Sophistication about symmetries

January 9, 2016

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 which traffics only in structures invariant under the relevant symmetry. In this paper, 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 widely held that the symmetries of a theory reveal "surplus structure": structure which, in some sense, the theory could do without. (For example, the boost symmetry of Newtonian mechanics indicates the superfluousness of absolute space; the gauge symmetry of electromagnetism reveals the superfluousness of absolute potentials; and so on and so forth.) In this paper, I compare and contrast two ways of taking that claim on board (although I do not intend to assess the scope or validity of the claim itself). The first is to replace the theory by (what I shall call) a *reduced* theory: a theory that deals only in quantities which 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 paper: 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 sophistication as an alternative way to enact those lessons. Finally, in section 5, I discuss the senses in which these really are alternatives (and how the original theory relates to its reduced and sophisticated versions). Section 6 concludes.

2. Symmetry

Here, I outline the kinds of symmetries that will be the topic of this paper. I consider symmetries for two kinds of theories: for theories formulated in terms of first-order 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.[1] 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 $\mathrm{Form}(\Sigma)$ 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).

The semantics for a language with signature Σ is given by Σ-*pictures*.[2] A Σ-picture M consists of a set $|M|$ (the *domain* of M), equipped with a function $\cdot^{\mathcal{M}}$ with domain Σ. For each n-ary predicate $\Pi \in \Sigma$, $\Pi^{\mathcal{M}}$ is a set of n-tuples with members drawn from $|M|$: that is, $\Pi^{\mathcal{M}} \subseteq |\mathcal{M}|^n$. A Σ-picture M determines the truth or falsity of elements of $\mathrm{Form}(\Sigma)$, 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 $\mathcal{M} \models_v \phi$; if ϕ is a sentence, then the variable-assignment no longer matters, and we write simply $\mathcal{M} \models \phi$.

A *theory* T in the signature Σ (for short, Σ-theory) is a set of Σ-sentences.[3] 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 $\mathrm{Mod}(T)$. Finally, T *entails* a Σ-sentence ϕ just in case $M \models \phi$ for every $M \in \mathrm{Mod}(T)$; this will be denoted by $T \vDash \phi$.[4] For example, consider the theory T_H of *handedness*. Letting $\Sigma_H = \{L, R\}$, T_H is the theory consisting of the following

[1] Notation and concepts mostly follows [Hodges, 1997].

[2] 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.

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

[4] The symbols for satisfaction and entailment are unfortunately similar: the former is $\models$, whilst the latter is $\vDash$. Context will make clear what is meant in any given case, however.

sentences:

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

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

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 relevant notion of symmetry is this: a symmetry is a translational equivalence between a theory and itself.[5] First, define a *dictionary map from* Σ_1 *to* Σ_2 to be any function $\mathfrak{D} : \Sigma_1 \to \text{Form}(\Sigma_2)$ 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 $(\mathfrak{D}\Pi)(y_1/x_1, \ldots, y_m/x_m)$.[6] Let us denote the result of applying such a 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 $\mathfrak{D} : T_1 \to T_2$. We can then say what it is for a pair of theories to be translationally equivalent:

Definition 1. Theories T_1 and T_2 are *translationally equivalent* if there are translations[7] $\mathfrak{D} : T_1 \to T_2$ and $\mathfrak{D}' : T_2 \to T_1$ such that, for any Σ_1-formula $\phi(x_1, \ldots, x_m)$, and any

[5] The notion of a translational equivalence is taken from [Barrett and Halvorson, 2015]; it should be noted that translational equivalence is, modulo trivial relabellings of predicates, equivalent to definitional equivalence.

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

[7] 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.

Σ_2-formula $\psi(x_1, \ldots, x_n)$,

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

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

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

A symmetry is then simply a translational equivalence under the special case $T_1 = T_2$. Of course, for any theory, the trivial translational equivalence $(\mathrm{Id}, \mathrm{Id})$ 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

$$\mathfrak{E}(L) = Rx_1 \tag{3a}$$

$$\mathfrak{E}(R) = Lx_1 \tag{3b}$$

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 $\mathbb{R}^q$. We will use the field-variables as 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 using the members of Ψ and standard differential operators. 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 $\nabla = (\partial/\partial x^1, \partial/\partial x^2, \partial/\partial x^3)$.

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 $g^{\mu\nu}$ (e.g. $A^\mu := g^{\mu\nu}A_\nu$), 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 symmetry is a vertical bundle automorphism of the theory's total space which maps solutions to solutions. (Thus, we are only considering so-called *internal* symmetries.) That is, regard the total space $\mathbb{R}^q \times \mathbb{R}^n$ as a (trivial) fibre bundle. Any vertical bundle automorphism can be expressed as a map $\eta : \mathbb{R}^n \to (\mathbb{R}^q \to \mathbb{R}^q)$.[8] That is, η assigns, to every point $p \in \mathbb{R}^n$, a map $\eta_p : \mathbb{R}^q \to \mathbb{R}^q$. 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 of solutions of T.

For example, in the theory T_P, let k be some real number, and for every $p \in \mathbb{R}^3$, let η_p be the map such that

$$\eta_p(\phi, \rho) = (\phi + k, \rho) \tag{7}$$

One can quickly verify that (7) transforms solutions of (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_p be the map such that

$$\eta_p(A_\mu, J^\mu) = (A_\mu + \partial_\mu\lambda|_p, J^\mu) \tag{8}$$

Again, one can verify (although a little less straightforwardly) that this transforms all and only solutions of (5) into other solutions.

