What Do Symmetries Tell Us About Structure?

Thomas William Barrett*

Abstract

Mathematicians, physicists, and philosophers of physics often look to the symmetries of an object for insight into the structure or constitution of the object. My aim in this paper is to explain why this practice is successful. In order to do so, I prove two theorems that are closely related to (and in a sense, generalizations of) Beth's and Svenonius' theorems.

1 Introduction

There is a famous idea about the relationship between the symmetries, or automorphisms, of a mathematical object and the structure of the object: An object's symmetries are often taken to provide us with significant information about its underlying structure. Hermann Weyl (1952, 144–5), for example, puts this idea as follows.

> A guiding principle in modern mathematics is this lesson: Whenever you have to do with a structure-endowed entity X, try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight into the constitution of X in this way.

Mathematicians, physicists, and philosophers of physics often employ Weyl's guiding principle. But justification for it is rarely offered,[1] and one is therefore left to wonder exactly *why* the automorphisms of X provide us with insight into the constitution of X. The aim of this paper is to answer this question.

We will begin by isolating a precise sense in which the automorphisms of an object encode significant information about the object. They provide a method of determining which structures are "definable" in terms of the basic structure of the object. Although there is a sense in which this method is imperfect, it

*I can be reached at thomaswbarrett@nyu.edu. Thanks to Neil Dewar, John Dougherty, Ben Feintzeig, Hans Halvorson, Phillip Kremer, Alex Meehan, and Jim Weatherall for comments and discussion on earlier versions of this paper.

[1] A notable exception is Dasgupta (2016). He understands symmetries in epistemic terms, rather than in "formal and mathematical" terms. He argues that the latter understanding of symmetries does not allow one to justify all of the inferences that are commonly made about symmetries. The discussion here can be understood as an exploration of how far we can get while still thinking about symmetries in the standard formal and mathematical terms.

suggests a more general way to learn about the structure of an object: Rather than only looking to automorphisms, one can use *all* of the structure-preserving maps between mathematical objects as a guide to the structure of the objects. We isolate a precise sense in which this more general method provides an even better guide to the structure of a mathematical object. One certainly can expect to gain a deep insight into the constitution of X by looking to X's automorphism group, but one can gain a deeper insight by looking to *all* of the structure-preserving maps between objects of the same type as X.

2 Symmetries and structure

Philosophers of physics often trace the standard method of reasoning about symmetries and structure back to the correspondence between Leibniz and Clarke on the nature of spacetime, and in particular, Leibniz's boost and shift arguments against Newtonian absolute space.[2] In their modern gloss, Leibniz's arguments aim to show that particular pieces of structure that Newton is committed to — namely, absolute position and absolute velocity — are not invariant under the symmetries of spacetime. Leibniz concludes from this that spacetime does not actually come equipped with those structure. Mathematicians, physicists, and philosophers of physics now reason about symmetries and structure in an analogous manner: After determining the symmetries of a particular mathematical object X, one will look for the structures on X that are "invariant under" or "preserved by" all of the symmetries of X. Those structures that are found to be invariant under the symmetries of X are often deemed to be "determined by" or "constructed from" or "come for free given" the basic structure of X. On the other hand, those structures that are found to be *not* invariant under the symmetries of X are not accorded this same status.

One can grasp the basic idea behind this method by considering the following examples. The first three are examples of structures that are invariant under the symmetries of the underlying mathematical object, and the latter three are examples of structures that are not.

Example 1. *The norm is invariant under the symmetries of an inner product space.* Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space, and consider the norm $||\cdot||$ on V associated with the inner product. Every automorphism f of $(V, \langle\cdot,\cdot\rangle)$ also preserves the norm, in the sense that $||v|| = ||fv||$ for all $v \in V$. ⌟

Example 2. *The metric topology is invariant under the symmetries of a metric space.* Let (X, d) be a metric space, and consider the metric topology τ_d on X. Every automorphism of (X, d) preserves the topology τ_d, in the sense that it is a homeomorphism. ⌟

Example 3. *The Levi-Civita derivative operator is invariant under the symmetries of a manifold with metric.* Let (M, g_{ab}) be a smooth manifold with metric,

[2] For discussion see Earman (1989), Baker (2010), Dasgupta (2015, 2016), and the references therein.

and consider the Levi-Civita derivative operator ∇ associated with g_{ab}. Every automorphism of (M, g_{ab}) preserves the derivative operator ∇, in the sense that $f^*(\nabla_n \lambda^{a_1 \ldots a_r}_{b_1 \ldots b_s}) = \nabla_n f^*(\lambda^{a_1 \ldots a_r}_{b_1 \ldots b_s})$ for all smooth tensor fields $\lambda^{a_1 \ldots a_r}_{b_1 \ldots b_s}$. $\lrcorner$

Example 4. *An order is not invariant under the symmetries of a set.* Let X be a set containing more than one element, and consider an arbitrary linear order $<$ on X. There is an automorphism of X (i.e. a bijection $X \to X$) that does not preserve $<$. $\lrcorner$

Example 5. *An inner product is not invariant under the symmetries of a vector space.* Let V be a vector space, and consider an arbitrary inner product $\langle \cdot, \cdot \rangle$ on V. One can easily show that there is an automorphism of V that does not preserve the inner product. $\lrcorner$

Example 6. *The Galilean temporal metric is not invariant under the symmetries of Minkowski spacetime.* Let $(\mathbb{R}^4, \eta_{ab})$ be Minkowski spacetime, and consider the standard temporal metric $t_{ab} = (d_a x^1)(d_b x^1)$ of Galilean spacetime. There are automorphisms of $(\mathbb{R}^4, \eta_{ab})$ that do not preserve t_{ab} (Barrett, 2015b, Proposition 2). $\lrcorner$

