Theories Are Not Partially Ordered

Thomas William Barrett and Hans Halvorson*

Abstract

This paper presents a simple example of first-order theories T_1 and T_2 such that i) T_1 can be embedded in T_2 and vice versa, ii) T_1 posits all of the structure of T_2 and vice versa, but iii) T_1 and T_2 are not equivalent. This shows that theories lack both the Cantor-Bernstein and co-Cantor-Bernstein properties and are neither partially ordered by the relation 'is embeddable in' nor by 'posits all of the structure of'. In addition, these results clarify the overall geography of notions of equivalence between theories and yield two philosophical payoffs related to the recent discussions of structure and equivalence.

1 Introduction

There is a long tradition in philosophy of comparing theories in terms of their ontological commitments and, in more recent literature, in terms of their structural commitments. The most famous case of this latter kind of comparison comes from the history of classical spacetime theories. It is standard to claim, for example, that the Galilean theory of spacetime posits less structure than the Newtonian theory of spacetime. The Newtonian theory comes equipped with the structure required to single out a preferred inertial frame as the rest frame, while the Galilean theory does not.[1]

It has recently been argued that we can appeal to this kind of structural comparisons between theories to inform our judgments about equivalence of theories (North, 2009; Barrett, 2019).[2] If two theories disagree in terms of their structural commitments, then we can infer that they are inequivalent. That is, if two theories posit different structures — like the Galilean and Newtonian theories of spacetime do — then they cannot 'have the same content' or 'say the same thing about the world'. It is uncontroversial that equivalent theories must agree in terms of their structural commitments. But there are some further natural principles relating structure and equivalence that are not as obvious. In this paper we will consider the following two.

*Draft of January 15, 2020. The authors can be reached at tbarrett@philosophy.ucsb.edu and hhalvors@princeton.edu. Thanks to Laurenz Hudetz, [[...]] for comments.

[1] See Geroch (1978), Friedman (1983), Earman (1989), and Maudlin (2012) for discussion.

[2] See Weatherall (2019b) for a review of the recent debate on equivalence of theories. See

Principle 1. If T_1 posits all of the structure of T_2 and T_2 posits all of the structure of T_1, then T_1 and T_2 are equivalent.

Principle 2. If T_1 can be embedded in of T_2 and T_2 can be embedded in T_1, then T_1 and T_2 are equivalent.

Principle 1 captures an intuitive idea that one might have about structure and equivalence. If one theory posits all of the structure of the other — in the sense that it can 'build' or 'define' all of that structure — and the other posits all of the structure of the one, then we would expect them to posit the same structure. Principle 2 is similarly intuitive. Suppose that one theory can be 'embedded' in the other and the other can be 'embedded' in the one. This means that we can think of the one as a 'sub-theory' of the other, and the other as a 'sub-theory' of the one. In this case, we would again expect the two theories to posit the same structure and to be equivalent. Surprisingly, however, neither Principle 1 nor Principle 2 is true. This means that theories are neither partially ordered by the relation 'posits all of the structure of' nor by the relation 'is embeddable in'. For reasons that we will discuss below, it is natural to call Principle 2 the 'Cantor-Bernstein property' of theories and Principle 1 the 'co-Cantor-Bernstein' property of theories.

The aim of this paper is to present a particularly simple example that demonstrates that first-order theories lack both the Cantor-Bernstein and co-Cantor-Bernstein properties. This means that comparing the structural commitments of theories requires much more care than might have initially been supposed.

2 Two theories

Consider the following two theories.[3]

The theory $\mathbf{T_1}$. T_1 is formulated in the signature $\Sigma_1 = \{\sigma, p_0, p_1, p_2, \ldots\}$ where σ is a sort symbol and each of the p_i is a unary predicate symbol. We define the theory as follows:

$$T_1 = \{\exists_{=1} x(x = x)\}$$

T_1 says that there is exactly one thing, but it is silent on whether or not that thing is p_i.

The theory $\mathbf{T_2}$. T_2 is formulated in the signature $\Sigma_2 = \{\sigma, q_0, q_1, q_2, \ldots\}$ where σ is a sort symbol and each of the q_i is a unary predicate symbol. We define the theory as follows:

$$T_2 = \{\exists_{=1} x(x = x), \forall x(q_0(x) \to q_1(x)), \forall x(q_0(x) \to q_2(x)), \ldots\}$$

[3] For preliminaries on model theory the reader is encouraged to consult Hodges (2008). For additional discussion of these two theories see Halvorson (2012) and Barrett and Halvorson (2016b).