[8] This is the same function-type as $\mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^q$; but expressing it in the curried form makes its conceptual import a little clearer.

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 remind ourselves that field-values serve to indicate 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. 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

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

$$(a_1, \ldots, a_n) \in \Pi^{\mathfrak{D}^*M} \iff (a_1, \ldots, a_n) \in (\lambda x_1 \ldots x_n.\mathfrak{D}\Pi)^M \tag{9b}$$

Finally, we have the following important result from model theory:[9]

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 iff $\mathfrak{D}^*$ is a bijection $\mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$, with $(\mathfrak{D}')^*$ as its inverse.

Proof. See Appendix A. □

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 characterisation of symmetries in local field theories as those vertical bundle automorphisms which 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 paper 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 setup complete, let us turn to the main task. In this paper, I will suppose that, at least under certain circumstances and for certain theories, the following claim is true:

[9] 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].

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

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.[10] However, rather than become involved in that debate, I wish to investigate the issue of what we should do next. That is, suppose that we do indeed have a theory which contains symmetries, and to which (for whatever reason) we have become convinced that the above applies. 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 which deals only in quantities which 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 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.

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 which 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 (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 which we use to specify our reduced theory is the *congruence* relation: that is, we introduce a relation

[10] The literature addressing these questions is very large: see, for example, [Saunders, 2003a], the essays in [Brading and Castellani, 2003], [Baker, 2010], [Dasgupta, 2014], [Caulton, 2015], and references therein.

C that is defined by

$$\forall x \forall y (Cxy \leftrightarrow ((Lx \wedge Ly) \vee (Rx \wedge Ry))) \tag{10}$$

Informally, congruence is just the relationship that holds between two objects iff they have the same handedness. Let us use θ_C as a shorthand for the formula (10). If we supplement T_H by this definition, then we get its definitional extension $T_H^+ := T_H \cup \{\theta_C\}$, in signature $\{L, R, C\}$. 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 $M = \mathfrak{E}^* N$.

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

$$\forall x Cxx \tag{11a}$$

$$\forall x \forall y (Cxy \rightarrow Cyx) \tag{11b}$$

$$\forall x \forall y \forall z ((Cxy \wedge Cyz) \rightarrow Cxz) \tag{11c}$$

$$\forall x \forall y \forall z ((\neg Cxy \wedge \neg Cyz) \rightarrow Cxz) \tag{11d}$$

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 $M \models T_H^+$; then M satisfies the sentences (1) and (10); but the sentences (11) of T_C are simply a consequence of those sentences, and so M must make (11) true as well; since these refer only to C, it follows that $M|_{\Sigma_C} \models T_C$. On the other hand, for any model N of T_C, there is a Σ_H^+-expansion N^+ of M (i.e., a Σ_H^+-picture N^+ such that $N^+|_{\Sigma_C} = N$) which is a model of T_H^+. Indeed, if N is a model of T_C, then it is clear from equations (11a)–(11c) that C^N is an equivalence relation over $|N|$, and from (11d) 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 (1) and (10), i.e. that $N^+ \models T_H^+$.

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 which agree with respect to all the structure invariant under $\mathfrak{E}^*$ 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

$$\mathbf{E} := \nabla\varphi \tag{12}$$

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 (12), 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,

$$\varphi^N = \varphi^M + k \tag{13}$$

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

$$\nabla \times \mathbf{E} = 0 \tag{14a}$$

$$\nabla \cdot \mathbf{E} = 4\pi\rho \tag{14b}$$

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 equations (14). This is obvious just from plugging the definition (12) into (14). 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

$$F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu \tag{15}$$

Let T_A^+ be the result of supplementing T_A with the definition (15). Once again, we observe first that the electromagnetic field determines all gauge-invariant quantities. That is, for any models $(A_\mu, F_{\mu\nu})$ and $(A'_\mu, F'_{\mu\nu})$ 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 equa-

tions are

$$F_{\mu\nu} = -F_{\nu\mu} \tag{16a}$$

$$\partial_{[\mu}F_{\nu\rho]} = 0 \tag{16b}$$

$$\partial_{\mu}F^{\mu\nu} = J^{\nu} \tag{16c}$$

where, again, indices (all of which range from 0 to 3) are raised using the Minkowski matrix (and the square bracket $[\ldots]$ indicates anti-symmetrisation). 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 (16); 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: i.e., 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 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.[11] But the more invariant quantities one wants to use in Q, the harder it is going to be to find a finitely or recursively axiomatisable theory out of them (satisfying both the Goldilocks conditions).[12]

As an illustration of these perils, consider the theory $\widetilde{T}_A$. 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 which 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

$$h(\gamma) = \exp\left(\oint_\gamma A_\mu \mathrm{d}x^\mu\right) \tag{17}$$

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 $\widetilde{T}_A^+$ (both with base space U) might agree on the former, yet disagree on the latter.[13] Of course, this does not mean that there can be no reduced theory of $\widetilde{T}_A$. It certainly doesn't mean that there is no complete set of invariant quantities for $\widetilde{T}_A$: in fact, it can be shown that the set of all holonomies comprises just such a complete set. 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^τ by (17)).[14]

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 well have explanatory deficits relative to the original theory. For the reduced theory treats the invariant quantities

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

[12] The rider "finitely or recursively axiomatisable" is necessary to rule out theories consisting simply of all the logical consequences of T expressible in terms of Q. 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.