There is a stark contrast between the first three examples and the latter three. The norm $|| \cdot ||$, the metric topology τ_d, and the Levi-Civita derivative operator ∇ are all determined by the basic structure of their respective mathematical objects. In fact, in each of the first three examples the basic structure of the mathematical object suffices to *define* the piece of invariant structure. In Example 1, for instance, one uses the inner product in the familiar way to define the norm $||v|| := \sqrt{\langle v, v \rangle}$ of a vector $v \in V$. The same holds of the metric topology and the Levi-Civita derivative operator. They are definable in terms of the metric d and the metric g_{ab}, respectively.

The structures on a mathematical object that are not invariant under the symmetries of the object, on the other hand, are not "determined by" the basic structure of the underlying mathematical object. In contrast to the invariant structures from Examples 1–3, the basic structure of the mathematical objects in Examples 4–6 does not suffice to define the new piece of structure. A set does not define a privileged ordering of its elements, the basic structure of a vector space does not suffice to define a privileged inner product, and the Minkowski metric famously does not define a notion of absolute simultaneity on spacetime.

Examples like the above six suggest the following "conjecture" about the relationship between symmetry and structure:

Conjecture. *A piece of structure is invariant under the symmetries of a mathematical object if and only if it is definable from the basic structure of the object.*

If true, this conjecture would explain why Weyl's guiding principle is successful. It is natural to think of mathematical objects as coming equipped not only with their "basic structure," but also with the structures that are definable in terms of their basic structure. If the symmetries of an object tell us which structures on the object are definable in terms of the object's basic structure, then they provide us with a guide to the structures that the object actually comes equipped with. Symmetries would then provide us with insight into a mathematical object's constitution because they tell us precisely which structures the object has and which the object lacks.

3 Two theorems

In order to consider whether this conjecture is true, we first need to clarify it. We do so by working in the framework of standard first-order logic, and in particular, the theory of definability. We will need the following basic preliminaries.[3]

A **signature** Σ is a set of predicate symbols, function symbols, and constant symbols. The Σ-terms, Σ-formulas, and Σ-sentences are recursively defined in the standard way. A **Σ-structure** A is a nonempty set in which the symbols of Σ have been interpreted. One recursively defines when a sequence of elements $a_1, \ldots, a_n \in A$ **satisfy** a Σ-formula $\phi(x_1, \ldots, x_n)$ in a Σ-structure A, written $A \vDash \phi[a_1, \ldots, a_n]$. We will use the notation ϕ^A to denote the set of tuples from the Σ-structure A that satisfy a Σ-formula ϕ. A **Σ-sentence** is a Σ-formula with no free variables. So if ϕ is a Σ-sentence, then $A \vDash \phi$ just in case the empty sequence satisfies ϕ in A. A **Σ-theory** T is a set of Σ-sentences. The sentences $\phi \in T$ are called the **axioms** of T. A Σ-structure M is a **model** of a Σ-theory T if $M \vDash \phi$ for all $\phi \in T$. Two Σ-theories are **logically equivalent** if they have the same class of models. A theory T **entails** a sentence ϕ, written $T \vDash \phi$, if $M \vDash \phi$ for every model M of T. If $\Sigma \subset \Sigma^+$ are signatures, we say that a Σ^+-theory T^+ is an **extension** of a Σ-theory T if $T \vDash \phi$ implies that $T^+ \vDash \phi$ for every Σ-sentence ϕ.

We can now clarify what it might mean for the basic structure of a mathematical object to "define" an additional piece of structure. The standard way to do this employs the following set-up.

- Let Σ be a signature. We think of the elements of Σ as the pieces of "basic structure" of the mathematical objects under consideration.
- Let r be a symbol that is not contained in Σ. We think of r as the additional piece of structure that we are investigating. It may or may not be invariant under the symmetries of the mathematical object. We will assume for simplicity that r is a unary predicate symbol, but everything that follows easily generalizes to the cases where r is not unary and where r is a function or constant symbol.
- Let T be a $\Sigma \cup \{r\}$-theory. We think of the theory T as picking out the "type of mathematical object" that we will be considering.

It is worth taking a moment here to state how this set-up relates to some of the above examples. In Example 3 "the language of manifolds with metric" plays the role of Σ, the derivative operator ∇ plays the role of r, and "the theory of

[3] The reader is encouraged to consult Hodges (2008) for further details.

manifolds with metric and Levi-Civita derivative operator" plays the role of T. Similarly, in Example 5 "the language of vector spaces" plays the role of Σ, the inner product $\langle \cdot, \cdot \rangle$ plays the role of r, and "the theory of inner product spaces" plays the role of T.

There are two particularly natural ways of making precise the idea that the theory T defines the structure r in terms of the basic structures in Σ. The first condition that we will consider is the following.

(E1) There is a Σ-formula ϕ such that $T \vDash \forall x(r(x) \leftrightarrow \phi(x))$.

When E1 holds, we say that the theory T **explicitly defines** r in terms of Σ. We call the sentence $\forall x(r(x) \leftrightarrow \phi(x))$ an **explicit definition** of r in terms of Σ. The condition E1 captures a sense in which r can be "constructed from" the basic structures in Σ.

The following condition provides us with a second natural way to explicate the idea that T defines r in terms of the basic structures in Σ.

