R), but with a new definition for homomorphism, namely, that a homomorphism h from m to n consists of a map $h_1 : |m| \to |n|$ and a bijection $h_2 : \Sigma_H \to \Sigma_H$, such that

$$h_1[L^m] = (h_2(L))^n, \tag{28a}$$

$$h_1[R^m] = (h_2(R))^n. \tag{28b}$$

In the case of electrostatics, I remarked in passing that the space Φ could be most elegantly defined as a set equipped with a free transitive additive action of $\mathbb{R}$; the external method of defining it would simply mean taking that set to be $\mathbb{R}$ itself, and the additive action to be exactly that expressed by Equation (7). The advantage of defining the new semantics externally is that it offers a relatively easy means of characterizing the objects of the semantics, and of the means by which they accord truth-values to sentences of the formal language: simply (as we saw for the handedness case) use the old semantics, then construct a supervaluationist semantics over the members of each equivalence class of isomorphic new objects. So defined, it will certainly meet the conditions required to be a sophistication.

The main disadvantage of this method is that it might seem far too easy. In general, the external method of defining some kind of mathematical structure might be thought to offer less insight into the nature of that structure: it is one thing to know that a vector space consists of precisely those features of $\mathbb{R}^k$ that are invariant under linear transformations, but another to see that those features are exactly the operations of addition and scalar multiplication, as codified by the axioms for a vector space. More ecumenically, one might think merely that both kinds of construction are important for fully understanding the structure—in which case, one would desire an internal construction as well. And it is often very opaque what kind of internal construction will correspond to an external construction. Electromagnetism makes this fairly clear. It is not at all obvious (I contend) that the features of maps $\mathbb{R}^4 \to \mathbb{R}^4$ preserved under gauge transformations in Equation (8) are precisely the features of vector potentials between connections on a $U(1)$ principal bundle.

---

[35] In category-theoretic terms, this amounts to introducing arrows into the category of models corresponding exactly to the symmetry transformations—which is precisely what Weatherall ([2016a]) proposes to do for (gauge) symmetry transformations. I expand upon the relation to Weatherall's proposal in Section 5.

Nevertheless, we could reason as follows: Assuming that one accepts the external method of definition as mathematically legitimate, then its application gives us a way of defining a sophisticated semantics for the theory, by brute force.[36] It then means that we do have a precise target for a sophisticated semantics that is internally defined: we are looking for some internal construction that delivers an equivalent class of structures.[37]

So, now that we have a decent grip on what sophistication means, we should consider its virtues (or vices). Let's begin by considering the two criticisms we levelled at reduced theories: that they are too hard to find, and that they carry an explanatory cost relative to their unreduced versions. Regarding the former, we have just seen that finding a sophisticated semantics will always be easy if we use the external method. And although we don't have any kind of general guarantee that we will thereby be able to find some kind of internal characterization of those structures, we do—as a matter of fact—generally seem to have success in finding them. This isn't terribly mysterious when one appreciates the role that symmetry considerations play in the construction of theories. If we are demanding that the equations of the theory manifest certain symmetries, then the easiest way to ensure that they do is to construct them as equations governing objects upon which the sought-for symmetries act as isomorphisms. As a result, modern theories are typically born sophisticated. (The paradigm case is the construction of Yang–Mills theories as theories governing connections on a principal G-bundle, which then ensures a sophisticated semantics with respect to G acting as a local gauge group.)

As to the latter, we see that the invariants remain definable, even using the sophisticated semantics: the fact that a sophisticated semantics determines unambiguous truth-values for invariant sentences of the language guarantees that the definitions will remain well posed. As a result, the explanation of why the invariants manifest such-and-such features is also preserved. In the handedness theory, for example, it remains the case that congruence is a matter of possessing the same handedness property—and, hence, that congruence is an equivalence relation. The electric field is still definable as the gradient of the potential, even if the latter is taking values in Φ rather than $\mathbb{R}$; so its irrotationality is still explicable as a consequence of its being a gradient. In the case of electromagnetism, one can still understand the definition in Equation (17) of $F_{\mu\nu}$ as the antisymmetric part of the four-gradient of the vector potential (of the target connection relative to an arbitrarily chosen flat reference connection); although it is more insightful to appreciate that this is precisely the definition of the curvature of the connection. Either way, however, the fact

---

[36] Which, to be clear, is in accord with standard mathematical practice.
[37] 'Equivalent' here meaning that they are isomorphic (not just equivalent) as categories.

that $F_{\mu\nu}$ is antisymmetric (Equation (18a)) and governed by the homogeneous Maxwell equation (Equation (18b)) receives a satisfying explanation.

However, sophistication also raises its own questions. The major issue is simply whether it really does succeed in implementing the idea that we should get rid of 'surplus' (that is, symmetry-variant) structure. After all (someone might say), surely the ontology postulated by the sophisticated version is mostly the same as that of the original theory: a pair of properties in the handedness case, an electrical potential in the electrostatic case, and a vector potential (up to arbitrary choice of reference connection) in the electromagnetic case? So how on earth could it be the case that the sophisticated theory is more parsimonious than the original, in the manner required by the symmetry-interpretation link?

There are two components to the answer: one mathematical, and one more metaphysical. The mathematical observation is that the standard way to explicate the idea of mathematical structure is via isomorphism: what it is for a pair of mathematical objects to have the same structure is for them to be isomorphic to one another (cf. Barrett [2015]; Swanson and Halvorson [unpublished]; Weatherall [2016b]). Thus, insofar as we want to defend sophistication's credentials as genuinely 'expurgating structure', we can invoke standard mathematical usage in support. This doesn't mean that there is no alternative construal of 'structure' that would not be so kind to the sophisticate; but the burden is on the opponent of sophistication to explain what that would be, and to justify their departure from its accepted mathematical meaning.

