THE HALVORSON EXAMPLE

TOBY MEADOWS

ABSTRACT. Halvorson has proposed an intriguing example of a pair of theories whose categories are equivalent but which are not themselves definitionally equivalent. Moreover, it seems obvious that these theories are not equivalent in any intuitive sense. We offer a new topological proof that these theories are not definitionally equivalent. However, the underlying theorem for this claim has a converse that shows a surprising collection of theories, which are superficially similar to those in Halvorson's example, turn out to be definitionally equivalent after all. This offers some new insight into what is going "wrong" in the Halvorson example.

In his seminal paper on scientific theories, Hans Halvorson puts forward an intriguing example of a pair of theories that are not definitionally equivalent, but which are equivalent as categories [Halvorson, 2012].[1] The theories involved in the *Halvorson example* are generally thought to be intuitively and somewhat obviously inequivalent. This thought has motivated pointed questions of the appropriateness of categorical equivalence as a tool for understanding when two theories are genuinely equivalent. It has also spawned a number of attempts to find more suitable halfway houses between categorical equivalence and definitional equivalence [Barrett and Halvorson, 2016, Hudetz, 2019, March, 2024].

My goal in this short paper is to provide a new topological proof that the theories involved in the Halvorson example are not definitionally equivalent. But more interestingly, the lemmas used in this proof have converses that can be used to show that a surprising variety of pairs of theories, which are similar to those of the Halvorson example, are – in fact – definitionally equivalent. These examples put some pressure on the thought that the original pair of theories were so obviously inequivalent after all. For this reason alone, it seems like a good idea to record this observation. However, the techniques used in this paper will also hopefully open a door to some new approaches to – the often difficult – problem establishing that a pair of theories are not definitionally equivalent.

The paper is organized as follows. We start in Section 1 by recalling the example and sketching the argument that the theories are equivalent as categories. We then use the Cantor-Bernstein theorem and an infinitary logic to show that there is a weak sense in which these theories are interdefinable. In Section 2, we give a new proof of Halvorson's result that the interdefinability

I'd like to thank Elliot Glazer for his invaluable assistance with Section 2.2. I'd also like to thank Jim Weatherall for some very helpful comments on a draft of this paper. I'd like to than Neil Dewar for some probing questions about this work. And finally, I'd like to thank the LPS community for showing up to a rather long impromptu talk about this material and other things in the Summer of 2024.

[1] The relevant definitions can be found in, for example, [Barrett and Halvorson, 2016].

cannot be executed in first order logic; i.e., the theories are not definitionally equivalent. We proceed by remolding the problem into a problem about propositional logic and then into a problem of topology. These results deliver a theorem from which Halvorson's result directly follows, but it also delivers a raft of, so to speak, converse cases establishing that a surprising variety of theories similar to those in the example are all definitionally equivalent. Finally in Section 3, we offer some informal discussion of the significance of these results.

1. The Example

There have been a number of presentations of the Halvorson Example, but we shall begin with a version in the setting of first order logic. Our theories T and S are described below:

- Let T be the theory in the language $\mathcal{L}_T = \{P_n, a\}_{n\in\omega}$ with one sort σ where each P_n has arity σ and a is a constant symbol. Let T consist of the axiom that says there is exactly one object.
- Let S be the theory in the language $\mathcal{L}_S = \{Q_n, b\}_{n\in\omega}$ with one sort τ where each Q_n has arity τ and b is a constant symbol. Let S say that there is exactly one object and include the following axioms, for all $n \in \omega$

$$(Q_0 b \to Q_n b).$$

A little informally, T talks about a single object, a, in language with infinitely many predicates and says absolutely nothing about whether they apply to a. S on the other hand, also talks about a single object, b, in a language with infinitely many predicates, but it says that if the first of those predicates holds of b, then so do all the rest. Given that T says almost nothing and S, says a little more, it seems natural to think that these theories are not equivalent. But how do we make that precise?

Proposition 1. *[Halvorson, 2012] T and S are not definitionally equivalent.*

So far so good. But definitional equivalence isn't the only criterion for theoretical equivalence in town. Moreover, in contexts where first order logic isn't obviously available – as is often the case in physics – it's not clear that definitional equivalence even makes sense. It would, thus, be pleasing to have an alternative and more general criterion for theoretical equivalence. One such criterion is *categorical equivalence* [Weatherall, 2021].