(I1) For all models M and N of T, if $M|_\Sigma = N|_\Sigma$, then $r^M = r^N$.

Here $M|_\Sigma$ and $N|_\Sigma$ are the Σ-structures obtained from M and N by "forgetting" the extension of the predicate r. When I1 holds, we say that T **implicitly defines** r in terms of Σ. Like E1, this captures a sense in which r is "determined by" the basic structures in Σ. The condition simply says that whenever two models agree on the structures in Σ, they must also agree on the structure r.

These basic intuitions behind explicit and implicit definition are confirmed by the following example.

Example 7. Let $\Sigma = \{p, q\}$ be a signature containing two unary predicate symbols, and consider the Σ-theory T with the following two axioms.

$$\forall x(p(x) \vee q(x)) \qquad \forall x \neg(p(x) \wedge q(x))$$

This theory says that there are two types of things — the p's and the q's — and everything is of one of these types, but not both. One verifies that $T \vDash \forall x(p(x) \leftrightarrow \neg q(x))$, which means that T explicitly defines p in terms of q. The predicate p can intuitively be "constructed" by taking the negation of the predicate q. One also shows that any two models M and N of T with $q^M = q^N$ must also satisfy $p^M = p^N$, which means that T implicitly defines p in terms of q. The extension of the predicate q "determines" the predicate p. (One can easily see that T also explicitly and implicitly defines q in terms of p.) ⌟

This example suggests that explicit definability and implicit definability are intimately related to one another. The following famous result establishes that this is indeed the case.

Beth's theorem. *E1 if and only if I1.*

Proof. See Hodges (2008, Theorem 6.6.4). □

With Beth's theorem in hand, we have the resources to address our conjecture from above. We would like to know what the relationship is between the definability conditions E1 and I1 and the invariance (or lack thereof) of the structure r under symmetries. The following condition is a natural way to make precise the idea that r is invariant under symmetries of the basic structures in Σ.

(S1) For any model M of T, if $h : M|_\Sigma \to M|_\Sigma$ is an automorphism, then $h[r^M] = r^M$.

An **automorphism** of a Σ-structure A is a bijection from A to itself that preserves the extensions of all of the predicates, functions, and constants in Σ. Automorphisms of a Σ-structure N are the maps from N to itself that preserve all of the basic structures in Σ. The condition S1 is therefore a straightforward way of saying that the symmetries of the basic structures of M preserve the structure r too.

We have the following simple result about the relationship between the conditions E1 and S1.

Proposition. *If E1, then S1.*

Proof. Let ϕ be the Σ-formula (whose existence is guaranteed by E1) that explicitly defines r, and let $h : M|_\Sigma \to M|_\Sigma$ be an automorphism. Since automorphisms preserve the extensions of all Σ-formulas, $h[\phi^M] = \phi^M$. E1 guarantees that $\phi^M = r^M$, which immediately implies S1. □

In conjunction with Beth's theorem, this result shows that I1 also implies S1. The most natural way to use this proposition is by appealing to its contrapositive. The contrapositive provides us with a simple way of showing that a piece of structure is *not* definable: If we can show that a piece of structure is not invariant under the symmetries of a mathematical object, then this proposition licenses us to conclude that the structure is not definable (neither explicitly nor implicitly) from the basic structures of the mathematical object.[4]

One can see this method in action by looking back to Examples 4–6. In Example 5, for instance, one shows that there is an automorphism of a vector space V that does not preserve the inner product. The contrapositive of this proposition (extrapolating beyond first-order logic) licenses one to conclude that an inner product is *not* defined by the basic structure of a vector space. There is no natural inner product that is "determined by" or "comes for free given" the basic structure of a vector space.

The proposition gives us the "if" half of our conjecture. If a piece of structure is definable from the basic structure of a mathematical object, then it is invariant under the symmetries of the object. But the proposition leaves open the "only if" half of the conjecture. We still do not know the extent to which we are

[4] In this respect the proposition is closely related to the "only if" half of Beth's theorem, which is sometimes called "Padoa's method."

justified in concluding that the pieces of invariant structure in Examples 1–3 *are* definable in terms of the basic structure of the underlying objects.

Unfortunately, the following example shows that this inference is not yet justified: *S1 does not imply E1.*

Example 8. Consider the signature $\Sigma = \{p\}$, where p is a unary predicate symbol, and let T be the $\Sigma \cup \{r\}$-theory with the one axiom $\exists_{=1}x(x = x)$. This theory says that there is one thing, but says nothing about whether it is p or r. The condition S1 holds of the theory T. Indeed, if M is a model of T and $h : M|_{\{p\}} \to M|_{\{p\}}$ is an automorphism, then it must be that h is the identity map, since M only contains one element. This immediately implies that $h[r^M] = r^M$. But I1 does not hold of T since there are models M and N with $p^M = p^N$ but $r^M \neq r^N$. Beth's theorem implies that E1 also fails to hold of T. ⌟

This example demonstrates a sense in which the "only if" half of the conjecture fails. Indeed, it seems that the symmetries of a mathematical object do not provide us with a complete guide to the definable structures of the object. A piece of structure's invariance under the automorphisms of an object does not necessarily imply that it is definable from the object's basic structure.

There is, however, a way to substantiate the conjecture: We can be more restrictive about what kinds of mathematical objects we are considering. We say that a Σ-theory T is **complete** if for every Σ-sentence ϕ, either $T \vDash \phi$ or $T \vDash \neg\phi$. When one restricts attention to complete theories, the converse of the above proposition holds.