The metaphysical answer is to get clear on what ontological commitment has been relinquished in the passage from an unsophisticated to a sophisticated semantics. Sophisticated substantivalism, the view that originally inspired us, reconciles the existence of spacetime points with a denial of world-multiplicity by appeal to anti-haecceitism.[38] Anti-haecceitists about spacetime points deny that spacetime points are 'modally robust': they deny that there are worlds that instantiate the same distribution of qualitative properties and relations over spacetime points, yet differ only over which spacetime points play which qualitative roles.[39] This suggests a correlative metaphysical manoeuvre here. We should be anti-quidditists, and deny that physical properties are modally robust. We should not believe that there are worlds that instantiate the same structure in their laws, and differ only over

---

[38] See (Pooley [2013]). Note that anti-haecceitism seems to be the doctrine relevant to applying these kinds of thoughts to external symmetries, and anti-quidditism the doctrine required to make this move for internal symmetries. I hope to expand upon this observation in future work.

[39] This formulation is a little unhappy, since it doesn't distinguish the anti-haecceitist from the essentialist. If there is a difference between them, it comes out in what they say about what one gets by 'permuting' the spacetime points whilst leaving the pattern of qualitative roles the same: roughly, the essentialist thinks that this delivers an impossibility, while the anti-haecceitist thinks that this delivers back the possibility with which we began.

which properties play which nomological roles (see, for example, Hawthorne [2001]; Lewis [2009]). As a result, when one has symmetries—that is, when multiple properties in the theory play the same nomological role—their permutation does not yield a new possibility.

Note that this should not be understood as the claim that symmetric properties ought to be identified with one another. The view is not that properties are individuated by nomological profile, so that there can be no two properties with the same profile.[40] Rather, the view is that when there are two properties with the same profile, there is no fact of the matter about which property-instantiation in a given possible world is an instantiation of which property. The handedness case illustrates this idea nicely: in each world, there are two classes of congruence counterparts, each of which is the extension of a handedness property; but there is no preferred way of matching up a congruence class in one world with one in another world, that is, of identifying such pairs of congruence classes as the extensions of 'the same' handedness property as one another.[41] That said, relative to an (arbitrarily chosen) identification of the congruence class in one world with a congruence class in another, there is a privileged way of identifying the remaining congruence classes: they had better be identified with each other, since the distinction between the classes in each model has to be preserved.

For local field theories, we think of the available values of a particular field as the determinates of a determinable property (so this is a property of space-time points); it is to these properties that we apply the anti-quidditist lesson. So, in the case of electrostatics, we are anti-quidditist about the different potential-values: we deny that there is a privileged way of identifying the potentials-properties in one world with those in another. As with handedness, this doesn't mean collapsing all these properties into one (that is, taking all points of Φ to represent the same property). It also doesn't mean denying that there might be privileged relative identifications (relative, that is, to some initial arbitrary identification). For although there is a Φ-automorphism relating any two chosen points of Φ, there is not always a symmetry relating any two chosen pairs of points in Φ (the pairs $\langle \phi, \phi' \rangle$ and $\langle \psi, \psi' \rangle$ can only be mapped to one another by Equation (7) if $\phi' - \phi = \psi' - \psi$). In the electromagnetism case, we can reckon that the available determinates for a spacetime point are represented by the points in the fibre over (the $\mathbb{R}^4$ point representing) that spacetime point. Note that if we do so, we not only deny that there are

---

[40] Compare the discussion in (Hawthorne [2001], Part 3). One could say that the two properties are 'weakly discernible' in (some appropriate generalization of) the sense of (Saunders [2003b]).

[41] Note that the distinction doing the work here is whether it is possible to engage in transworld identification of properties, not whether this transworld identification is mediated by a quiddity or taken as primitive. This suggests that the 'quidditism without quiddities' of (Locke [2012]) is not importantly different from quidditism with quiddities.

privileged ways to identify such properties across worlds—we also deny that there is a privileged way to identify such properties across spacetime points![42]

5 Equivalence

We've now seen three forms that a theory can take (or more carefully, that a formally interpreted theory can take): an unreduced and unsophisticated form (let's call it the vulgar form), in which there are symmetries relating non-isomorphic models; a reduced form, in which there are no symmetries; and a sophisticated form, in which symmetries relate isomorphic models. I now want to look more closely at the relationships between these three forms. In particular, let us look at the question of whether, and to what extent, these theories can be regarded as equivalent.

The only formal criterion of equivalence that we have so far met with in this article is that of translational equivalence. This criterion can only be applied to theories formulated in the framework of first-order model theory.[43] The only two examples that we have of theories in this framework are the (vulgar) handedness theory, T_H, and its reduced counterpart, the congruence theory T_C. For these theories, we can make the following judgement: they are not translationally equivalent, at least not under the dictionary map

$$\mathfrak{F}(C) = ((Lx \wedge Ly) \vee (Rx \wedge Ry)). \tag{29}$$

For, as is easily seen, $\mathfrak{F}^*$ is not a bijection. This is as far as translational equivalence (strictly understood) can take us. None of the electrostatic or electromagnetic theories were formulated in the framework of first-order model theory, nor was the sophisticated handedness theory (since its semantics are different). We therefore seek a more general framework, into which both the first-order and field-theoretic cases might be enfolded.

Weatherall ([2016a]) has recently observed that category theory offers just such a framework. In order to apply category-theoretic resources, we must specify how to characterize the category of models for each theory; this amounts to specifying what counts as a morphism between models. There is a reasonably

---

[42] cf. (Maudlin [2007], Chapter 3). This will be a somewhat strange metaphysics: a possible world does not consist in a distribution of properties over spacetime points (which would correspond to a section of the bundle), but rather—very roughly—in a distribution of local counterpart relations between infinitesimally nearby points (corresponding to a connection on the bundle). However, this is an artefact of the fact that a connection on a principal bundle represents a 'pure' gauge field: a gauge field represented independently of any matter whose dynamics is conditioned by the field. So it should be unsurprising that a solution of this pure theory turns out to represent a pretty strange kind of world.