[13] This fact is the essential kernel of the Aharonov-Bohm effect [Aharonov and Bohm, 1959]; for further details, see [Healey, 2007].

[14] See [Loll, 1994] for discussion.

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.

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 (10) entails that it is an equivalence relation. For example, the claim that C is symmetric becomes, when translated using (10), the tautology

$$\forall x \forall y (((Lx \wedge Ly) \vee (Rx \wedge Ry)) \rightarrow ((Ly \wedge Lx) \vee (Ry \wedge Rx))) \tag{18}$$

In the case of electrostatics, one can see that the equation (14b) in T_E corresponds to the equation (4) of T_P. Equation (14a) is a new addition, however; again, the reason it is not needed in T_P is because, translated using (12), it becomes the mathematical truth that

$$\nabla \times \nabla \varphi = 0 \tag{19}$$

This example also demonstrates that this phenomenon is part of what makes finding a reduced theory so hard. In trying to find the reduced version of T_P, one might be encouraged by the observation that ϕ only ever appears in (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 $\mathbf{E}$ for $\nabla\phi$ in (4); one also has to add in further equations to recapture conditions such as (19).

For electromagnetism, it is the equations (16a) and (16b) which have no counterpart in the unreduced theory T_A; for in that theory, they reduce to the mathematical trivialities that

$$\partial_\mu A_\nu - \partial_\nu A_\mu = -(\partial_\nu A_\mu - \partial_\mu A_\nu) \tag{20a}$$

$$\partial_{[\mu} \partial_\nu A_{\rho]} = 0 \tag{20b}$$

The list goes on. Any attempt to reduce $\widetilde{T}_A$ to holonomies must stipulate that the holonomies obey various identities; attempting to reduce a non-Abelian gauge theory to so-called "Wilson loops" (the relevant analogue of the holonomies for the non-Abelian case) requires positing an even more restrictive set of conditions still.[15] Or con-

[15] See [Arntzenius, 2012, chap. 6].

sider relationalist theories of space, which must posit constraints amongst the spatial relations (e.g. the Triangle Inequality) that merely follow from the definitions of those relations on substantivalist views.[16] Why is it bad for the reduced theory to introduce these extra conditions as primitive posits? Part of the issue is just that it adds to the complexity of those theories. More significantly, though, it seems to remove a certain *explanatory* virtue from the original formulation of the theory. In the unreduced theory, there is a good answer to the question of *why* the invariant quantities obey these conditions: they obey these conditions because of how they are built up out of other kinds of structure in the theory. In the unreduced theory, it seems, we get some kind of insight into these conditions—an insight that risks being lost, or occluded, if we insist that the reduced theory is the be-all and end-all.

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 distinct 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-member multiset[17] $\mathbf{2}^m$, each element of which is a subset of $|m|$

The point of doing so comes in the introduction of a new definition of "homomorphism" for such pictures: we take a homomorphism $h : m \to n$ to comprise a map

[16] See [Maudlin, 2007, chap. 3].

[17] A multiset is like a set, except that elements of a multi-set may occur more than once [Blizard, 1988]. The idea of using multisets in models of this kind is taken from [Lutz, 2015].

$h_1 : |m| \to |n|$ and a bijection $h_2 : \mathbf{2}^m \to \mathbf{2}^n$, such that for each $i \in \mathbf{2}^m$ and any $a \in |m|$,

$$\text{if } a \in i, \text{ then } h_1(a) \in h_2(i) \tag{21}$$

In other words, we relax the requirement that isomorphisms must preserve the extensions of predicates: instead, they may map the extension of one predicate to the extension of the other. To compose a pair of such homomorphisms, simply compose the components.

We now remark on how de-handed pictures determine truth-values. 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 $\exists x Lx$, for example, there is no privileged way to determine which of the two "extensions" in the 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 which 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) relativised to a bijection $V : \{L, R\} \to \mathbf{2}^m$. 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 $\mathbf{2}^m$); 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:

$$\begin{aligned} m \models_{V,v} Lx \text{ iff } v(x) \in V(L) \\ m \models_{V,v} Rx \text{ iff } v(x) \in V(R) \end{aligned} \tag{22}$$

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 following result.

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,

$$m \models_{V,v} \phi \text{ iff } m \models_{V',v} \phi \tag{23}$$

Proof. See Appendix A. □

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 which make T_H true. We then obtain the following.

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

$$m \models_{V,v} \phi \text{ iff } m \models_{V',v} \phi \tag{24}$$

Proof. See Appendix A. □

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 (i.e. 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 $\nabla^2\phi$ to be straightforwardly defined. We can therefore continue to use the equation (4), interpreted as equations governing models of this kind rather than the original kind. The transformation (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.[18] Moreover, note that it's not just that the symmetry transformations of the form (7) are automorphisms of Φ: *every* automorphism of Φ is a transformation of the form (7).

Finally, consider the electromagnetic theory. This time, models of the theory are to be connections on a principal $U(1)$-bundle over $\mathbb{R}^4$.[19] Once more, we retain the equations (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

[18] 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.