Svenonius' theorem. *If T is complete, then E1 if and only if S1.*

Proof. See Hodges (2008, Corollary 10.5.2). □

Svenonius' theorem shows that our conjecture holds for complete theories. For this restricted class of mathematical objects — that is, models of complete theories — it is the case that a piece of structure is definable if and only if it is invariant under symmetry. But this way of substantiating the conjecture leaves something to be desired. Completeness is a strong condition to impose on a theory. Most first-order theories are not complete, and one wonders what the relationship is between symmetry and structure for these more general theories.

Fortunately, there is another way to try to overcome the fact that S1 does not imply E1. So far, when asking what symmetries tell us about structure, we have only allowed ourselves to consider the automorphisms of a mathematical object. But automorphisms are just one particular kind of structure-preserving map between mathematical objects — namely, the ones from an object to itself. This observation suggests a more general way to learn about the definable structures on a mathematical object X: Rather than only looking to the automorphisms of X, one can look to *all* of the structure-preserving maps between objects of the same kind as X. It is natural to wonder how much information this larger class of maps encodes about the definable structures on X.

The following two generalizations of the condition S1 are in line with this new approach.

(S2) For all models M and N of T, if $h : M|_\Sigma \to N|_\Sigma$ is an elementary embedding, then $h[r^M] = r^N$.

(S3) For all models M and N of T, if $h : M|_\Sigma \to N|_\Sigma$ is an isomorphism, then $h[r^M] = r^N$.

Two clarifications are in order about these conditions. First, an **elementary embedding** between Σ-structures A and B is a map $h : A \to B$ that satisfies

$$A \vDash \phi[a_1, \ldots, a_n] \text{ if and only if } B \vDash \phi[h(a_1), \ldots, h(a_n)]$$

for all Σ-formulas $\phi(x_1, \ldots, x_n)$ and elements $a_1, \ldots, a_n \in A$. And second, an **isomorphism** $h : A \to B$ between the Σ-structures A and B is a bijection that preserves the extensions of all predicates, functions, and constant symbols in Σ. Every automorphism is an isomorphism, and every isomorphism is an elementary embedding, but in general the converses do not hold.

The conditions S2 and S3 differ from S1 only in that they appeal to elementary embeddings and isomorphisms instead of automorphisms. These conditions provide two straightforward ways of saying that maps between models of T that preserve their basic structure also preserve the structure r. It turns out that S2 and S3 are both equivalent to E1 and I1. By themselves, the automorphisms of a mathematical object did not provide us with all of the information about the definable structure on that object. But once we allow ourselves to look at these larger classes of maps — that is, elementary embeddings or isomorphisms — we are provided with all of the information about definable structure.

Theorem 1. *E1 if and only if S2.*

Proof. Suppose first that E1 holds. Let M and N be models of T with $h : M|_\Sigma \to N|_\Sigma$ an elementary embedding. We immediately see that

$$h[r^M] = h[\phi^M] = \phi^N = r^N$$

where ϕ is the Σ-formula (whose existence is guaranteed by E1) that explicitly defines r. The first and third equalities follow from E1, while the second equality holds since h is an elementary embedding. This implies S2.

Now suppose that S2 holds. Let M and N be models of T with $M|_\Sigma = N|_\Sigma$. The identity map $1 : M|_\Sigma \to N|_\Sigma$ is an elementary embedding, so by S2 it must be that $1[r^M] = r^N$. This immediately implies that $r^M = r^N$ and so $M = N$. We have therefore shown I1. Beth's theorem then implies E1. □

Theorem 2. *E1 if and only if S3.*

Proof. Suppose that E1 holds. Theorem 1 implies that S2 must hold, and since every isomorphism is an elementary embedding we immediately see that S3 holds too. On the other hand, if S3 holds, one establishes E1 by arguing exactly as in the "if" half of Theorem 1. □

4 Philosophical payoffs

The relationships between these different notions of definability and invariance under symmetry are summarized in the following figure.

$$\begin{tikzcd}
E1 \arrow[r, shift left] & I1 \arrow[l, shift left] \arrow[r, shift left] \arrow[d, shift right] & S2 \arrow[l, shift left] \arrow[r, shift left] & S3 \arrow[l, shift left] \\
& S1 \arrow[u, shift right, "/" marking] & &
\end{tikzcd}$$

Beth's theorem establishes the left-most equivalence in the top row, while the other two follow from Theorems 1 and 2. The relationship between S1 and the conditions in the top row follows from our proposition and Example 8.

These results come to bear on our earlier conjecture (which we restate here for convenience) in a straightforward manner.

Conjecture. *A piece of structure is invariant under the symmetries of a mathematical object if and only if it is definable from the basic structure of the object.*

The fact that S1 does not entail the conditions E1 and I1 shows that automorphisms by themselves do not encode all of the information about definable structure. So there is a sense in which the "only if" half of the conjecture does not hold. But the equivalence of S2 and S3 with E1 and I1 does establish a slightly weaker form of the conjecture. These equivalences show that the class of all structure-preserving maps between mathematical objects encodes all of the information about which structures are and are not definable on the objects. These results therefore suggest an amendment to Weyl's guiding principle about symmetry and structure: *One can gain insight into the constitution of a mathematical object by looking to the class of structure-preserving maps between objects of the same kind.*