[43] Although Glymour ([1970]), in the course of defending translational equivalence as necessary for theoretical equivalence, does make some remarks on how it might be extended to local field theories. (To be pedantic, Glymour's concern is with definitional equivalence rather than translational equivalence; but as can be seen in (Barrett and Halvorson [2016]), the two notions coincide for theories with disjoint vocabulary.)

obvious candidate for the morphisms between models of our first-order theories: the relevant notion of homomorphism, whether vulgar or sophisticated. So first, consider the relationship between $\text{Mod}(T_H)$ and $\text{Mod}(T_C)$, considered as categories in this way. We know that the dictionary map $\mathfrak{F}$ induces a map $\mathfrak{F}^*$ on models. Given any $h : M \to M'$ in $\text{Mod}(T_H)$, let $\mathfrak{F}^*h$ just be h itself (considered as a function on the base set; this prescription works because $|\mathfrak{F}^*M| = |M|$). So defined, $\mathfrak{F}^*$ is easily shown to be a functor from $\text{Mod}(T_H)$ to $\text{Mod}(T_C)$. However, it is not an equivalence of categories.[44] More specifically, it is not full: that is, there are objects M, M' of $\text{Mod}(T_H)$ such that the induced map $h \in \text{Hom}(M, M') \mapsto \mathfrak{F}^*h \in \text{Hom}(\mathfrak{F}^*M, \mathfrak{F}^*M')$ is not surjective.

Proposition 4: $\mathfrak{F}^* : \text{Mod}(T_H) \to \text{Mod}(T_C)$ is not full.

See the appendix for proof.

Second, consider the relationship between $\text{Mod}(T_C)$ and $\text{mod}(T_H)$—that is, between the category of models of T_C and the category of sophisticated models of T_H. Again, we can regard the dictionary map $\mathfrak{F}$ as inducing a functor from $\text{mod}(T_H) \to \text{Mod}(T_C)$; just to maintain notational hygiene, call this functor $\mathfrak{F}^\dagger$. Explicitly, for any $m \in \text{mod}(T_H)$, let $\mathfrak{F}^\dagger m$ be the Σ_C-picture such that

- $|\mathfrak{F}^\dagger m| = |m|$
- For any $a, b \in |\mathfrak{F}^\dagger m|$, $\langle a, b \rangle \in C^{\mathfrak{F}^\dagger m}$ if and only if a and b are members of the same element of $\epsilon^m[\chi^m]$

For any $h : m \to n$, let $\mathfrak{F}^\dagger h$ be the map $H : |\mathfrak{F}^\dagger m| \to |\mathfrak{F}^\dagger n|$ such that $H = h_1$. It is straightforward to verify that $\mathfrak{F}^\dagger m \in \text{Mod}(T_C)$, and that $\mathfrak{F}^\dagger h$ is a Σ_C-homomorphism—that is, that $\mathfrak{F}^\dagger$ really is a functor. This time, however, we have

Proposition 5: $\mathfrak{F}^\dagger$ is an equivalence of categories: it is full, faithful, and essentially surjective.

See the appendix for proof.

Finally, what about the relationship between $\text{Mod}(T_H)$ and $\text{mod}(T_H)$? For any vulgar model M, let $\mathfrak{I}^*M$ be the sophisticated model such that

$$
\begin{aligned}
|\mathfrak{I}^*M| &= |M|, \\
\chi^{\mathfrak{I}^*M} &= \{L, R\}, \\
\epsilon^{\mathfrak{I}^*M}(L) &= L^M, \\
\epsilon^{\mathfrak{I}^*M}(R) &= R^M,
\end{aligned}
\tag{30}
$$

---

[44] Just to be clear, this is a distinct result from the fact that T_H and T_C are not definitionally equivalent.

and for any $H : M \to N$, let $\Im^* H$ be such that $(\Im^* H)_1 = H$ (considered as maps on sets), and $(\Im^* H)_2 = \mathrm{Id}_{\{L,R\}}$. We then find:

Proposition 6: $\Im^* : \mathrm{Mod}(T_H) \to \mathrm{mod}(T_H)$ is not full.

See the appendix for proof.

So, we have the following results. First, $\mathrm{mod}(T_H)$ and $\mathrm{Mod}(T_C)$ are equivalent as categories; second, although we don't have a demonstration that $\mathrm{Mod}(T_H)$ is inequivalent to either $\mathrm{mod}(T_H)$ or $\mathrm{Mod}(T_C)$ (since we have not ruled out there is some appropriate functor between them), we have at least shown that the obvious functors will not do the job. Let us now consider local field theories. The relevant notion of morphism between functions $f : U \to V$ and $f' : U' \to V'$ (where U and U', and V and V', are spaces in the same category) is that of a pair of morphisms $\alpha : U \to U'$ and $\beta : V \to V'$; morphisms, that is, in the ambient categories of the relevant spaces. Thus, for the case of a local field theory equipped with a vulgar semantics (that is, interpreted with respect to functions of type $\mathbb{R}^n \to \mathbb{R}^q$), we find the following: the only morphisms are pairs of the kind $\alpha = \mathrm{Id}_{\mathbb{R}^n}$ and $\beta = \mathrm{Id}_{\mathbb{R}^q}$![45] That is, we find that the category of models of such a theory is always a discrete category.[46]

Let's see how this plays out in the case of our electrostatic theories. First, consider the relationship between the categories $\mathrm{Mod}(T_P)$ and $\mathrm{Mod}(T_E)$. Let $\mathfrak{G}^*$ be the functor $\mathrm{Mod}(T_P) \to \mathrm{Mod}(T_E)$ whose action on models is given by (taking the dual of) the definition of Equation (14); its action on morphisms is simply $\mathfrak{G}^*(\mathrm{Id}_M) = \mathrm{Id}_{\mathfrak{G}^* M}$, as required by functoriality (which suffices to determine $\mathfrak{G}^*$, given that we are working with discrete categories). Second, consider the relationship between $\mathrm{mod}(T_P)$ and $\mathrm{Mod}(T_E)$. Let $\mathfrak{G}^\dagger$ be the functor $\mathrm{mod}(T_P) \to \mathrm{Mod}(T_E)$ that acts on models via the dual of Equation (14), and whose action on non-identity morphisms is as follows: given such a morphism $k : M \to M'$, it must be the case that k is a global potential shift, Equation (7), so that $\mathfrak{G}^\dagger M = \mathfrak{G}^\dagger M'$; we take $\mathfrak{G}^\dagger k := \mathrm{Id}_{\mathfrak{G}^\dagger M}$. Finally, let $K : \mathbb{R} \to \Phi$ be any bijection such that K^{-1} is a bijective embedding of Φ into $\mathbb{R}$. We can then define $\mathfrak{K}^* : \mathrm{Mod}(T_P) \to \mathrm{mod}(T_P)$ as the functor such that $\phi^{\mathfrak{K}^* M} = K \circ \phi^M$ (and whose action on the only morphisms in $\mathrm{Mod}(T_P)$—the identity morphisms—is to take them to the corresponding identity morphisms in $\mathrm{mod}(T_P)$).