[19] See [Baez, 1994], [Healey, 2007], or [Weatherall, 2014a] for an introduction to the fibre bundle formalism.

bundle; it is straightforward to show that any two such flat connections will be related by a gauge transformation (a vertical automorphism of the bundle), and hence that it doesn't matter which flat connection we choose as a reference-point. And, since gauge transformations are vertical automorphisms of the bundle, the action of (8) on the target connection will yield a model isomorphic to the original. Note that the extension to the theory $\widetilde{T}_A$ is straightforward: we simply take models to be connections on principal $U(1)$-bundles over $U \subseteq \mathbb{R}^4$. And these models do indeed contain all the same gauge-invariant quantities as the unsophisticated models: in particular, the such a connection fixes the values of all the holonomies.

Hopefully, these examples make clear enough what is intended; let us now seek a general characterisation. Note that the proposal on the table—that we can do justice to a symmetry using isomorphism rather than invariance—is a generalisation of the "sophisticated substantivalist" method for dealing with spacetime symmetries.[20] With that in mind, let us refer to theories equipped with semantics of this sort as *sophisticated* (rather than reduced) theories. In general, we will suppose that the hallmark of a sophisticated theory is that one leaves the syntactic structure well alone, but alters the semantic structure which is used to interpret those syntactical constructions ("interpret" here meaning merely assign truth-values to sentences, rather than anything more philosophically substantive). That is, suppose that we have some theory T, which (as before) is subject to some group G of symmetries. Let's use the term "picture" to mean an object which (like a Σ-picture or a Ψ-field) can be used to systematically determine the truth-values of sentences in the language of T. Then a *sophistication of T by G* of T's semantics is the "forgetting" of the G-variant structure (but *only* the G-variant structure) from each picture in T's original semantics, thereby obtaining a semantics which is adequate to assign truth-values to the G-invariant sentences of T's language—and which has the feature that if F is the forgetful map, then for any original pictures M and M', $M' = g^*M$ (for some $g \in G$) iff $F(M) = F(M')$.

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

[20] [Pooley, 2006, Pooley, 2013]

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 program for geometry, and the internal method as falling more under the Riemannian tradition.[21]

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 symmetry transformations are now going to "count" as isomorphisms.[22] 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, the re-characterisation of models in terms of multisets was essentially just a means of legitimating the new definition of homomorphism. 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 (7). The advantage of defining the new semantics externally is that it offers a relatively easy means of characterising 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 $\mathbb{R}^k$ which are invariant under linear transformations, but another to see that those features are exactly the op-

[21] See [Wallace, 2015] for a detailed defence of using the external method for defining spacetime geometry, and for an expansion on the connection to Klein and Riemann.

[22] 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, 2015a] proposes to do for (gauge) symmetry transformations. I expand upon the relation to Weatherall's proposal in section 5.

erations 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 (8) are precisely the features of vector potentials between connections on a $U(1)$ principal bundle. Nevertheless, we could reason as follows. Assuming that one accepts the external method of definition as mathematically legitimate,[23] then its application gives us a way of defining a sophisticated semantics for the theory, by brute force. It then means that we do have a precise target for a sophisticated semantics which is internally defined: we are looking for some internal construction which delivers an equivalent class of structures.[24]

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 characterisation 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

[23] Which, to be clear, is in accord with standard mathematical practice.

[24] "Equivalent" here meaning that they are isomorphic (not just equivalent) as categories.

such-and-such features are 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 (15) 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); however, it is more insightful to appreciate that this is precisely the definition of the curvature of the connection. Either way, however, the fact that $F_{\mu\nu}$ is antisymmetric (16a) and governed by the homogeneous Maxwell equation (16b) 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" (i.e., 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.[25] 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 which originally inspired us, reconciles the existence of spacetime points with a denial of world-multiplicity by appeal to *anti-*

[25] cf. [Barrett, 2014]; [Swanson and Halvorson, 2012]; [Weatherall, 2015b].

haecceitism.[26] Anti-haecceitists about spacetime points deny that spacetime points are "modally robust": they deny that there are worlds which instantiate the same distribution of qualitative properties and relations over spacetime points, yet differ only over which spacetime points play which qualitative roles.[27] This suggests a correlative metaphysical manoeuvre here. We should be *anti-quidditists*,[28] and deny that physical properties are modally robust: we should not believe that there are worlds which instantiate the same structure in their laws, and differ only over which properties play which nomological roles. As a result, when one has symmetries—i.e., 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.[29] 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.[30] 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.

[26] 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.

[27] 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, whilst the anti-haecceitist thinks that this delivers back the possibility with which we began.

[28] See e.g. [Lewis, 2009], [Hawthorne, 2001].

[29] Compare the discussion in [Hawthorne, 2001, Part Three]. One could say that the two properties are "weakly discernible" in (some appropriate generalisation of) the sense of [Saunders, 2003b].

[30] 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.

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 spacetime 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 (i.e. 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 (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 privileged ways to identify such properties across worlds—we also deny that there is a privileged way to identify such properties across spacetime points![31]

5. Equivalence

We've now seen three forms a theory can take (or more carefully, which 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 essay is that of translational equivalence. This criterion can only be applied to theories formulated in the framework of first-order model theory.[32] The only two examples

[31] cf. [Maudlin, 2007, chap. 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 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.