In addition to allowing us to improve upon Weyl's guiding principle, these results help to justify a number of arguments in philosophy of physics and the foundations of mathematics. It is common practice in philosophy of physics to use the symmetries of a particular physical theory as a means of examining the structure of the theory. For example, symmetries are often used in debates between substantivalists and relationalists about the structure of spacetime. Relationalists will often argue that a particular piece of structure — like "absolute position" or "absolute velocity" — is not invariant under the symmetries of spacetime. This type of argument can be understood as an appeal to the "if" half of our conjecture. If one can show that a piece of structure is not invariant under the symmetries of spacetime, then one is licensed to conclude that the structure is not definable in terms of the basic structure of spacetime. This in turn provides a strong sense in which spacetime simply does not come equipped with that structure.[5]

[5] For discussion of spacetime symmetries, see Earman (1989), Dasgupta (2015, 2016), and the references therein. Weatherall (2017b) contains an argument about when the "extra

The "only if" half of the conjecture is sometimes employed more explicitly. One particularly famous example appears in the literature on the conventionality of simultaneity in special relativity. Philosophers of physics believed for many years that special relativity did not come equipped with a privileged notion of observer-relative simultaneity; the standard special relativistic notion of simultaneity was instead thought to be merely a convention. Malament (1977) was able to show, however, that the standard simultaneity relation on Minkowski spacetime is the only non-trivial equivalence relation that is invariant under the symmetries of special relativity. Malament explicitly appeals to the "only if" half of our conjecture to explain why his result is so powerful: It implies that the standard simultaneity relation is the only non-trivial equivalence relation that is definable in terms of the basic structure of Minkowski spacetime. In other words, Minkowski spacetime does not come equipped with any other candidate for a simultaneity relation.

The conjecture also comes to bear on two broader issues in philosophy of physics: the question of how to compare structure between theories and the question of how to assess whether or not two theories are equivalent. We conclude by discussing these two topics in turn.

Structure

The history of classical spacetime theories is often viewed as a progression towards a "less structured" spacetime. Aristotelian spacetime posits more structure than Newtonian spacetime, which in turn posits more structure than Galilean spacetime.[6] Intuitively, each of these spacetimes is obtained by taking something away from its predecessor. Galilean spacetime, for example, is obtained by taking away the preferred rest frame from Newtonian spacetime.

In order to capture the relationship that these different spacetime theories bear to one another, one needs a precise method of comparing "amounts of structure." Such a method would also be useful when diagnosing whether the models of a particular physical theory have "surplus structure" or when a theory is a "gauge theory."[7] It has recently been suggested that symmetries can be used to compare amounts of structure between mathematical objects. The basic idea behind this suggestion is that since the automorphisms of an object are the invertible structure-preserving maps from the object to itself, an object with "more automorphisms" must have "less structure" that these automorphisms are required to preserve.[8] The amount of structure that an object has is (in some sense) inversely proportional to the size of the object's automorphism

facts" that the substantivalist demands are definable in a particular mathematical structure, connecting the issues discussed in this paper to some of the classic works on substantivalism, relationalism, and symmetry. For general discussion of symmetries in philosophy of physics, see Belot (2003, 2013), Brading and Castellani (2007), Baker (2010), and Dewar (2015).

[6] See Maudlin (2012), Geroch (1978), and Barrett (2015b) for discussion.

[7] Weatherall (2016b) discusses the relationship between these two notions.

[8] For discussion of symmetries, automorphisms, and amounts of structure see Earman (1989), Ismael and van Fraassen (2003), North (2009), Halvorson (2011), Swanson and Halvorson (2012), Curiel (2014), Barrett (2015b,a), Weatherall (2016b), and Dewar (2016).

group. The following criterion has been proposed by Swanson and Halvorson (2012) and Barrett (2015a,b) to make this idea precise.

SYM*: X has more structure than Y if the automorphism group of X is a proper subset of the automorphism group of Y.

Our results about definability and invariance under symmetry lend support to this kind of criterion. If an object has "more automorphisms," then it is more difficult for a new piece of structure to be invariant under these automorphisms. The size of an object's automorphism group therefore provides us with a guide to the amount of definable structure that the object has. And indeed, SYM* makes the intuitive verdicts when presented with many classic examples. A topological space has more structure than a bare set, an inner product space has more structure than a bare vector space, and a manifold with metric has more structure than a bare manifold. In addition, SYM* makes the correct verdicts when applied to classical spacetime theories (Barrett, 2015b).

The problem with SYM*, however, stems from the fact that S1 does not entail E1. The automorphisms of an object do not provide a complete guide to definable structures on that object. This worry can be made precise by considering again the situation from Example 8. One can easily verify that for every model M of T, it is *not* the case that the structure $M|_{\{p\}}$ has less structure than M according to SYM*. Indeed, the two objects have precisely the same automorphism group. This is an undesirable verdict. The object M comes equipped with the structure provided by predicate r, and this is structure that the object $M|_{\{p\}}$ does not have. The fact that E1 does not hold of T shows that r is not even definable in terms of the structure on $M|_{\{p\}}$. Intuitively, M therefore has more structure than $M|_{\{p\}}$. The criterion SYM* makes the wrong verdict in this case.

The results above suggest that we can obtain a better guide to the amount of structure that an object has by looking to the class of *all* structure-preserving maps between objects rather than merely the automorphisms. And in fact, a method of comparing amounts of structure that employs exactly this idea has already been proposed. Baez et al. (2006) have suggested that one can compare amounts of structure between mathematical objects by looking to the *categories* in which the objects reside.[9]

In order to explain this method of comparing amounts of structure, we need the following simple category-theoretic machinery.[10] A first-order theory T has a category of models. A **category** C is a collection of objects with arrows between the objects that satisfy some basic properties. We will use the notation $\mathrm{Mod}(T)$ to denote the **category of models** of T. An object in $\mathrm{Mod}(T)$ is a model M of T, and an arrow $f : M \to N$ between objects in $\mathrm{Mod}(T)$ is an elementary embedding $f : M \to N$ between the models M and N. A functor $F : C \to D$ between categories C and D is a structure-preserving map between

[9] See also Barrett and Halvorson (2013) and Weatherall (2016b).