---

[45] The fact that there are so few morphisms between unsophisticated models is, of course, a product of our decision to work with coordinatized spaces. Since such spaces are very highly structured, there are very few structure-preserving maps. However, I don't believe that the results below hinge on this decision. (Indeed, the remarkable thing is that even in a relatively austere categorical environment, we are still able to establish useful results.)

[46] A category is discrete (at least, as I am using the term here) if and only if its only morphisms are identity morphisms.

We then obtain the following results, in analogy with Propositions 4, 5, and 6:

Proposition 7: $\mathfrak{G}^* : \mathrm{Mod}(T_P) \to \mathrm{Mod}(T_E)$ is not full.[47]

Proposition 8: $\mathfrak{G}^\dagger : \mathrm{mod}(T_P) \to \mathrm{Mod}(T_E)$ is full, faithful, and surjective—that is, it is an equivalence of categories.[48]

Proposition 9: $\mathfrak{K}^* : \mathrm{Mod}(T_P) \to \mathrm{mod}(T_P)$ is not full.

All proofs are given in the appendix. We can do more or less the same thing for electromagnetism, establishing the same trinity of results.[49] Again, these do not indicate that there are no categorical equivalences between $\mathrm{Mod}(T_P)$ and either $\mathrm{Mod}(T_E)$ or $\mathrm{mod}(T_P)$. Indeed, it seems plausible that there will be some functors between these categories that enact such an equivalence. For instance, any functor between $\mathrm{Mod}(T_P)$ and $\mathrm{Mod}(T_E)$ that is bijective on objects will be an equivalence. However, I suspect that any functor that is describable in appropriately systematic terms (that is, that meshes appropriately with respect to the non-categorical characterization of the models) will not be an equivalence. (Proving this formally would have to await a precisification of 'appropriately systematic' or 'meshes appropriately'.) And we do unambiguously have the result that the categories of sophisticated models come out equivalent to the relevant category of reduced models.

All of this suggests some general (if vague) conjectures. Suppose that a theory T admits some group G of symmetries (and that T is unsophisticated with respect to G). Let T' be a reduction of T to some complete set of G-invariants. Let $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T')$ be the categories of models for T and T', respectively, and let $\mathrm{mod}(T)$ be a category of sophisticated models for T. Finally, let's say that a 'reasonable' functor is one that meshes appropriately with the architecture of the models (whatever exactly that gets made out to mean).[50] Then the following conjectures seem plausible:

- There is a reasonable functor $F : \mathrm{mod}(T) \to \mathrm{Mod}(T')$ that is full, faithful, and essentially surjective.
- There are no reasonable functors from $\mathrm{Mod}(T)$ to either $\mathrm{Mod}(T')$ or $\mathrm{mod}(T)$ that are full, faithful, and essentially surjective (or perhaps the stronger claim: there are no such functors that are full).

---

[47] cf. (Weatherall [2016b], Proposition 1).
[48] cf. (Weatherall [2016b], Proposition 2).
[49] For an explicit discussion of the case of electromagnetism, see (Weatherall [2016a], [2016b]).
[50] At least in our examples, reasonableness seems to be a matter of being definable in terms of the syntactic content (for example, being generated by a translation between two theories). Hence, my emphasis on reasonableness accords with recent work on the sometimes-neglected virtues of the syntactic view of theories (Halvorson [2012], [2013]; Lutz [2014a], [2014b]).

Making these conjectures precise would require (i) a more thorough treatment of how to characterize reduction and sophistication in category-theoretic terms, and (ii) a clarification of the notion of 'reasonableness'. I defer doing so to future work; instead, let us consider the philosophical implications of these technical observations.

Begin with the inequivalence between the reduced and unreduced theories (under vulgar semantics). *Prima facie*, this may seem in tension with Weatherall's ([2016a], Section 5) claim that categorical equivalence (of categories of models) is 'a criterion of equivalence that does capture the sense in which [electromagnetism in terms of fields] and [electromagnetism in terms of potentials] are synonymous'. However, there is no serious disagreement here. The equivalence that Weatherall describes is between electromagnetism formulated in terms of fields—what we have been calling T_F—and electromagnetism formulated in terms of potentials, when gauge transformations are counted as morphisms in its category of models. In other words, the equivalence described by Weatherall is precisely the equivalence between the reduced theory, on the one hand, and the unreduced theory under the sophisticated semantics, on the other.

However, this does highlight a reason why one has to be careful in the use of categorical equivalence as a criterion for theory equivalence. Categorical equivalence does not straightforwardly pronounce on the equivalence of theories (conceived of syntactically, as sets of sentences), but rather on the equivalence of theories relative to a certain way of characterizing the models of a theory as a category. In other words, categorical equivalence is a criterion that applies to theories together with a choice of semantics: change the semantics (from a vulgar to a sophisticated semantics, for example) and one will, in general, change the category of models. To be clear, all of this is present in Weatherall's discussion, albeit in a slightly different form. Whereas I have emphasized the need to specify (not just a theory, but also) the semantic structures one intends to use in formally interpreting the theory, Weatherall ([2016a], Section 5) speaks of constructing the category of models of a theory in such a way that we appropriately privilege 'maps that preserve the "physical structure" of a model, in the sense that two models related by such a map are physically equivalent'.[56] I take these to be two ways of getting at the same idea. If one intends to renounce commitment to a certain amount of structure in one's models as 'unphysical', then one had better also think that the role such structure plays in determining the semantic content of the theory is inessential and/or the product of arbitrary convention.