T_2 says that there is exactly one thing, and that if that thing is q_0, then it is q_i for all of the other i too. One can think of the predicate q_0 as a kind of 'light switch' that turns on all of the other predicates q_i.

The first task of this paper is to examine these theories and the relationships between them. In particular, we will demonstrate the following five claims.

1. T_1 and T_2 are not definitionally equivalent.
2. T_1 and T_2 are not Morita equivalent.
3. There are conservative translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$.
4. There are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$.
5. There is no essentially surjective and conservative translation between T_1 and T_2 (in either direction).

It is natural to separate these claims into two kinds. The first two claims are about what kinds of extensions exist for T_1 and T_2 (i.e. they have no common definitional nor Morita extension), while the last three are about what kinds of translations exist between T_1 and T_2. After proving these claims, we will return to the general philosophical issues mentioned above and discuss how these results demonstrate the falsity of Principles 1 and 2.

3 Extension

Let $\Sigma \subset \Sigma^+$ be signatures with $p \in \Sigma^+ - \Sigma$ an n-ary predicate symbol. Recall that an **explicit definition** of p in terms of Σ is a Σ^+-sentence of the form

$$\forall x_1 \dots \forall x_n (p(x_1, \dots x_n) \leftrightarrow \phi(x_1, \dots, x_n))$$

where $\phi(x_1, \dots, x_n)$ is a Σ-formula. A **definitional extension** of a Σ-theory T to the signature Σ^+ is a Σ^+-theory

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

such that for each predicate symbol $s \in \Sigma^+ - \Sigma$, the sentence δ_s is an explicit definition of s in terms of Σ. One can also define new function and constant symbols, but for our purposes this will not be important.

Recall that two theories are **definitionally equivalent** if they have a 'common definitional extension'. More precisely: if T_1 is a Σ_1-theory and T_2 is a Σ_2-theory, T_1 and T_2 are definitionally equivalent if there is a definitional extension T_1^+ of T_1 to the signature $\Sigma_1 \cup \Sigma_2$ and a definitional extension T_2^+ of T_2 to the signature $\Sigma_1 \cup \Sigma_2$ such that T_1^+ and T_2^+ are logically equivalent (in the sense that they have the same class of models).

We now have the following simple proposition.

Proposition 1. *T_1 and T_2 are not definitionally equivalent.*

Proof. Suppose for contradiction that T is a common definitional extension of T_1 and T_2. Since T defines each of the predicate symbols of T_2, there is a Σ_1-sentence ϕ such that $T \vDash \forall y q_0(y) \leftrightarrow \phi$. Now one can verify that the sentence ϕ has the following property: If ψ is a Σ_1-sentence and $T_1 \vDash \psi \rightarrow \phi$, then either (i) $T_1 \vDash \neg\psi$ or (ii) $T_1 \vDash \phi \rightarrow \psi$. But ϕ cannot have this property. Consider the Σ_1-sentence

$$\phi \wedge \forall x p_i(x)$$

where p_i is a predicate symbol that does not occur in ϕ. We trivially see that $T_1 \vDash (\phi \wedge \forall x p_i(x)) \rightarrow \phi$, but neither (i) nor (ii) hold of $\phi \wedge \forall x p_i(x)$. This is a contradiction, so T_1 and T_2 are not definitionally equivalent. □

It has recently been suggested that definitional equivalence is too strict a standard of equivalence between theories, in the sense that it judges theories to be inequivalent that we have good reason to consider equivalent. For example, Euclidean geometry can be formulated with only a sort of points (Tarski, 1959), with only a sort of lines (Schwabhäuser and Szczerba, 1975), or with both a sort of points and a sort of lines (Hilbert, 1930).[4] Since these formulations use different sort symbols, and we have so far provided no way of defining new sort symbols, definitional equivalence does not capture any sense in which they are equivalent. In order to address this shortcoming of definitional equivalence, a more liberal standard of equivalence has been proposed. It has been called "Morita equivalence" (by Barrett and Halvorson (2016b, 2017a,b)) and "generalized definitional equivalence" (by Andréka et al. (2008)).[5]

The precise details of Morita equivalence are not important for our purposes here, but the basic idea is simple. Morita equivalence allows one to define new *sort* symbols — in addition to new predicate, function, and constant symbols — using some basic construction rules. Two theories are then said to be Morita equivalent if they have a 'common Morita extension', which is just like a common definitional extension except that it might define new sorts. One can show that geometry formulated in terms of points is Morita equivalent to geometry formulated in terms of lines; one uses the sort of lines to build the sort of points, and vice versa (Barrett and Halvorson, 2017a).