[10] The reader is encouraged to consult Mac Lane (1971) or Borceux (1994) for further details. We take for granted the definitions of a category and of a functor.

categories. When T^+ is an extension of a Σ-theory T, we can define the functor $\Pi : \mathrm{Mod}(T^+) \to \mathrm{Mod}(T)$ by

$$\Pi(M) = M|_\Sigma \qquad \Pi(h) = h$$

for every model M of T^+ and elementary embedding h between models of T^+. One can easily verify that Π is a functor. We say that a functor $F : C \to D$ is **full** if for all objects c_1, c_2 in C and arrows $g : Fc_1 \to Fc_2$ in D there exists an arrow $f : c_1 \to c_2$ in C with $Ff = g$. F is **faithful** if for all objects c_1, c_2 in C and arrows $f, g : c_1 \to c_2$, $Ff = Fg$ implies that $f = g$. And F is **essentially surjective** if for every object d in D there is an object c in C such that Fc is isomorphic to d. A functor that is full, faithful, and essentially surjective is called an **equivalence** of categories.

Baez et al. (2006) classify functors between categories based on "what they forget." Most importantly for our purposes, when a functor $F : C \to D$ is not full it is said to **forget structure**. The existence of a functor $F : C \to D$ that forgets structure captures a sense in which (relative to the comparison generated by F) objects of D have less structure than objects of C. One can see the idea behind this method by considering the following example. It is standard to recognize a sense in which topological spaces have more structure than sets, and the Baez method of comparing amounts of structure allows one to recover this sense.

Example 9. Consider the categories Set and Top. The objects of Set are sets and the arrows are functions between sets. The objects of Top are topological spaces and the arrows are continuous functions. One particularly natural functor $U : \mathrm{Top} \to \mathrm{Set}$ is defined by

$$U : (X, \tau) \longmapsto X \qquad U : f \longmapsto f$$

for all topological spaces (X, τ) and continuous functions f. One can easily verify that U is a functor. It converts a topological space into a set by "forgetting" about the topology. Since there are functions between some topological spaces that are not continuous, U trivially is not full and therefore forgets structure. ⌟

The motivation behind the Baez method of comparing amounts of structure is the same as that behind SYM*. Since the functor $U : \mathrm{Top} \to \mathrm{Set}$ is not full, this provides a sense in which there are "more arrows" (relative to the comparison given by U) between objects in the category Set than there are between objects in the category Top. The arrows in these categories are structure-preserving maps between the objects. Therefore, since there are "more structure-preserving maps" between the objects of Set than there are between the objects of Top, the former must have less structure that these maps are required to preserve.

Theorem 1 yields a corollary that concretely justifies the Baez method of comparing amounts of structure.

Corollary 1. *Let T^+ be a $\Sigma \cup \{r\}$-theory that is an extension of the Σ-theory T. The functor $\Pi : Mod(T^+) \to Mod(T)$ forgets structure if and only if E1 does not hold of T^+.*

Proof. It is easy to verify that Π is full if and only if S2 holds of T^+. Theorem 1 then immediately implies the corollary. □

Extrapolating beyond the case of first-order theories, this corollary tells us that a functor from C to D forgets structure if and only if the objects in C have structure that is not definable from the structure of the objects in D. Note that this feature of the Baez method is an improvement upon the criterion SYM*. As we saw when we revisited Example 8, it can be the case that objects of C and D have the same amount of structure according to SYM*, even though the objects of C come equipped with some structure that is *not* definable in terms of the structure that objects of D have. The automorphism group of a mathematical object therefore does not provide us with a perfect guide to the amount of definable structure than an object has. Corollary 1, however, shows that the *category* in which the object resides does provide us with such a guide.

Equivalence

These results also come to bear on a particular approach to theoretical equivalence, a topic which has recently received significant attention from philosophers of science.[11] We would like to know the conditions under which two theories should be considered equivalent. It has recently been suggested that category theory provides us with a standard for equivalence of theories: Two theories T_1 and T_2 are **categorically equivalent** if their categories of models $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are "structurally identical", i.e. if there is a functor F between them that is an equivalence of categories. This criterion is supposed to capture a sense in which the two theories might be considered "intertranslatable."

One would therefore hope that categorically equivalent theories T_1 and T_2 are such that the structures of T_1 are definable in terms of T_2 and vice versa. Unfortunately, this is not the case. Barrett and Halvorson (2016b, Theorem 5.2) provide an example of theories T_1 and T_2 that are categorically equivalent, but T_1 is unable to define the structures of T_2 and vice versa. This demonstrates the following:

> It is not the case that an equivalence of categories between $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ implies that the structures of T_1 are definable in terms of the structures of T_2.