With these clarifications to hand, it does seem right to say that the reduced and unreduced theories are not equivalent. Electromagnetism with fields and electromagnetism with potentials can only feasibly be regarded as equivalent if gauge symmetries are regarded as relating physically equivalent models; but to

compulsion towards setting the theory up one way rather than another, or towards accepting one pattern of explanation among its parts as uniquely privileged.[52]

On either account, though, the case can be made for valuing sophistication over reduction. On some more realist account of explanation (for example, the grounding account), the explanatory virtues of sophistication make it more likely to be the correct account of the (objective) grounding structure of the world. On a more deflationary picture, those virtues make it a more helpful or convenient way of characterizing the structure of the world; even if a reduced theory is picking out the same structure, it will generally do so in a less tractable way. And of course, both accounts will appreciate the fact that sophistication is typically easier to come by than reduction.

6 Conclusion

To wrap up, I will make two remarks about what I have sought to do in this article—and a comment on something I've not done, but that would be worth doing. The main aim has simply been to convince you that fixating on reduction as the only acceptable means of dealing with symmetries is a mistake.[53] If, as I've argued, sophistication rather than reduction is a legitimate way to seek to expurgate symmetry-variant structure, then a number of interesting consequences follow. One is that carrying out that expurgation becomes (in general) somewhat more straightforward: if all we are required to do is provide a sophisticated understanding of the theory (especially if we do so using the external method), then our lives are made substantially easier than if we need to find a reduced theory. Moreover, with more expurgatory options on the table, we can open up new approaches to classic problems concerning symmetry. The debate on the Aharonov–Bohm effect, for example, is often characterized as requiring us to choose between a trilemma of unpalatable ontologies: a locally acting and separable (but not gauge-invariant) ontology of potentials; a locally acting and gauge-invariant (but non-separable) ontology of holonomies; or a separable and gauge-invariant (but non-locally acting) ontology of fields.[54] But the argument here suggests another option: adopting the 'sophisticated' ontology of connections of a principal bundle (or, more carefully, of whatever the metaphysical correlate of such a connection is).[55] I don't claim that doing so will

---

[52] I read Weatherall's 'puzzleball' account of the foundations of physical theories (Weatherall [2012]) as expressing this kind of picture; it is also closely related to Cartwright's 'dappled-world' conception of inter-theoretic relationships (Cartwright [1999]).

[53] In this regard, cf. (Pooley [2013]; Weatherall [2016b], [forthcoming]).

[54] In the sense of having no 'action at a distance'.

[55] cf. Footnote 42.

magically resolve these problems; but it at least enlivens the conceptual geography.[56]

Second, on a more methodological note, I claim that the above illustrates the value of an eclectic approach to formalisms. Rather than alighting on some framework—first-order logic, differential geometry, category theory, or whatever—as the be-all and end-all, we should be pluralistic about what tools are best applied to the formal study of scientific theories. For example, if we want a tight grip on how the derivable consequences of some axioms relate to the models of those axioms, then we should make use of model theory; but, we should bear in mind that virtually no realistic theory will be expressible in those terms. If we want to abstract away and apply a uniform condition for equivalence, then we should characterize our theories as categories; but we should bear in mind that not all of the essential information about a theory is likely to reside in that category we have rendered it as. By shifting between methods and means as circumstances demand, we can discern similarities and analogues between different formalisms, and use these to cross-fertilize our investigations into one area with insights from another.

Finally, something I've not touched on here is the issue of quantization. In the context of gauge theory, one standard reason to reduce a theory is in preparation for quantizing it. How this process should be undertaken, and how it is to be understood, are longstanding topics of discussion in philosophy of physics.[57] So a natural question is how sophistication and quantization relate to one another. Can a sophisticated theory be quantized in the same way a reduced theory can? And if so, is the same quantum theory obtained as a result? Although these are important and interesting questions, there is no space to address them here; I leave them for future work.

Appendix

Proposition 1: Suppose that we have translations $\mathfrak{D} : T_1 \to T_2$ and $\mathfrak{D}' : T_2 \to T_1$. Then $\mathfrak{D}$ and $\mathfrak{D}'$ implement a translational equivalence between T_1 and T_2 if and only if $\mathfrak{D}^*$ is a bijection $\mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$, with $(\mathfrak{D}')^*$ as its inverse.

---

[56] In the Aharonov–Bohm case, for instance, there will be significant subtleties about the sense in which connections are separable: note that specifying a connection on a region U, and a connection on an overlapping region V, generally underdetermines the connection on $U \cup V$ (absent information about how things stand in $U \cup V$).

[57] See, for example, (Henneaux and Teitelboim [1992]; Thébault [2012]; Pitts [2014]).

Proof

First, it is easy to establish by induction that for any Σ_2-picture M, any Σ_1-formula ϕ, and any variable-assignment v over $|M|$,

$$M \vDash_v \mathfrak{D}\phi \iff \mathfrak{D}^*M \vDash_v \phi. \tag{31}$$

Now, assume first that $\mathfrak{D}$ and $\mathfrak{D}'$ implement a translational equivalence between T_1 and T_2. I show that for any $M \in \mathrm{Mod}(T_2)$, $(\mathfrak{D}')^*\mathfrak{D}^*M = M$, that is, that $(\mathfrak{D}')^*\mathfrak{D}^*$ acts on $\mathrm{Mod}(T_2)$ as the identity. The proof that $\mathfrak{D}^*(\mathfrak{D}')^*$ acts on $\mathrm{Mod}(T_1)$ as the identity goes similarly.

So consider any such M. We have immediately that $|(\mathfrak{D}')^*\mathfrak{D}^*M| = |\mathfrak{D}^*M| = |M|$. So now consider any relation-symbol $\pi \in \Sigma_2$. By the above lemma,

$$(\mathfrak{D}')^*\mathfrak{D}^*M \vDash_v \pi\mathbf{x} \text{ if and only if } \mathfrak{D}^*M \vDash_v \mathfrak{D}'\pi\mathbf{x}$$

$$\text{if and only if } M \vDash_v \mathfrak{D}\mathfrak{D}'\pi\mathbf{x}.$$

But since $M \vDash T_2$ and $\mathfrak{D}$, $\mathfrak{D}'$ implement a translational equivalence,

$$M \vDash \forall\mathbf{x}(\pi\mathbf{x} \leftrightarrow \mathfrak{D}\mathfrak{D}'\pi\mathbf{x}),$$

and so $M \vDash_v \mathfrak{D}\mathfrak{D}'\pi\mathbf{x}$ if and only if $M \vDash_v \pi\mathbf{x}$, for any assignment v. Thus, $\pi^{(\mathfrak{D}')^*\mathfrak{D}^*M} = \pi^M$. Thus, $(\mathfrak{D}')^*\mathfrak{D}^*M = M$.