[32] Although [Glymour, 1970], in the course of defending translational equivalence as necessary for the-

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 judgment: they are not translationally equivalent, at least not under the dictionary map

$$\mathfrak{F}(C) = ((Lx \wedge Ly) \vee (Rx \wedge Ry)) \tag{25}$$

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 has recently observed that category theory offers just such a framework.[33] In order to apply category-theoretic resources, we must specify how to characterise the category of models for each theory; this amounts to specifying what counts as a morphism between models. There is a reasonably 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 $\mathrm{Mod}(T_H)$ and $\mathrm{Mod}(T_C)$, 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 $\mathrm{Mod}(T_H)$, 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 $\mathrm{Mod}(T_H)$ to $\mathrm{Mod}(T_C)$. However, it is not an equivalence of categories.[34] More specifically, it is not *full*: that is, there are objects M, M' of $\mathrm{Mod}(T_H)$ such that the induced map $h \in \mathrm{Hom}(M, M') \mapsto \mathfrak{F}^*h \in \mathrm{Hom}(\mathfrak{F}^*M, \mathfrak{F}^*M')$ is not surjective.

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

Proof. See Appendix A. □

Second, consider the relationship between $\mathrm{Mod}(T_C)$ and $\mathrm{mod}(T_H)$—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 $\mathrm{mod}(T_H) \to \mathrm{Mod}(T_C)$;

oretical 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, 2015], the two notions coincide for theories with disjoint vocabulary.)

[33] [Weatherall, 2015a]

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

just to maintain notational hygiene, call this functor $\mathfrak{F}^\dagger$. Explicitly, for any $m \in \mathrm{mod}(T_H)$, let $\mathfrak{F}^\dagger m$ be the Σ_C-picture such that

- $|\mathfrak{F}^\dagger m| = |m|$
- For any $a, b \in |\mathfrak{F}^\dagger m|$, $\langle a, b\rangle \in C^{Fm}$ iff a and b are members of the same element of $\mathbf{2}^m$

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 $\mathfrak{F}^\dagger m \in \mathrm{Mod}(T_C)$, 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.

Proof. See Appendix A. □

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

$$\begin{aligned} |\mathfrak{I}^* M| &= |M| \\ \mathbf{2}^{\mathfrak{I}^* M} &= [L^M, R^M] \end{aligned} \tag{26}$$

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

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

Proof. See Appendix A. □

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),[35] we have at least shown that the obvious functors will not do the job. Let us now consider local field theories. As mentioned earlier, the relevant notion of morphism between functions $f : U \to V$ and $f' : U' \to V'$ (where

[35] Here is a means by which one could seek to demonstrate it: by showing that neither $\mathrm{Mod}(T_C)$ nor $\mathrm{mod}(T_H)$ have a terminal object. Since $\mathrm{Mod}(T_H)$ does have a terminal object (a/the model containing exactly one element that is L and exactly one that is R), doing so would suffice to show that the categories are inequivalent. This strikes me as a good proof-method, but one which I lack the categorical expertise to execute.

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 (i.e., 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 $\beta = \mathrm{Id}_{\mathbb{R}^q}$! That is, we find that the category of models of such a theory is always a *discrete* category.[36] For a local field theory equipped with some more sophisticated semantics (that is, done in a manner that doesn't take coordinates so seriously), one finds that the morphisms are somewhat more liberalised, and hence that the category of models is somewhat more interesting.[37]

Let's see how this plays out in the case of our electrostatic theories. First, consider the relationship between the (discrete) 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 (12); 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 $\mathrm{mod}(T_P) \to \mathrm{Mod}(T_E)$ which acts on models via the (dual of) (12), 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 (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 $\mathrm{mod}(T_P)$).

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.[38]

Proposition 8. $\mathfrak{G}^\dagger : \mathrm{mod}(T_P) \to \mathrm{Mod}(T_E)$ is full, faithful and surjective; i.e., it is an equivalence of categories.[39]

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

[37] The fact that there are so few morphisms between unsophisticated models is, of course, a product of our decision to work with coordinatised 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.)

[38] cf. [Weatherall, 2015b, Proposition 1].

[39] cf. [Weatherall, 2015b, Proposition 2].

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

All proofs are given in Appendix A.

We can do more or less the same thing for electromagnetism, establishing the same trinity of results.[40] 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 which enact such an equivalence: for instance, any functor between $\mathrm{Mod}(T_P)$ and $\mathrm{Mod}(T_E)$ which is bijective on objects will be an equivalence. However, I suspect that any functor which is describable in appropriately systematic terms (i.e. which meshes appropriately with respect to the non-categorical characterisation 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 which meshes appropriately with the architecture of the models (whatever exactly that gets made out to mean).[41] Then the following conjectures seem plausible:

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

Making these conjectures precise would require (a) a more thorough treatment of how to characterise reduction and sophistication in category-theoretic terms, and (b) 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.

[40] For an explicit discussion of the case of electromagnetism, see [Weatherall, 2015b] and [Weatherall, 2015a].

[41] At least in our examples, reasonableness seems to be a matter of being definable in terms of the *syntactic* content (e.g. 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], [Halvorson, 2013], [Lutz, 2014a], [Lutz, 2014b]).

Begin with the inequivalence between the reduced and unreduced theories (under vulgar semantics). Prima facie, this may seem in tension with Weatherall's 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."[42] 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 characterising 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 emphasised the need to specify (not just a theory, but also) the semantic structures one intends to use in formally interpreting the theory, Weatherall 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."[43] 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 judge that they do

[42] [Weatherall, 2015a, p. 15]
[43] [Weatherall, 2015a, p. 17]

so is precisely to affirm a commitment to sophisticated rather than vulgar semantics as embodying the true commitments of the theory.