One might wonder whether our theories T_1 and T_2 are Morita equivalent. The following simple result has already been demonstrated.[6]

Proposition 2. *T_1 and T_2 are not Morita equivalent.*

Proof. The proof proceeds in precisely the same manner as the proof of Proposition 1. Barrett and Halvorson (2016b, Theorem 5.2) give the precise details. □

[4] See Szczerba (1977) and Schwabhäuser et al. (1983, Proposition 4.59, Proposition 4.89).

[5] See also Hudetz (2017a,b) and Tsementzis (2015).

[6] The result goes through even if one formulates T_1 and T_2 using two different sort symbols, instead of the same sort symbol σ.

4 Translation

Our next three claims are about the kinds of 'translations' that exist between these theories. We need some basic preliminaries.

Let Σ_1 and Σ_2 be signatures. A **reconstrual** F of Σ_1 into Σ_2 is a map from the predicates in the signature Σ_1 to Σ_2-formulas that takes an n-ary predicate symbol $p \in \Sigma_1$ to a Σ_2-formula $Fp(x_1, \ldots, x_n)$ with n free variables.[7] One can think of the Σ_2-formula $Fp(x_1, \ldots, x_n)$ as the 'translation' of the Σ_1-formula $p(x_1, \ldots, x_n)$ into the signature Σ_2. We will use the notation $F : \Sigma_1 \to \Sigma_2$ to denote a reconstrual F of Σ_1 into Σ_2.

A reconstrual $F : \Sigma_1 \to \Sigma_2$ extends to a map from arbitrary Σ_1-formulas to Σ_2-formulas in the natural recursive manner. In the case where one is only considering signatures with predicate symbols (as we are here), this map is particularly easy to describe. Let $\phi(x_1, \ldots, x_n)$ be a Σ_1-formula. We define the Σ_2-formula $F\phi(x_1, \ldots, x_n)$ recursively as follows.

- If $\phi(x_1, \ldots, x_n)$ is $x_i = x_j$, then $F\phi(x_1, \ldots, x_n)$ is the Σ_2-formula $x_i = x_j$.
- If $\phi(x_1, \ldots, x_n)$ is $p(x_1, \ldots, x_n)$, where $p \in \Sigma_1$ is an n-ary predicate symbol, then $F\phi(x_1, \ldots, x_n)$ is the Σ_2-formula $Fp(x_1, \ldots, x_n)$.
- If $F\phi$ and $F\psi$ have already been defined for Σ_1-formulas ϕ and ψ, then we define the Σ_2-formula $F(\neg\phi)$ to be $\neg F\phi$, $F(\phi \wedge \psi)$ to be $F\phi \wedge F\psi$, $F(\forall x\phi)$ to be $\forall x F\phi$, etc.

Suppose that T_1 and T_2 are theories in the signatures Σ_1 and Σ_2, respectively. We say that a reconstrual $F : \Sigma_1 \to \Sigma_2$ is a **translation** $F : T_1 \to T_2$ if $T_1 \vDash \phi$ implies that $T_2 \vDash F\phi$ for every Σ_1-sentence ϕ. A translation F gives rise to a map $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$, which takes models of the theory T_2 to models of the theory T_1. For every model A of T_2 we first define a Σ_1-structure $F^*(A)$ as follows.

- $\mathrm{dom}(F^*(A)) = \mathrm{dom}(A)$.
- $(a_1, \ldots, a_n) \in p^{F^*(A)}$ if and only if $A \vDash Fp[a_1, \ldots, a_n]$.

A straightforward argument demonstrates that $F^*(A)$ is indeed a model of T_1 (Barrett and Halvorson, 2016a, §4).

A translation $F : T_1 \to T_2$ is **conservative** if $T_2 \vDash F\phi$ implies that $T_1 \vDash \phi$ for any Σ_1-sentence ϕ. We make two brief remarks about this concept. First, one can think of a conservative translation as a kind of 'injection' or 'embedding' on sentences. Indeed, one can easily verify that the following simple condition is equivalent to a translation $F : T_1 \to T_2$ being conservative.

- For any Σ_1-sentences ϕ_1 and ϕ_2, if $T_2 \vDash F\phi_1 \leftrightarrow F\phi_2$ then $T_1 \vDash \phi_1 \leftrightarrow \phi_2$.

[7] This notion naturally extends to signatures that contain function and constant symbols, but that will be unimportant for our purposes. See Hodges (2008), Button and Walsh (2018), and Barrett and Halvorson (2016a) for details.

This condition is clearly capturing a kind of injectivity on sentences: If a translation maps two Σ_1-sentences to equivalent Σ_2-sentences, then they must have been equivalent to begin with. Second, one can easily verify that if a translation $F : T_1 \to T_2$ is such that $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ is surjective, then F is conservative.