Now, assume that $\mathfrak{D}^*$ and $(\mathfrak{D}')^*$ are mutually inverse. I show that for any Σ_2-formulae ψ,

$$T_2 \vdash \forall\mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathfrak{D}\mathfrak{D}'\psi(\mathbf{x})). \tag{32}$$

So suppose that Equation (32) did not hold. Then there would be some model M of T_2 such that $M \vDash \forall\mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathfrak{D}\mathfrak{D}'\psi(\mathbf{x}))$—that is, such that for some variable-assignment v for M, either $M \vDash_v \psi$ and $M \nvDash_v \mathfrak{D}\mathfrak{D}'\psi$, or vice versa. But by the above lemma, $M \vDash_v \psi$ if and only if $M \vDash_v \mathfrak{D}\mathfrak{D}'\psi$. So by reductio, Equation (32) holds. By similar reasoning, we can show that the parallel claim for T_1 holds; hence, $\mathfrak{D}$ and $\mathfrak{D}'$ are a translational equivalence. $\square$

Proposition 2: Suppose that ϕ is logically equivalent to $\mathfrak{C}\phi$, let m be a dehanded picture, and let v be a first-order variable-assignment for m. Then for any second-order variable-assignments V and V' for m,

$$m \vDash_{V, v} \phi \text{ if and only if } m \vDash_{V', v} \phi. \tag{25}$$

Proof

Clearly, there only are two second-order variable-assignments for m (since χ^m has only two members); so unless $V = V'$ (in which case the proposition is

trivial),   then   we   have   that $V(L) = V'(R)$ and $V(R) = V'(L)$.   Let M and M' be Σ_H-pictures defined as follows:

$$|M| = |M'| = |m|, \tag{33a}$$

$$L^M = \epsilon^m(V(L)), \tag{33b}$$

$$R^M = \epsilon^m(V(R)), \tag{33c}$$

$$L^{M'} = \epsilon^m(V'(L)) = \epsilon^m(V(R)) = R^M, \tag{33d}$$

$$R^{M'} = \epsilon^m(V'(R)) = \epsilon^m(V(L)) = L^M. \tag{33e}$$

In other words, $M' = \mathfrak{C}^* M$. But clearly, $m \vDash_{V,v} \phi$ if and only if $M \vDash_v \phi$, and $m \vDash_{V',v} \phi$ if and only if $M' \vDash_v \phi$. Hence (suppressing reference to v), $m \vDash_V \phi$ if and only if $M \vDash \phi$ if and only if $M \vDash \mathfrak{C}\phi$ if and only if $\mathfrak{C}^* M \vDash \phi$ if and only if $M' \vDash \phi$ if and only if $m \vDash_{V'} \phi$. □

> Proposition 3: Suppose that ϕ is equivalent modulo T_H to $\mathfrak{C}\phi$, let m be a de-handed model of T_H, and let v be a first-order variable-assignment for m.[58] Then for any second-order variable-assignments V and V' for m,

$$m \vDash_{V, v} \phi \text{ if and only if } m \vDash_{V', v} \phi. \tag{26}$$

Proof

As above, but restricting to models of T_H.

> Proposition 4: $\mathfrak{F}^* : \mathrm{Mod}(T_H) \to \mathrm{Mod}(T_C)$ is not full.

Proof

Let M be as follows:

$$|M| = \{0, 1, 2\},$$
$$L^M = \{0\},$$
$$R^M = \{1, 2\}.$$

Since $\mathfrak{F}^* M = \mathfrak{F}^*(\mathfrak{C}^* M)$, we know that $\mathrm{Id}_{\mathfrak{F}^* M} \in \mathrm{Hom}(\mathfrak{F}^* M, \mathfrak{F}^*(\mathfrak{C}^* M))$. If there was some $h : M \to \mathfrak{C}^* M$ such that $\mathfrak{F}^* h = \mathrm{Id}_{\mathfrak{F}^* M}$, then h would need to act as the identity on the underlying set $|M|$. But there is no homomorphism from M to $\mathfrak{C}^* M$ that does this. So there is no such h; thus, the induced map is not surjective. □

Where, again, $\mathfrak{C}$ is the 'handedness swap' map in Equation (3).

Proposition 5: $\mathfrak{F}^{\dagger}$ is an equivalence of categories: it is full, faithful, and essentially surjective.

Proof

First, I introduce a helpful abbreviation. For any $m \in \mathrm{mod}(T_H)$, and any $a \in |m|$, a is in the extension of one element of χ^m or the other (but not both). So let a^m denote the element of χ^m of whose extension a is a member (that is, let a^m denote the chirality of a).

Now, consider any $m, n \in \mathrm{mod}(T_H)$, and let H be any homomorphism from $\mathfrak{F}^{\dagger}m$ to $\mathfrak{F}^{\dagger}n$. Now define $h_1 : |m| \rightarrow |n|$ by the condition that $h_1 = H$ (as a map on sets). Then, letting a be some arbitrary element of $|m|$, define $h_2 : \chi^m \rightarrow \chi^n$ as the (unique) bijection such that $h_2(a^m) = (h_1(a))^n$; it is easily seen that this does indeed uniquely determine h_2, and that it does so independently of the choice of a. So, for each $q \in \chi^m$ and any $b \in |m|$, if $b \in \epsilon^m(q)$, then $b^m = q$, so $h_2(q) = (h_1(b))^n$, so $h_1(b) \in h_2(q)$. Thus, $h := (h_1, h_2)$ is a homomorphism $m \rightarrow n$, and $H = \mathfrak{F}^{\dagger}h$. So $\mathfrak{F}^{\dagger}$ induces a surjective map on morphisms between any m and n, that is, $\mathfrak{F}^{\dagger}$ is full.