However, what of the relationship between the reduced and sophisticated categories of models? In what sense are sophistication and reduction equivalent? In particular, one might be worried by the fact that I (apparently) introduced sophistication about theories as an *alternative* to reduction—and, I suggested, a superior one! So if there is indeed something to choose between them, surely they can't be equivalent after all?

Here is what seems to me like the right thing to say: the two theories are equivalent in terms of their *intensional ontology*, in terms of the kinds of structures that they postulate as present in any world aptly described by them; but they differ in their *explanatory* structure. Electrostatics in terms of sophisticated potentials, and electrostatics in terms of fields, both agree that there is a physically significant irrotational vector field; and both agree that this field (as with any such field) is representable as the gradient of a scalar field—provided that that scalar field is defined only up to potential shifts, or (equivalently) that it take values in Φ rather than $\mathbb{R}$. However, they disagree over what kind of explanation can be given of why this vector field is irrotational. For the theory in terms of fields, its irrotationality is simply a brute fact—a fact which usefully permits the field's representation as a certain kind of gradient, but not arising from anything else. For the theory in terms of sophisticated potentials, the field is the derivative object, and so admits of an explanation in terms of what is fundamental (i.e., the potential): it is irrotational *because* it is a gradient, and gradients always have vanishing curl.

As a result, whether the two theories are "really" equivalent will turn on what one wants to say about the role of explanation in theory equivalence. On some accounts,[44] two theories cannot be equivalent if they offer different explanations of the phenomena. This will be particularly true if one is inclined to view explanations of this sort as arising from some kind of ontological structure out there in the world, such as if one is committed to some notion of *grounding*—conceived of as a genuine part of the world's architecture, and responsible for answering in-virtue-of questions (e.g. "in virtue of what is the electric field irrotational?").[45] If, however, one is sceptical of grounding (and cognate notions), then there is space for some more quietist or deflationary attitude towards the relevant explanations. On this kind of view, there need not always

[44] e.g. [Putnam, 1983]

[45] See e.g. the essays in [Correia and Schnieder, 2012].

be some fact of the matter about what kind of explanatory architecture is correct. It is certainly illuminating to see that some feature in a theory *can* be explained by another, if the theory is set up a particular way; but (in general) there is no compulsion towards setting the theory up one way rather than another, or towards accepting one pattern of explanation amongst its parts as uniquely privileged.[46]

On either account, though, the case can be made for valuing sophistication over reduction. On some more realist account of explanation (e.g. 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 characterising 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 paper. 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.[47] 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 characterised as requiring us to choose between a trilemma of unpalatable ontologies: a locally acting[48] 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. But the argument here

[46] 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].

[47] In this regard, cf. [Pooley, 2013], [Weatherall, 2014b], and [Weatherall, 2015b].

[48] In the sense of having no "action at a distance".

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).[49] I don't claim that doing so will magically resolve these problems;[50] but it at least enlivens the conceptual geography.

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 characterise 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-fertilise our investigations into one area with insights from another.

References

[Aharonov and Bohm, 1959] Aharonov, Y. and Bohm, D. (1959). Significance of Electromagnetic Potentials in the Quantum Theory. *Physical Review*, 115(3):485–491.

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

[Baez, 1994] Baez, J. (1994). *Knots and Quantum Gravity*. Clarendon Press, Oxford.

[Baker, 2010] Baker, D. J. (2010). Symmetry and the Metaphysics of Physics. *Philosophy Compass*, 5(12):1157–1166.

[49] cf. footnote 31.

[50] 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 $U \cup V$).

[Barrett, 2014] Barrett, T. W. (2014). On the Structure of Classical Mechanics. *The British Journal for the Philosophy of Science*.

[Barrett and Halvorson, 2015] Barrett, T. W. and Halvorson, H. (2015). Glymour and Quine on Theoretical Equivalence. Unpublished draft.

[Blizard, 1988] Blizard, W. D. (1988). Multiset theory. *Notre Dame Journal of Formal Logic*, 30(1):36–66.

[Brading and Castellani, 2003] Brading, K. and Castellani, E., editors (2003). *Symmetries in physics : philosophical reflections*. Cambridge University Press, Cambridge.

[Cartwright, 1999] Cartwright, N. (1999). *The dappled world: a study of the boundaries of science*. Cambridge University Press, Cambridge, UK; New York, NY.

[Caulton, 2015] Caulton, A. (2015). The role of symmetry in the interpretation of physical theories. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52, Part B:153–162.

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

[Dasgupta, 2014] Dasgupta, S. (2014). Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science*, Forthcoming. References are to the preprint available at www.shamik.net.

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

[Glymour, 1970] Glymour, C. (1970). Theoretical Realism and Theoretical Equivalence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1970:275–288.

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

[Halvorson, 2013] Halvorson, H. (2013). The Semantic View, If Plausible, Is Syntactic. *Philosophy of Science*, 80(3):475–478.

[Hawthorne, 2001] Hawthorne, J. (2001). Causal Structuralism. *Noûs*, 35:361–378.

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

[Hodges, 1997] Hodges, W. (1997). *A shorter model theory*. Cambridge University Press, Cambridge; New York.