This result is a definite mark against categorical equivalence as a general standard for equivalence of theories. Our discussion here, however, does suggest that one might be able to strengthen categorical equivalence in a way that does

[11] For example, see Quine (1975), Sklar (1982), Halvorson (2012, 2013, 2016), Glymour (2013), Van Fraassen (2014), and Coffey (2014) for general discussion of theoretical equivalence in philosophy of science. Glymour (1977), Knox (2014), Weatherall (2016a, 2017a), North (2009), Swanson and Halvorson (2012), Curiel (2014), Barrett (2015a, 2016), Rosenstock et al. (2015), Rosenstock and Weatherall (2016), Rosenstock (2015), Hudetz (2015) discuss particular physical theories. For various logical results proven about criteria for equivalence see Barrett and Halvorson (2016b) and the references therein.

allows it to better capture facts about definability. In particular, we have another simple corollary. In order to state this corollary we need one definition. Let $\Sigma \subset \Sigma^+$ be signatures. A **definitional extension** of a Σ-theory S to the signature Σ^+ is a Σ^+-theory that is logically equivalent to the theory

$$S^+ = S \cup \{\delta_s : s \in \Sigma^+ - \Sigma\},$$

where for each symbol $s \in \Sigma^+ - \Sigma$, the sentence δ_s is an explicit definition of s in terms of Σ. Two theories are **definitionally equivalent** if they have a common definitional extension.[12]

Corollary 2. *Let T^+ be a $\Sigma \cup \{r\}$-theory that is an extension of the Σ-theory T. The functor $\Pi : Mod(T^+) \rightarrow Mod(T)$ is an equivalence if and only if T^+ is a definitional extension of T.*[13]

Proof. The proof of the "if" half is familiar; it follows from Theorem 5.1 of Barrett and Halvorson (2016b). Assume then that Π is an equivalence. Since Π is full, Corollary 1 implies that E1 holds of T^+, so

$$T^+ \vDash \forall x(\phi(x) \leftrightarrow r(x))$$

for some Σ-formula ϕ. Now using the fact that Π is essentially surjective, one easily verifies that $T \cup \{\forall x(\phi(x) \leftrightarrow r(x))\}$ is logically equivalent to T^+. $\square$

The existence of an arbitrary equivalence between the categories of models of two theories does not guarantee that the two theories can define one another's structures. But if Π is an equivalence, then Corollary 2 implies that the two theories can do precisely this. It is natural to wonder whether there is some special property $\mathfrak{P}$ of the functor Π that allows it to encode more about definable structure than an arbitrary functor does. This suggests a family of conjectures of the following form:

> If there is a functor F that (i) is an equivalence of categories between $\text{Mod}(T_1)$ and $\text{Mod}(T_2)$ and (ii) has property $\mathfrak{P}$, then T_1 and T_2 are definitionally equivalent.[14]

[12] Definitional extensions and definitional equivalence have received attention in logic and philosophy of science. For example, see de Bouvére (1965), Kanger (1968), Glymour (1971, 1977, 1980, 2013), Pinter (1978), Pelletier and Urquhart (2003), Andréka et al. (2005), Friedman and Visser (2014), and Barrett and Halvorson (2016a,b, 2017a,b), and the references therein.

[13] It is important to mention a subtlety about the statement of this corollary. Note that since we are working in the framework of single-sorted logic, the functor Π is always guaranteed to be faithful. In the many-sorted framework, Π can fail to be faithful, and therefore, this corollary fails in this more general framework. I conjecture that the following result holds universally: Π *is an equivalence if and only if T^+ is a Morita extension of T*, in the sense defined by Barrett and Halvorson (2016b).

[14] As in footnote 13, the more general conjecture replaces definitional equivalence with Morita equivalence.

Results of this form would take a significant step towards improving upon categorical equivalence as a general standard of equivalence between theories. And indeed, work in this direction is currently being done by Hudetz (2016).[15]

The results here take a step towards providing justification for the use of category theoretic tools when examining the relationships between theories. The tools seems to be particularly well-suited to capture (i) when models of one theory have less structure than models of another theory, and (ii) when two theories are equivalent. More generally, these results provide support for one of the primary motivations behind category theory. The idea at the heart of category theory is simple: Mathematical objects can be thought of, not in terms of their "internal structure," but rather in terms of the relations that they bear to other objects. For example, from the category theoretic perspective one sees a group not as a set with a binary operation, but instead as an object in a particular network of arrows. This viewpoint has proven useful over the course of the last sixty years, yielding applications in many branches of mathematics and computer science. The extent to which the perspective is justified, however, depends on precisely how much information about mathematical objects is encoded by the arrows — that is, the structure-preserving maps — between the objects. Theorems 1 and 2 take a step towards justifying this perspective: The structure-preserving maps between mathematical objects encode *all* of the information about which structures are and are not definable on the objects.

References

Andréka, H., Madarász, J. X., and Németi, I. (2005). Mutual definability does not imply definitional equivalence, a simple example. *Mathematical Logic Quarterly*, 51(6):591–597.

Awodey, S. and Forssell, H. (2010). First-order logical duality. *Manuscript.*

Baez, J., Bartels, T., Dolan, J., and Corfield, D. (2006). Property, structure and stuff. Available at http://math.ucr.edu/home/baez/qg-spring2004/discussion.html.

Baker, D. (2010). Symmetry and the metaphysics of physics. *Philosophy Compass.*

Barrett, T. W. (2015a). On the structure of classical mechanics. *The British Journal for the Philosophy of Science*, 66(4):801–828.

Barrett, T. W. (2015b). Spacetime structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 51:37–43.

Barrett, T. W. (2016). Equivalent and inequivalent formulations of classical mechanics. *Manuscript.*

[15] See also the classic work of Makkai (1991) and Awodey and Forssell (2010).