We now have our first simple result concerning the existence of translations between our theories T_1 and T_2.

Proposition 3. *There are conservative translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$.*

Proof. Consider the reconstruals $F : \Sigma_1 \to \Sigma_2$ and $G : \Sigma_2 \to \Sigma_1$ defined by

$$F : p_i \longmapsto q_{i+1} \qquad\qquad G : q_i \longmapsto p_0 \vee p_i$$

It is trivial that $F : T_1 \to T_2$ is a translation. Since G maps the Σ_2-sentence $\forall x(q_0(x) \to q_i(x))$ to $\forall x((p_0(x) \vee p_0(x)) \to (p_0(x) \vee p_i(x)))$ and T_1 entails this latter sentence, it follows that G is a translation too.

It remains to show that F and G are conservative. One does this by showing that F^* and G^* are surjective. Suppose that M is a model of T_1. Then M is completely determined by which of the p_i hold of the one thing. We let N be the model of T_2 defined as follows: N has the same domain as M, q_0 does not hold of the one thing in N, and q_{i+1} holds of the one thing in N if and only if p_i holds of the only thing in M. One trivially sees that $F^*(N) = M$, so F^* is surjective. A similar argument shows that G^* is surjective. □

Since they are conservative, the translations F and G from Proposition 3 can be thought of as injections or embeddings between the theories T_1 and T_2. But there is a natural sense in which neither is 'surjective'. For example, F does not map any Σ_1-formula to any logical equivalent of the Σ_2-formula $q_0(x)$. The following property makes this thought precise. We say that a translation $F : T_1 \to T_2$ is **essentially surjective** if for every Σ_2-formula ψ there is a Σ_1-formula ϕ such that $T_2 \vDash \forall x_1 \ldots \forall x_n(\psi(x_1, \ldots, x_n) \leftrightarrow F\phi(x_1, \ldots, x_n))$. One can easily verify that neither F nor G from Proposition 3 are essentially surjective. But we do have the following simple result.

Proposition 4. *There are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$.*

Proof. Consider the reconstruals $H : \Sigma_1 \to \Sigma_2$ and $K : \Sigma_2 \to \Sigma_1$ defined by

$$H : p_i \longmapsto q_i \qquad\qquad K(q_i) = \begin{cases} p_0 \wedge \neg p_0 & \text{if } i = 0 \\ p_{i-1} & \text{otherwise} \end{cases}$$

It is trivial that H is an essentially surjective translation. Since K maps the Σ_2-sentence $\forall x(q_0(x) \to q_i(x))$ to $\forall x((p_0(x) \wedge \neg p_0(x)) \to p_{i-1}(x))$ and T_1 entails this latter sentence, it follows that K is a translation. It is easy to see that K is essentially surjective. □

One can verify that the theory T_1 does not entail the sentence $\forall x(p_0(x) \rightarrow p_1(x))$, but T_2 does entail $H(\forall x(p_0(x) \rightarrow p_1(x)))$. Similarly, the theory T_2 does not entail $\neg\forall x q_0(x)$, but T_1 does entail $K(\neg\forall x q_0(x))$. This means that neither H nor K is conservative.

One therefore wonders whether there is any translation between T_1 and T_2 that is essentially surjective and conservative. One might consider a translation like this to be an 'isomorphism' between the two theories. The following proposition settles this issue.

Proposition 5. *There is no essentially surjective and conservative translation between T_1 and T_2 (in either direction).*

Proof. Suppose that there is a translation $F : T_1 \rightarrow T_2$ that is essentially surjective and conservative. We show that this implies that T_1 and T_2 are definitionally equivalent, which contradicts Proposition 1. (The same argument demonstrates that there is no essentially surjective and conservative translation $T_2 \rightarrow T_1$.)

Let q_i be one of the predicate symbols in Σ_2. Since F is essentially surjective, there is some Σ_1-formula ϕ_i such that $T_2 \vDash \forall x(q_i(x) \leftrightarrow \phi_i(x))$. This allows us to define a reconstrual $G : \Sigma_2 \rightarrow \Sigma_1$ by setting $G(q_i) = \phi_i$. One then uses the fact that F is conservative to verify that $G : T_2 \rightarrow T_1$ is a translation, and to verify that both $T_1 \vDash \phi \leftrightarrow GF\phi$ and $T_2 \vDash \psi \leftrightarrow FG\psi$ for all Σ_1-formulas ϕ and Σ_2-formulas ψ. This means that T_1 and T_2 are intertranslatable, and so (Barrett and Halvorson, 2016a, Theorem 2) implies that they are definitionally equivalent. □

5 Conclusion

We conclude by discussing three philosophical payoffs that these technical claims yield.