When considered as categories, it turns out that T and S *are* equivalent. We shall deliver a quick rehearsal of that argument as it will also give us the opportunity to introduce some helpful terminology and offer a simpler perspective on the problem. In fact, we'll show that T and S are categorically isomorphic, which is generally stronger than categorical equivalence. Recall that the theory category $mod(T)$ of T is the category whose objects are models of T and whose arrows are elementary embeddings between them. We define $mod(S)$ similarly.

$mod(T)$ and $mod(S)$ are isomorphic if there is a functor F between them which is a bijection on both objects and arrows.[2]

Note that models of T and S both have an unusual – and for our purposes, helpful – feature: they all have exactly one object in them. This tells us that: there is at most one arrow between any pair of models in these categories; and as such, every arrow is an isomorphism between those models. This simplifies the problem greatly. In order to establish an isomorphism between these categories we just need to show there is a bijection between the models of T and S such that any pair of models from one isomorphism class from $mod(T)$ get sent to models in the same isomorphism class of $mod(S)$. So *really* we just want to show that there is a bijection between the isomorphism types of T and the isomorphism types of S. At this point, one might simply observe that T and S both have continuum many isomorphism types and so such a bijection obviously exists.[3]

Proposition 2. *[Halvorson, 2012] T and S have isomorphic theory categories.*

However, it will serve us well to dig a little further into the weeds here. For some motivation, the argument we have given thus far just shows that *there is* a bijection between their isomorphism types. It tells us nothing about what it is like. Given that we obtained it from crude cardinality facts, we've obtained our bijection using the Axiom of Choice. This may give some people reason to pause. Regardless, we can avoid Choice altogether and provide a very natural definition of such a bijection.[4]

To define this bijection, we change perspective and give a natural representation for the isomorphism types of T and S. Recall first that a model of T is of the form

$$\mathcal{A} = \langle \{a^{\mathcal{A}}\}, P_n^{\mathcal{A}}, a^{\mathcal{A}} \rangle_{n \in \omega}$$

where $\{a^{\mathcal{A}}\}$ is its domain and each $P_n^{\mathcal{A}} \subseteq \{a^{\mathcal{A}}\}$. It's not difficult to see that $\mathcal{A}$ can be represented[5] as a pair of an object $a^{\mathcal{A}}$ and a function $f_{\mathcal{A}} : \omega \to 2$ such that for all $n \in \omega$

$$a^{\mathcal{A}} \in P^{\mathcal{A}} \Leftrightarrow f_{\mathcal{A}}(n) = 1.$$

Now given we aren't interested so much in $\mathcal{A}$ as its isomorphism class, we can throw $a^{\mathcal{A}}$ away and use $f_{\mathcal{A}}$ for this representational purpose. To put it bluntly, each isomorphism type for models of T and S is essentially a function $f : \omega \to 2$.

This means we can represent (the isomorphism types of) these theories using subsets of 2^{ω}: the set of functions $f : \omega \to 2$. Recall that 2^{ω} is commonly known as *Cantor Space*, and

[2] See [Awodey, 2006] for a definition of a functor. It is essentially the obvious rendering of an order preserving map between categories.

[3] This is essentially the proof given in [Barrett and Halvorson, 2016].

[4] Similar work to these preliminaries, which goes into more detail, can be found in [Barrett and Halvorson, manuscript].

[5] That is, $\mathcal{A}$ and $\langle a^{\mathcal{A}}, f_{\mathcal{A}} \rangle$ can be simply defined from each other. Given the topic of this paper, one might worry whether or not they are "equivalent." This is not important in this argument, since the theory categories for T and S are unusual in the ways noted above.

following a convention in mathematical logic, I'm going to call elements of 2^ω reals. Thus, we let $X_T = 2^\omega$ (i.e., all of them); and

$$X_S = 2^\omega \backslash \{f \in 2^\omega \mid f(0) = 1 \wedge \exists n \in \omega\ f(n) = 0\}$$

(i.e., those reals that don't start with a 1 and have a 0 later on). We wish to define a bijection between X_T and X_S. First we note that we can easily define injections in either direction.

- Let $\rho : X_S \to X_T$ be the identity.
- Let $\sigma : X_T \to X_s$ be such that for all $f \in 2^\omega$ and for all $n \in \omega$

$$\sigma(f)(n) = \begin{cases} 0 & \text{if } n = 0 \\ f(n-1) & \text{otherwise} \end{cases}.$$