Now consider any $m, n \in \mathrm{mod}(T_H)$, and let $h, h' : m \rightarrow n$ such that $\mathfrak{F}^{\dagger}h = \mathfrak{F}^{\dagger}h'$. Clearly, $h_1 = h_1'$. Furthermore, since h and h' are homomorphisms, it follows that for any $a \in |m|$, $h_2(a^m) = (h_1(a))^n = (h_1'(a))^n = h_2'(a^m)$; hence, $h_2' = h_2$. So $h = h'$. So $\mathfrak{F}^{\dagger}$ induces an injective map on morphisms between any m and n, that is, $\mathfrak{F}^{\dagger}$ is faithful.

Finally, let M by any member of $\mathrm{Mod}(T_C)$. Define a de-handed picture m by setting $|m| = |M|$, letting the members of χ^m be the two congruence classes in M, and setting ϵ^m to the identity. Clearly, $\mathfrak{F}^{\dagger}m = M$. So $\mathfrak{F}^{\dagger}$ is surjective, and therefore essentially surjective. □

Proposition 6: $\mathfrak{I}^* : \mathrm{Mod}(T_H) \rightarrow \mathrm{mod}(T_H)$ is not full.

Proof

It is clear by inspection that $\mathfrak{I}^* = \left(\mathfrak{F}^{\dagger}\right)^{-1} \circ \mathfrak{F}^*$; hence, since $\mathfrak{F}^*$ is not full and $\left(\mathfrak{F}^{\dagger}\right)^{-1}$ is an equivalence, $\mathfrak{I}^*$ is not full either. □

Proposition 7: $\mathfrak{G}^* : \mathrm{Mod}(T_P) \rightarrow \mathrm{Mod}(T_E)$ is not full.[59]

Proof

Let $\mathfrak{C}^*$ be the functor on $\mathrm{Mod}(T_P)$ induced by the symmetry transformation in Equation (7). Let M be any model of T_P. Given the set-up, we know that $\mathfrak{C}^*M \neq M$, and hence that $\mathrm{Hom}(M, \mathfrak{C}^*M) = \varnothing$. Yet we also know that $\mathfrak{G}^*(\mathfrak{C}^*M) = \mathfrak{G}^*M$, and hence that $\mathrm{Hom}(\mathfrak{G}^*(\mathfrak{C}^*M), \mathfrak{G}^*M) \neq \varnothing$ (since it

---

[59] cf. (Weatherall [2016b], Proposition 1).

contains $\text{Id}_{\mathfrak{G}^*M}$). So, the map on arrows induced by $\mathfrak{G}^*$ is not surjective for the pair of objects $M, \mathfrak{C}^*M$—that is, $\mathfrak{G}^*$ is not full. $\qquad\square$

Proposition 8: $\mathfrak{G}^\dagger : \text{mod}(T_P) \to \text{Mod}(T_E)$ is full, faithful, and surjective—that is, it is an equivalence of categories.[60]

Proof

Consider any $m, n \in \text{mod}(T_P)$, and let H by any morphism from $\mathfrak{G}^\dagger m$ to $\mathfrak{G}^\dagger n$. It must be the case that $H = \text{Id}_{\mathfrak{G}^\dagger m}$ (since $\text{Mod}(T_E)$ is discrete). So there are two possibilities: either $m = n$, or m and n are related by some global potential shift k. If the former, then $\mathfrak{G}^\dagger \text{Id}_m = \text{Id}_{\mathfrak{G}^\dagger m}$; if the latter, then $\mathfrak{G}^\dagger k = \text{Id}_{\mathfrak{G}^\dagger m}$. Either way, therefore, $\mathfrak{G}^\dagger$ induces a surjective map on arrows between m and n; so $\mathfrak{G}^\dagger$ is full.

Now consider any $m, n \in \text{mod}(T_P)$, and any morphisms $h, h' : m \to n$. If $m = n$, then $\text{Hom}(m, n) = \{\text{Id}_m\}$; if $m \neq n$, then $\text{Hom}(m, n) = \{k\}$ where k is the (unique) global potential shift relating them; either way, $h = h'$. So (trivially) if $\mathfrak{G}^\dagger h = \mathfrak{G}^\dagger h'$, then $h = h'$. So $\mathfrak{G}^\dagger$ induces an injective map on arrows between m and n; so $\mathfrak{G}^\dagger$ is faithful.

Finally, let $M \in \text{Mod}(T_E)$. As already discussed, for any such model there is some $m \in \text{mod}(T_P)$ such that $\mathfrak{G}^\dagger m = M$. So $\mathfrak{G}^\dagger$ is surjective, and hence essentially surjective. $\qquad\square$

Proposition 9: $\mathfrak{K}^* : \text{Mod}(T_P) \to \text{mod}(T_P)$ is not full.

Proof

It is clear by inspection that $\mathfrak{K}^* = \left(\mathfrak{G}^\dagger\right)^{-1} \circ \mathfrak{G}^*$. Since $\mathfrak{G}^*$ is not full, and $\left(\mathfrak{G}^\dagger\right)^{-1}$ is an equivalence, $\mathfrak{K}^*$ is not full. $\qquad\square$

Acknowledgements

---

[60] cf. (Weatherall [2016b], Proposition 2).

Munich Center for Mathematical Philosophy
Ludwig-Maximilians Universität München
Munich, Germany
neil.a.dewar@gmail.com

References

Aharonov, Y. and Bohm, D. [1959]: 'Significance of Electromagnetic Potentials in the Quantum Theory', *Physical Review*, **115**, pp. 485–91.

Arntzenius, F. [2012]: *Space, Time, and Stuff*, Oxford: Oxford University Press.

Baez, J. C. and Muniain, J. P. [1994]: *Gauge Fields, Knots, and Gravity*, Singapore: World Scientific.

Baker, D. J. [2010]: 'Symmetry and the Metaphysics of Physics', *Philosophy Compass*, **5**, pp. 1157–66.

Barrett, T. W. [2015]: 'On the Structure of Classical Mechanics', *British Journal for the Philosophy of Science*, **66**, pp. 801–28.

Barrett, T. W. and Halvorson, H. [2016]: 'Glymour and Quine on Theoretical Equivalence', *Journal of Philosophical Logic*, **45**, pp. 467–83.