Mutual faithful interpretability and Morita equivalence

The first concerns the recent debate about when two theories should be considered equivalent. A significant amount of attention has been devoted to better understanding the relationships between different standards of equivalence of theories and to applying these standards to various cases of interest.[8] The simple example that we have considered here allows us to further clarify the overall geography of standards of equivalence between theories.

[8] For discussion of these relationships, see for example Barrett and Halvorson (2016a,b) and (Button and Walsh, 2018, p. 118). In addition to the recent articles on equivalence already cited, see Barrett (2017), Coffey (2014), Curiel (2014), Halvorson (2013), Glymour (2013), Hudetz (2015, 2017a), Knox (2011, 2014), North (2009), Rosenstock et al. (2015), Rosenstock and Weatherall (2016), Teh and Tsementzis (2017), Van Fraassen (2014), and Weatherall (2016, 2017), and the references therein. See Weatherall (2019a) for a survey of recent work.

Recall that T_1 and T_2 are **mutually faithfully interpretable** if there are conservative translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$.[9] It is already well known that definitional equivalence is a stricter standard of equivalence than mutual faithful interpretability. More precisely, any pair of theories that are definitionally equivalent are mutually faithfully interpretable, but there are mutually faithful interpretable theories that are not definitionally equivalent (Andréka et al., 2005; Button and Walsh, 2018). Morita equivalence, however, is a more liberal notion of equivalence than definitional equivalence. So it is natural to ask the following question.

> *If two theories T_1 and T_2 are mutually faithfully interpretable, then are T_1 and T_2 also Morita equivalent?*

If the answer to this question were yes — that is, if mutual faithful interpretability entailed Morita equivalence — this would be a mark against Morita equivalence as a plausible standard of equivalence between theories. It is widely accepted that mutual faithful interpretability is too liberal a standard of equivalence; it considers theories to be equivalent that we have good reason to consider inequivalent (Szczerba, 1977). So if the answer to the above question were yes, this would immediately imply that Morita equivalence is too liberal a standard as well.[10]

The answer to the above question, however, is no. It is not the case that Morita equivalence is entailed by mutual faithful interpretability. Our discussion here has demonstrated precisely this. It follows from Proposition 3 that our theories T_1 and T_2 are mutually faithfully interpretable, but as was stated in Proposition 2, they are not Morita equivalent.

The second and third payoffs of this discussion are of a more straightforwardly philosophical nature. This example yields the following two surprising conclusions about theories. It can be that one theory is embeddable in another and the other is embeddable in the one, but the two theories are not equivalent. And it can be that one theory 'posits all of the structure' of another and the other 'posits all of the structure' of the one, but the two theories do not posit the same structure.

The Cantor-Bernstein property

We begin by discussing the former. Recall Principle 2 from above.

Principle 2. If T_1 can be embedded in T_2 and T_2 can be embedded in T_1, then T_1 and T_2 are equivalent.

Proposition 3 captures a sense in which our theory T_1 can be embedded in T_2 and our theory T_2 can be embedded in T_1. The existence of the conservative

[9] See Button and Walsh (2018, p. 117). Our presentation here is slightly less general, since we have not discussed translations with parameters.

[10] See McEldowney (2019), for example, for worries about Morita equivalence along these lines.

translation $F : T_1 \to T_2$ shows us that T_1 can be 'viewed as a part of' T_2, or in other words, that T_1 is a 'sub-theory' of T_2. It is easy to see how this is the case: T_1 can be viewed as the part of T_2 that is constructed from the predicate symbols $q_1, q_2, \ldots$ and so on. And similarly the existence of the conservative translation $G : T_2 \to T_1$ shows us that one can view T_2 as a part of T_1; it can be viewed as the part of T_1 that is constructed from the Σ_1-formulas $p_0, p_0 \vee p_1, p_0 \vee p_2, \ldots$ and so on. Despite the fact that each of these theories can be embedded in the other, they are not the same theory. Propositions 1, 2, and 5 substantiate this intuition. Our theories T_1 and T_2 are not definitionally equivalent, Morita equivalent, nor does there exist an essentially surjective and conservative translation between them.