Barrett, T. W. and Halvorson, H. (2013). How to count structure. *Manuscript.*

Barrett, T. W. and Halvorson, H. (2016a). Glymour and Quine on theoretical equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

Barrett, T. W. and Halvorson, H. (2016b). Morita equivalence. *The Review of Symbolic Logic*, 9(3):556–582.

Barrett, T. W. and Halvorson, H. (2017a). From geometry to conceptual relativity. *Forthcoming in Erkenntnis.*

Barrett, T. W. and Halvorson, H. (2017b). Quine's conjecture on many-sorted logic. *Forthcoming in Synthese.*

Belot, G. (2003). Notes on symmetries. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*. Cambridge.

Belot, G. (2013). Symmetry and equivalence. In *The Oxford Handbook of Philosophy of Physics*. Oxford.

Borceux, F. (1994). *Handbook of Categorical Algebra*, volume 1. Cambridge University Press.

Brading, K. and Castellani, E. (2007). Symmetries and invariances in classical physics. In Butterfield, J. and Earman, J., editors, *Handbook of the Philosophy of Science, Philosophy of Physics, Part B.*

Coffey, K. (2014). Theoretical equivalence as interpretative equivalence. *The British Journal for the Philosophy of Science*, 65(4):821–844.

Curiel, E. (2014). Classical mechanics is Lagrangian; it is not Hamiltonian. *The British Journal for the Philosophy of Science*, 65(2):269–321.

Dasgupta, S. (2015). Substantivalism vs relationalism about space in classical physics. *Philosophy Compass*, 10(9):601–624.

Dasgupta, S. (2016). Symmetry as an epistemic notion (twice over). *The British Journal for the Philosophy of Science*, 67(3):837–878.

de Bouvére, K. L. (1965). Synonymous theories. In *Symposium on the Theory of Models*, pages 402–406. North-Holland Publishing Company.

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

Dewar, N. (2016). Sophistication about symmetries. *Manuscript.*

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

Friedman, H. M. and Visser, A. (2014). When bi-interpretability implies synonymy. *Logic Group Preprint Series*, 320:1–19.

Geroch, R. (1978). *General Relativity from A to B.*

Glymour, C. (1971). Theoretical realism and theoretical equivalence. In *PSA 1970*, pages 275–288. Springer.

Glymour, C. (1977). The epistemology of geometry. *Noûs*, 11:227–251.

Glymour, C. (1980). *Theory and Evidence.* Princeton University Press.

Glymour, C. (2013). Theoretical equivalence and the semantic view of theories. *Philosophy of Science*, 80(2):286–297.

Halvorson, H. (2011). Natural structures on state space. *Manuscript.*

Halvorson, H. (2012). What scientific theories could not be. *Philosophy of Science*, 79(2):183–206.

Halvorson, H. (2013). The semantic view, if plausible, is syntactic. *Philosophy of Science*, 80(3):475–478.

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

Hodges, W. (2008). *Model Theory.* Cambridge University Press.

Hudetz, L. (2015). Linear structures, causal sets and topology. *Studies in History and Philosophy of Modern Physics*, pages 294–308.

Hudetz, L. (2016). Definable categorical equivalence. *Manuscript.*

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

Kanger, S. (1968). Equivalent theories. *Theoria*, 34(1):1–6.

Knox, E. (2014). Newtonian spacetime structure in light of the equivalence principle. *The British Journal for the Philosophy of Science*, 65(4):863–880.

Mac Lane, S. (1971). *Categories for the working mathematician.* Springer.

Makkai, M. (1991). *Duality and definability in first order logic*, volume 503. American Mathematical Society.

Malament, D. B. (1977). Causal theories of time and the conventionality of simultaneity. *Noûs*, pages 293–300.

Maudlin, T. (2012). *Philosophy of Physics: Space and Time.*

North, J. (2009). The 'structure' of physics: A case study. *The Journal of Philosophy*, 106:57–88.

Pelletier, F. J. and Urquhart, A. (2003). Synonymous logics. *Journal of Philosophical Logic*, 32(3):259–285.

Pinter, C. C. (1978). Properties preserved under definitional equivalence and interpretations. *Mathematical Logic Quarterly*, 24(31-36):481–488.

Quine, W. V. O. (1975). On empirically equivalent systems of the world. *Erkenntnis*, 9(3):313–328.

Rosenstock, S. (2015). On holonomy and fiber bundle interpretations of Yang-Mills theory. Unpublished manuscript.

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

Rosenstock, S. and Weatherall, J. O. (2016). A categorical equivalence between generalized holonomy maps on a connected manifold and principal connections on bundles over that manifold. *Journal of Mathematical Physics*, 57(10). arXiv:1504.02401 [math-ph].

Sklar, L. (1982). Saving the noumena. *Philosophical Topics*, pages 89–110.

Swanson, N. and Halvorson, H. (2012). On North's 'The structure of physics'. *Manuscript*.

Van Fraassen, B. C. (2014). One or two gentle remarks about Hans Halvorson's critique of the semantic view. *Philosophy of Science*, 81(2):276–283.

Weatherall, J. O. (2016a). Are Newtonian gravitation and geometrized Newtonian gravitation theoretically equivalent? *Erkenntnis*, 81(5):1073–1091.

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

Weatherall, J. O. (2017a). Category theory and the foundations of classical field theories. In Landry, E., editor, *Forthcoming in Categories for the Working Philosopher*. Oxford University Press.

Weatherall, J. O. (2017b). Regarding the 'hole argument'. *Forthcoming in the British Journal for the Philosophy of Science*.

Weyl, H. (1952). *Symmetry*. Princeton University Press.