Brading, K. and Castellani, E. [2003]: *Symmetries in Physics: Philosophical Reflections*, Cambridge: Cambridge University Press.

Cartwright, N. [1999]: *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Caulton, A. [2015]: 'The Role of Symmetry in the Interpretation of Physical Theories', *Studies in History and Philosophy of Modern Physics*, **52**, pp. 153–62.

Correia, F. and Schnieder, B. [2012]: *Metaphysical Grounding*, Cambridge: Cambridge University Press.

Dasgupta, S. [2016]: 'Symmetry as an Epistemic Notion (Twice Over)', *British Journal for the Philosophy of Science*, **67**, pp. 837–78.

de Bouvère, K. [1965]: 'Synonymous Theories', in J. W. Addison, L. Henkin and A. Tarski (*eds*), *The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley*, Amsterdam: North-Holland.

Glymour, C. [1970]: 'Theoretical Realism and Theoretical Equivalence', *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **1970**, pp. 275–88.

Halvorson, H. [2012]: 'What Scientific Theories Could Not Be', *Philosophy of Science*, **79**, pp. 183–206.

Halvorson, H. [2013]: 'The Semantic View, if Plausible, Is Syntactic', *Philosophy of Science*, **80**, pp. 475–8.

Hawthorne, J. [2001]: 'Causal Structuralism', *Noûs*, **35**, pp. 361–78.

Healey, R. [2007]: *Gauging What's Real*, Oxford: Oxford University Press.

Henneaux, M. and Teitelboim, C. [1992]: *Quantization of Gauge Systems*, Princeton, NJ: Princeton University Press.

Hodges, W. [1997]: *A Shorter Model Theory*, Cambridge: Cambridge University Press.

Huggett, N. [2017]: 'Target Space = Space', *Studies in History and Philosophy of Modern Physics*, **59**, pp. 81–8.

Lewis, D. [2009]: 'Ramseyan Humility', in D. Braddon-Mitchell and R. Nola (*eds*), *Conceptual Analysis and Philosophical Naturalism*, Cambridge, MA: MIT Press, pp. 203–22.

Locke, D. [2012]: 'Quidditism without Quiddities', *Philosophical Studies*, **160**, pp. 345–63.

Loll, R. [1994]: 'The Loop Formulation of Gauge Theory and Gravity', in J. C. Baez (*ed*.), *Knots and Quantum Gravity*, Oxford: Oxford University Press, pp. 1–20.

Lutz, S. [2014a]: 'Empirical Adequacy in the Received View', *Philosophy of Science*, **81**, pp. 1171–83.

Lutz, S. [2014b]: 'What's Right with a Syntactic Approach to Theories and Models?', *Erkenntnis*, **79**, pp. 1475–92.

Maudlin, T. [2007]: *The Metaphysics within Physics*, Oxford: Oxford University Press.

Møller-Nielsen, T. [2017]: 'Invariance, Interpretation, and Motivation', *Philosophy of Science*, **84**, pp. 1253–64.

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**, pp. 382–406.

Pooley, O. [2006]: 'Points, Particles, and Structural Realism', in D. Rickles, S. French and J. Saatsi (*eds*), *The Structural Foundations of Quantum Gravity*, Oxford: Oxford University Press, pp. 83–120.

Pooley, O. [2013]: 'Substantivalist and Relationalist Approaches to Spacetime', in R. Batterman (*ed*.), *The Oxford Handbook of Philosophy of Physics*, Oxford: Oxford University Press, pp. 522–86.

Putnam, H. [1983]: 'Equivalence', in his *Philosophical Papers*, Volume 3, Cambridge: Cambridge University Press, pp. 26–45.

Redhead, M. L. G. [1975]: 'Symmetry in Intertheory Relations', *Synthese*, **32**, pp. 77–112.

Rickles, D. [2017]: 'Dual Theories: "Same but Different" or "Different but Same"?', *Studies in History and Philosophy of Modern Physics*, **59**, pp. 62–7.

Rosenstock, S. and Weatherall, J. O. [2016]: 'A Categorical Equivalence between Generalized Holonomy Maps on a Connected Manifold and Principal Connections on Bundles over that Manifold', *Journal of Mathematical Physics*, **57**, p. 102902.

Saunders, S. [2003a]: 'Indiscernibles, General Covariance, and Other Symmetries: The Case for Non-eliminativist Relationalism', in A. Ashtekar, D. Howard, J. Renn, S. Sarkar and A. Shimony (*eds*), *Revisiting the Foundations of Relativistic Physics: Festschrift in Honour of John Stachel*, Dordrecht: Kluwer.

Saunders, S. [2003b]: 'Physics and Leibniz's Principles', in K. Brading and E. Castellani (*eds*), *Symmetries in Physics: Philosophical Reflections*, Cambridge: Cambridge University Press, pp. 289–308.

Schroeren, D. [unpublished]: 'Theoretical Equivalence as Explanatory Equivalence'.

Swanson, N. and Halvorson, H. [unpublished]: 'On North's "The Structure of Physics"'.

Thébault, K. P. [2012]: 'Symplectic Reduction and the Problem of Time in Nonrelativistic Mechanics', *British Journal for the Philosophy of Science*, **63**, pp. 789–824.

Wallace, D. [forthcoming]: 'Who's Afraid of Coordinate Systems? An Essay on Representation of Spacetime Structure', *Studies in History and Philosophy of Modern Physics*.

Weatherall, J. O. [2012]: 'Inertial Motion, Explanation, and the Foundations of Classical Spacetime Theories', in D. Lehmkuhl, G. Schiemann and E. Scholz (*eds*), *Towards a Theory of Spacetime Theories*, Basel: Birkhäuser.

Weatherall, J. O. [2016a]: 'Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent?', *Erkenntnis*, **81**, pp. 1073–91.

Weatherall, J. O. [2016b]: 'Understanding Gauge', *Philosophy of Science*, **83**, pp. 1039–49.

Weatherall, J. O. [forthcoming]: 'Regarding the "Hole Argument"', *British Journal for the Philosophy of Science*.

Weatherall, J. O. [unpublished]: 'On the Field Concept in Classical Physics'.