What we have shown here is that Principle 2 is false and that theories lack the Cantor-Bernstein property. The Cantor-Bernstein theorem famously says that if there are injections $f : X \to Y$ and $g : Y \to X$ between sets X and Y, then there is a bijection between X and Y. Because of this theorem we say that sets have the Cantor-Bernstein property. And indeed, it makes sense to talk about the Cantor-Bernstein property for any category. One says that a category C has the Cantor-Bernstein property if for any objects c and d of C whenever there is a monomorphism (i.e. a generalization of the concept of an injection) from $c \to d$ and a monomorphism $d \to c$, the objects c and d are isomorphic. The property captures a basic intuition one might have about the notion of 'being a part of': If X is a part of Y and Y is a part of X, then X and Y are the same. Our example here demonstrates that the category of theories does not have the Cantor-Bernstein property.[11] In other words, the analogue of the Cantor-Bernstein theorem does not hold of theories: T_1 and T_2 can be embedded into one another, but nonetheless they are not the same.[12]

This means that theories are not partially ordered by the relation 'is embeddable in'. Recall that a partial order is one that is reflexive, anti-symmetric, and transitive. It is trivially the case that the relation 'is embeddable in' is reflexive and transitive on first-order theories. Since the identity translation is a conservative translation, every theory is embeddable in itself. And if T_1 is embeddable in T_2 and T_2 is embeddable in T_3, then — since the composition of conservative translations is indeed a conservative translation — T_1 is embeddable in T_3. But the results presented here imply that 'is embeddable in' is not anti-symmetric. In order to be anti-symmetric, it would have to be the case that there are not *distinct* theories T_1 and T_2 that are each embeddable in the other. The results above demonstrate that there *are* theories T_1 and T_2 that are embeddable in one another and are distinct, in the sense that they are not equivalent.

[11] There are various ways to 'categorify' the collection of first-order theories. The reader is encouraged to consult Visser (2006) for details.

[12] Another example of the failure of the Cantor-Bernstein property is the relation between classical and intuitionistic logic. Obviously, intuitionistic logic can be embedded in classical logic. The famous Gödel translation shows that intuitionistic logic can be embedded in classical logic. But it is nonetheless unnatural to think of the two logics as equivalent, as was recently pointed out, for example, by Dewar (2018).

The Co-Cantor Bernstein property

We now turn to our final payoff. It can be that one theory 'posits all of the structure' of another and the other 'posits all of the structure' of the one, but the two theories do not posit the same structure. In other words, Principle 1 is false.

Principle 1. If T_1 posits all of the structure of T_2 and T_2 posits all of the structure of T_1, then T_1 and T_2 are equivalent.

This means that — just like the relation 'is embeddable in' — the relation 'posits all of the structure of' is not a partial order on theories. It is reflexive and transitive, but it is not anti-symmetric. There are inequivalent theories T_1 and T_2 that both posit all of the structure of the other. It has already been argued that there are theories that posit 'incomparable' amounts of structure, in the sense that neither posits more nor less structure than the other (Barrett, 2015a,b). But this example shows that the structural commitments of theories behave even more strangely than was initially thought. It is worth taking a moment to discuss this in detail.

One might put this payoff as follows: It can be that one theory 'is as ideologically rich as' another and the other 'is as ideologically rich as' the one, but the two theories are not ideologically equivalent. The ideology of a theory is usually thought of as the range of concepts that are expressible in the language in which the theory is formulated (Quine, 1951). It takes a moment to explain why our example yields this surprising conclusion about structure and ideology.

The existence of an essentially surjective translation $H : T_1 \to T_2$ captures a sense in which T_1 has all the structure of (or is as ideologically rich as) T_2. The idea here is simple. When H is essentially surjective, any formula ψ that one can formulate in the language of T_2 is expressible using the language of T_1; that is precisely what the essential surjectivity of H guarantees. There is some formula ϕ in the language of T_1 that H translates to a logical equivalent of ψ. Intuitively, this means that the theory T_1 can define all of the structures that T_2 has. Or in other words, T_1 can express all of the same concepts that T_2 can. Quine (1951, p. 15) himself remarks that one can investigate the ideology of a theory by examining what kinds of translations exist between theories: "Much that belongs to ideology can be handled in terms merely of the translatability of notations from one language into another; witness the mathematical work on definability by Tarski and others."

One can grasp the basic idea here by considering the following two simple examples. Suppose that T is a Σ-theory and consider the signature $\Sigma^+ = \Sigma \cup \{p\}$, where p is a new unary predicate symbol not contained in Σ. We consider the following two cases.

1. Suppose that T^+ is the Σ^+-theory that has precisely the same axioms as T. There is a natural sense in which T^+ has all of the structure of T, but not vice versa; T^+ has the new piece of structure p that T lacks. One

can capture this basic idea by looking to facts about essentially surjective translations between these two theories.