[Lewis, 2009] Lewis, D. (2009). Ramseyan Humility. In Braddon-Mitchell, D. and Nola, R., editors, *Conceptual Analysis and Philosophical Naturalism*, pages 203–222. Mit Press.

[Locke, 2012] Locke, D. (2012). Quidditism without quiddities. *Philosophical Studies*, 160(3):345–363.

[Loll, 1994] Loll, R. (1994). The Loop Formulation of Gauge Theory and Gravity. In *Knots and Quantum Gravity*, pages 1–20. Clarendon Press, Oxford.

[Lutz, 2014a] Lutz, S. (2014a). Empirical Adequacy in the Received View. *Philosophy of Science*, 81(5):1171–1183.

[Lutz, 2014b] Lutz, S. (2014b). What's Right with a Syntactic Approach to Theories and Models? *Erkenntnis*, pages 1–18.

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

[Maudlin, 2007] Maudlin, T. (2007). *The Metaphysics Within Physics*. Oxford University Press, Oxford.

[Pooley, 2006] Pooley, O. (2006). Points, particles, and structural realism. In Rickles, D., French, S., and Saatsi, J., editors, *The Structural Foundations of Quantum Gravity*, pages 83–120. Oxford University Press, Oxford.

[Pooley, 2013] Pooley, O. (2013). Substantivalist and relationalist approaches to spacetime. In *The Oxford Handbook of Philosophy of Physics*, pages 522–586. Oxford University Press, Oxford.

[Putnam, 1983] Putnam, H. (1983). Equivalence. In *Realism and Reason*, volume 3 of *Philosophical Papers*, pages 26–45. Cambridge University Press, Cambridge.

[Saunders, 2003a] Saunders, S. (2003a). Indiscernibles, general covariance, and other symmetries: the case for non-eliminativist relationalism. In Ashtekar, A., Howard, D., Renn, J., Sarkar, S., and Shimony, A., editors, *Revisiting the Foundations of Relativistic Physics: Festschrift in Honour of John Stachel*. Kluwer, Dordrecht.

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

[Swanson and Halvorson, 2012] Swanson, N. and Halvorson, H. (2012). On North's "The Structure of Physics". Unpublished note.

[Wallace, 2015] Wallace, D. (2015). Who's Afraid of Coordinate Systems?

[Weatherall, 2012] Weatherall, J. O. (2012). Inertial motion, explanation, and the foundations of classical spacetime theories. In Lehmkuhl, D., Schiemann, G., and Scholz, E., editors, *Towards a theory of spaceme theories*, number 13 in Einstein Studies. Birkhäuser, Basel. Draft of July 2012.

[Weatherall, 2014a] Weatherall, J. O. (2014a). Fibre Bundles, Yang-Mills, and GR. *The British Journal for the Philosophy of Science*.

[Weatherall, 2014b] Weatherall, J. O. (2014b). Regarding the "Hole Argument". Unpublished draft.

[Weatherall, 2015a] Weatherall, J. O. (2015a). Are Newtonian gravitation and geometrized Newtonian gravitation theoretically equivalent? *Erkenntnis*. Available from http://arxiv.org/abs/1411.5757.

[Weatherall, 2015b] Weatherall, J. O. (2015b). Understanding Gauge. Unpublished draft; paper given as part of a symposium on Formal Methods in Philosophy of Science at PSA 2014.

A. Proofs of propositions

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 iff $\mathfrak{D}^*$ is a bijection $\mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$, with $(\mathfrak{D}')^*$ as its inverse.

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 $|M|$,

$$M \models_v \mathfrak{D}\phi[\mathbf{a}] \iff \mathfrak{D}^*M \models_v \phi[\mathbf{a}] \tag{27}$$

Now, assume first that $\mathfrak{D}$ and $\mathfrak{D}'$ implement a translational equivalence between T_1 and T_2. I show that for any $M \in \mathrm{Mod}(T_2)$, $(\mathfrak{D}')^*\mathfrak{D}^*M = M$, i.e., 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,

$$\begin{aligned}(\mathfrak{D}')^*\mathfrak{D}^*M \models \pi[\mathbf{a}] \text{ iff } \mathfrak{D}^*M &\models \mathfrak{D}'\pi[\mathbf{a}] \\ \text{iff } M &\models \mathfrak{D}\mathfrak{D}'\pi[\mathbf{a}]\end{aligned}$$

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

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

and so $M \models \mathfrak{D}\mathfrak{D}'\pi[\mathbf{a}]$ iff $M \models \pi[\mathbf{a}]$. Thus, $\pi^{(\mathfrak{D}')^*\mathfrak{D}^*M} = \pi^M$. A similar proof goes to show that for any function-symbol $\mu \in \Sigma_2$, $\mu^{(\mathfrak{D}')^*\mathfrak{D}^*M} = \mu^M$. 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,

$$T_2 \vDash \forall\mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathfrak{D}\mathfrak{D}'\psi(\mathbf{x})) \tag{28}$$

So suppose that (28) did not hold. Then there would be some model M of T_2 such that $M \not\models \forall\mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathfrak{D}\mathfrak{D}'\psi(\mathbf{x}))$; i.e., such that for some $\mathbf{a}$ from M, either $M \models \psi[\mathbf{a}]$ and $M \not\models \mathfrak{D}\mathfrak{D}'\psi[\mathbf{a}]$, or vice versa. But by the above lemma, $M \models \psi[\mathbf{a}]$ iff $M \models \mathfrak{D}\mathfrak{D}'\psi[\mathbf{a}]$. So by reductio, (28) 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. □