First, there is an essentially surjective translation $T^+ \rightarrow T$. The translation simply maps each symbol in Σ to itself and maps p to any Σ-formula with one free variable. It is trivial to verify that this is indeed an essentially surjective translation. This makes precise our basic intuition that T^+ is as ideologically rich as T. Second, since there is no Σ-formula that is logically equivalent to p, the translation $T \rightarrow T^+$ that maps every element of Σ to itself is not essentially surjective. This makes precise our intuition that T is *not* as ideologically rich as T^+.

2. Now suppose instead that T^+ is a definitional extension of T to the signature Σ^+. In this case there is a strong sense in which T and T^+ have precisely the same structure and ideological richness; p is not a piece of structure that is new to T^+. Rather, it is explicitly definable in terms of those structures that T posits, so there is a strong sense in which T itself already posits the structure p.

   This basic intuition can be made precise by noticing that there are essentially surjective translations $T \rightarrow T^+$ and $T^+ \rightarrow T$. It follows from our discussion in point 1 above that there is an essentially surjective translation $T^+ \rightarrow T$. And since T^+ is a definitional extension of T, the translation $T \rightarrow T^+$ that maps every element in Σ to itself is in this case essentially surjective. The Σ-formula ϕ that defines p is logically equivalent to p.

The existence of an essentially surjective translation from T_1 to T_2 captures a strong sense in which T_1 has all of the structure or ideological richness of T_2. The two examples above show that this idea is a direct generalization of the kinds of simple examples that formed our intuitions about amounts of structure and ideological richness in the first place.

In addition, the tools that have recently been used to compare amounts of structure between theories cohere well with the idea that essentially surjective translations capture facts about structure. It has been suggested — by Earman (1989), and more recently by North (2009), Swanson and Halvorson (2012), and Barrett (2015a,b) — that the size of an object's automorphism group can be used as a guide to the amount of structure that the object has. An automorphism of an object is a structure-preserving bijection from the object to itself. If an object has more automorphisms, therefore, that suggests that the object has less structure that these automorphisms are required to preserve. The important fact for our purposes here is the following: If there is an essentially surjective translation from T_1 to T_2, this captures a sense in which the models of these two theories have automorphism groups of precisely the same size.

It only takes a moment to make this claim precise. An **automorphism** of a Σ-structure M is a bijection $f : M \rightarrow M$ that satisfies $M \vDash p[a_1, \ldots, a_n]$ if and only if $M \vDash p[f(a_1), \ldots f(a_n)]$ for any predicate symbol $p \in \Sigma$ and elements $a_1, \ldots, a_n \in M$. One can easily prove the following.[13]

[13] A similar argument actually demonstrates an even stronger result: If F is essentially

Proposition 6. *Let $H : T_1 \to T_2$ be an essentially surjective translation with N a model of T_2. Then $H^*(N)$ and N have the same automorphism group.*

Proof. If $f : N \to N$ is an automorphism of N, one can easily verify using the definition of H^* that f is also an automorphism of $H^*(N)$. Suppose that $f : H^*(N) \to H^*(N)$ is an automorphism and let $q \in \Sigma_2$ be a predicate symbol in the signature of T_2. We immediately see that for any $a \in N$:

$$N \vDash q[a] \iff H^*(N) \vDash \phi[a] \iff H^*(N) \vDash \phi[f(a)] \iff N \vDash q[f(a)]$$

Here ϕ is a Σ_1-formula such that $H\phi$ is logically equivalent to q, whose existence is guaranteed by the essential surjectivity of H. The first and third biconditionals follow from this choice of ϕ and the definition of H^*, while the second follows from the fact that f is an automorphism of $H^*(N)$. This means that f is an automorphism of N. □

Our current best methods of comparing amounts of structure therefore lend further support to the following claim: The existence of an essentially surjective translation from T_1 to T_2 captures a sense in which T_1 has all of the structure or ideological richness of T_2.

Now that we have this idea on the table, we can return to consider our pair of theories T_1 and T_2. As we have seen in Proposition 4, there are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$. This captures a sense in which T_1 has all of the structure of (or is as ideologically rich as) T_2 and T_2 has all of the structure of (or is as ideologically rich as) T_1. But Propositions 1, 2, and 5 imply that T_1 and T_2 do not have the *same structure* or the *same ideology*. This is completely intuitive: T_2 has a special 'light switch' predicate q_0 that 'turns on' all of the other predicates q_i, while T_1 contains no such structure.