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,

$$m \models_{V,v} \phi \text{ iff } m \models_{V',v} \phi \tag{23}$$

Proof. Clearly, there only are two second-order variable-assignments for m (since $\mathbf{2}^m$

has only two members); so if $V \neq V'$, then we have that $V(L) = V'(R)$ and $V(R) = V'(L)$. Let M and M' be Σ_H-pictures defined as follows:

$$|M| = |M'| = |m| \tag{29a}$$
$$L^M = V(L) \tag{29b}$$
$$R^M = V(R) \tag{29c}$$
$$L^{M'} = V'(L) = V(R) = R^M \tag{29d}$$
$$R^{M'} = V'(R) = V(L) = L^M \tag{29e}$$

In other words, $M' = \mathfrak{E}^* M$. But clearly, $m \models V, v\phi$ iff $M \models_v \phi$, and $m \models_{V',v} \phi$ iff $M' \models_v \phi$. Hence (suppressing reference to v), $m \models_V \phi$ iff $M \models \phi$ iff $M \models \mathfrak{E}\phi$ iff $\mathfrak{E}^* M \models \phi$ iff $M' \models \phi$ iff $m \models_{V'} \phi$. □

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

$$m \models_{V,v} \phi \text{ iff } m \models_{V',v} \phi \tag{24}$$

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 $\mathfrak{F}^* M = \mathfrak{F}^*(\mathfrak{E}^* M)$, we know that $\mathrm{Id}_{\mathfrak{F}^* M} \in \mathrm{Hom}(\mathfrak{F}^* M, \mathfrak{F}^*(\mathfrak{E}^* M))$. If there was some $h : M \to \mathfrak{E}^* M$ 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 $\mathfrak{E}^* M$ which does this. So there is no such h; thus, the induced map is not surjective. □

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 one element of $\mathbf{2}^m$ or the other (but not both). So let a^m denote the element of $\mathbf{2}^m$ of which a is a member.

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 $h_1 : |m| \to |n|$ by the condition that $h_1 = H$ (as a map on sets). Then, letting a be some arbitrary element of $|m|$, define $h_2 : \mathbf{2}^m \to \mathbf{2}^n$ 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 $i \in \mathbf{2}^m$ and any $b \in |m|$, if $b \in i$, then $b^m = i$, so $h^2(i) = (h_1(b))^n$, so $h_1(b) \in h_2(i)$. Thus, $h := (h_1, h_2)$ is a homomorphism $m \to n$, and $H = \mathfrak{F}^\dagger h$. So $\mathfrak{F}^\dagger$ induces a surjective map on morphisms between any m and n, i.e. $\mathfrak{F}^\dagger$ is full.

Now consider any $m, n \in \mathrm{mod}(T_H)$, and let $h, h' : m \to n$ 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, i.e., $\mathfrak{F}^\dagger$ is faithful.

Finally, let M by any model of T_H. Define a de-handed picture m by setting $|m| = |M|$, and letting the members of $\mathbf{2}^m$ be the two congruence classes of C^M. Clearly, $\mathfrak{F}^\dagger m = M$. So $\mathfrak{F}^\dagger$ is surjective, and therefore essentially surjective. □

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

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

□

Proposition 7. $\mathfrak{G}^* : \mathrm{Mod}(T_P) \to \mathrm{Mod}(T_E)$ is not full.[51]

Proof. Let $\mathfrak{C}^*$ be the functor on $\mathrm{Mod}(T_P)$ induced by the symmetry transformation (7). Let M be any model of T_P. Given the setup, 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 contains $\mathrm{Id}_{\mathfrak{G}^* M}$). 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. □

Proposition 8. $\mathfrak{G}^\dagger : \mathrm{mod}(T_P) \to \mathrm{Mod}(T_E)$ is full, faithful and surjective; i.e., it is an equivalence of categories.[52]

[51] cf. [Weatherall, 2015b, Proposition 1].
[52] cf. [Weatherall, 2015b, Proposition 2].

Proof. Consider any $m, n \in \mathrm{mod}(T_P)$, and let H by any morphism from $\mathfrak{G}^\dagger m$ to $\mathfrak{G}^\dagger n$. It must be the case that $H = \mathrm{Id}_{\mathfrak{G}^\dagger m}$ (since $\mathrm{Mod}(T_E)$ 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 $\mathfrak{G}^\dagger \mathrm{Id}_m = \mathrm{Id}_{\mathfrak{G}^\dagger m}$; if the latter, then $\mathfrak{G}^\dagger k = \mathrm{Id}_{\mathfrak{G}^\dagger m}$. 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 $m, n \in \mathrm{mod}(T_P)$, and any morphisms $h, h' : m \to n$. If $m = n$, then $\mathrm{Hom}(m, n) = \{\mathrm{Id}_m\}$; if $m \neq n$, then $\mathrm{Hom}(m, n) = \{k\}$ 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 $M \in \mathrm{Mod}(T_E)$. As already discussed, for any such model there is some $m \in \mathrm{mod}(T_P)$ such that $\mathfrak{G}^\dagger m = M$. So $\mathfrak{G}^\dagger$ is surjective, and hence essentially surjective. □

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

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