This shows that Principle 1 is false and theories lack what one might call the 'co-Cantor-Bernstein property'. It is easy to verify that if there are surjections $f : X \to Y$ and $g : Y \to X$ between sets X and Y, then there is a bijection between X and Y. One might call this the 'co-Cantor-Bernstein theorem', and therefore say that sets have the co-Cantor-Bernstein property. Once again, it makes sense to talk about the co-Cantor-Bernstein property for any category. We can say that a category C has the co-Cantor-Bernstein property if for any objects c and d of C whenever there is an epimorphism (i.e. a generalization of the concept of an surjection) from $c \to d$ and an epimorphism $d \to c$, the objects c and d are isomorphic. Our example here demonstrates that the category of theories does not have this proeprty. Our two theories T_1 and T_2 can be surjected onto one another, but nonetheless, as Propositions 1, 2, and 5 tell us, they are not equivalent. These concepts of amount of structure and ideological richness are therefore much more subtle than one might have initially thought.

One might wonder how these results come to bear on the current debates about structure and equivalence in philosophy of science. Real scientific theories are, of course, much more complicated than theories in first-order logic.

surjective translation, then F^* is a full functor between the categories of models of T_2 and T_1 (Halvorson, 2019; Barrett, 2018).

We therefore should be careful not to assume that the 'category of all scientific theories' has all the features that the category of first-order theories has. Nonetheless, we should take 'no go' results like the ones we have presented here seriously. If the category of first-order theories fails to have some simple feature — like the Cantor-Bernstein or co-Cantor-Bernstein property — then that gives us reason to doubt that the category of all scientific theories will have that feature. For example, it would seem natural to think that if T_1 is reducible to T_2, and T_2 is reducible to T_1, then T_1 and T_2 are equivalent theories. This is exactly the kind of inference, however, that the results presented here about first-order theories call into question.

References

Andréka, H., Madarasz, J., and Németi, I. (2008). Defining new universes in many-sorted logic. *Mathematical Institute of the Hungarian Academy of Sciences, Budapest*, 93.

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.

Barrett, T. (2019). Structure and equivalence. *Forthcoming in Philosophy of Science*.

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. (2017). Equivalent and inequivalent formulations of classical mechanics. *Forthcoming in the British Journal for the Philosophy of Science*.

Barrett, T. W. (2018). 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. *Erkenntnis*, 82(5):1043–1063.

Barrett, T. W. and Halvorson, H. (2017b). Quine's conjecture on many-sorted logic. *Synthese*, 194(9):3563–3582.

Button, T. and Walsh, S. (2018). *Philosophy and Model Theory*. Oxford University Press.

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.

Dewar, N. (2018). On translating between logics. *Analysis*, 78(4):622–630.

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

Friedman, M. (1983). *Foundations of space-time theories: Relativistic physics and philosophy of science*. Princeton University Press.

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

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

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. (2019). *The Logic in Philosophy of Science*. Cambridge University Press.

Hilbert, D. (1930). *Grundlagen der Geometrie*. Teubner.

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. (2017a). Definable categorical equivalence: Towards an adequate criterion of theoretical intertranslatability. *Forthcoming in Philosophy of Science*.

Hudetz, L. (2017b). The semantic view of theories and higher-order languages. *Forthcoming in Synthese*.

Knox, E. (2011). Newton-Cartan theory and teleparallel gravity: The force of a formulation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42(4):264–275.

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

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

McEldowney, P. A. (2019). On Morita equivalence and interpretability. *Forthcoming in The Review of Symbolic Logic*.

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

Quine, W. V. O. (1951). Ontology and ideology. *Philosophical Studies*, 2(1):11–15.

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].

Schwabhäuser, W. and Szczerba, L. (1975). Relations on lines as primitive notions for Euclidean geometry. *Fundamenta Mathematicae*.

Schwabhäuser, W., Szmielew, W., and Tarski, A. (1983). *Metamathematische Methoden in der Geometrie*. Springer.

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

Szczerba, L. (1977). Interpretability of elementary theories. In *Logic, Foundations of Mathematics, and Computability Theory*. Springer.

Tarski, A. (1959). What is elementary geometry? In *The Axiomatic Method With Special Reference to Geometry and Physics*. North-Holland.

Teh, N. and Tsementzis, D. (2017). Theoretical equivalence in classical mechanics and its relationship to duality. *Forthcoming in Studies in History and Philosophy of Modern Physics*.

Tsementzis, D. (2015). A syntactic characterization of Morita equivalence. *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.

Visser, A. (2006). Categories of theories and interpretations. In *Logic in Tehran. Proceedings of the workshop and conference on Logic, Algebra and Arithmetic, held October 18–22, 2003*. ASL.

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

Weatherall, J. O. (2017). 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. (2019a). Classical spacetimes. In Knox, E. and Wilson, A., editors, *The Routledge Companion to Philosophy of Physics*. Routledge.

Weatherall, J. O. (2019b). Theoretical equivalence in physics. *Forthcoming in Philosophy Compass*.