ρ works since its obvious that $X_S \subsetneq X_T$. To see how σ works we might be more succinct and clear by writing $\sigma(f) = 0^\frown f$ to indicate that we are just putting a 0 at the front of the f sequence. This is not a surjection since the constant function $f : \omega \to \{1\}$ is in X_S but not in the range of σ. Since we have injections in both directions but neither of them are surjections. Regardless, the Cantor-Bernstein theorem tells us that since we have these injections, there is a bijection $\pi : X_T \to X_S$. Moreover, the Axiom of Choice is not required. Now we've established categorical equivalence twice over, but there's still a little more depth to be plumbed: the definition of π is exceedingly simple. For $f \in 2^\omega$, we let

$$\pi(f) = \begin{cases} f & \text{if } \exists n \in \omega\ f = 0^{n\frown}\bar{1} \\ 0^\frown f & \text{otherwise} \end{cases}$$

where $0^n = \langle \underbrace{0, ..., 0}_{\text{n-timtes}} \rangle$ and $\bar{1} : \omega \to \{1\}$. Intuitively speaking, we check if f has a finitely many 0s following by 1s forever. If so, we leave it alone; if not we put a 0 in front of it.

I think it's worth pausing to remark on this observation. While it might see obvious that T and S are inequivalent, there is a way of relating them that is – conceptually speaking – very simple: the function between them can be described with a quick sentence. We noted above and we'll show below that T and S are not definitionally equivalent, but one might wonder if they are, in some sense, close to being definitionally equivalent. One way to see that they are is to move to a stronger logic. $\mathcal{L}_{\omega_1\omega}$ is the logic extending first order logic with the conjunctions and disjunctions of countable length that we denote by $\bigwedge$ and $\bigvee$ respectively. In this setting, we may define translations back and forth from $\mathcal{L}_T$ and $\mathcal{L}_S$ witnessing definitional equivalence. To illustrate this we just describe a translation t from $\mathcal{L}_S$ into $\mathcal{L}_T$ as it is most like the function π above. Given the atomic formula $Q_n x$ in $\mathcal{L}_S$ for $n > 0$ we translate that into $\mathcal{L}_T$ in the logic $\mathcal{L}_{\omega_1\omega}$ as

$$(\bigvee_{n\in\omega}(\bigwedge_{i\leq n}\neg P_n x \wedge \bigwedge_{i>n} P_n x) \to P_n x) \wedge$$
$$(\neg \bigvee_{n\in\omega}(\bigwedge_{i\leq n}\neg P_n x \wedge \bigwedge_{i>n} P_n x) \to P_{n-1} x).$$

The busy looking antecedent of the first conjunct is intended to mirror the first clause in the definition of $\pi(f)$; in the second conjunct, we use its negation to represent the "otherwise" clause. We leave the similar definition of Q_0 to the reader and note that b is clearly translated as a. The translation in the other direction is also left to the reader, however, it should be clear that in the stronger setting of $\mathcal{L}_{\omega_1\omega}$ it is not hard to make translations that mirror the bijection described above; and thus, in this infinitary setting, T and S are definitionally equivalent.

What should we make of this? I think it just gives us reason to take another look at these problems. While we do have something like definitional equivalence, we cannot – as finite beings – really work in the logic of infinite conjunctions and disjunctions. As such, it's not so obvious how we can really make use of this translation; and so we don't have a particularly compelling reason to take this result very seriously.[6] But one might wonder what we might see if we look a little closer.

2. A new proof

In this section, we revisit the claim that T and S are not definitionally equivalent through a topological lens. In the previous section, we saw that the isomorphism types of models of T and S can be helpfully represented by logicians' reals: i.e., functions $f : \omega \to 2$. This suggests that a topological perspective might be helpful.

2.1. **Propositional logic and topology.** Before we set up our topological apparatus, first note that the theories T and S, while articulated in first order logic have very natural counterparts in propositional logic.[7] In particular, we might articulate T in propositional logic as the empty propositional theory in the variables $\{p_n\}_{n\in\omega}$ and S as

$$q_0 \to q_n$$

for all $n \in \omega$ in the language with variables $\{q_n\}_{n\in\omega}$. With the clutter of quantification removed, we define a natural topology on 2^ω by using the following "cylinder" sets as a basis. For each pair s, t of finite subsets of ω with empty over lap, we let

$$C_{s,t} = \{f \in 2^\omega \mid \forall n \in s\ f(n) = 1 \wedge \forall n \in t\ f(n) = 0\}.$$

[6] With some hindsight, I think that the results of this paper suggest that definitional equivalences based on infinitary techniques deserve another look. We have developed a framework emerging out of Hudetz's program that can be used to comprehensively subsume such equivalences [Hudetz, 2017]. We save the exposition of that work for a future date.

[7] This was observed by Halvorson in [2012]. Indeed, that article begins with the propositional representation.

It can be seen that these sets form a clopen basis for a topology on 2^ω. Moreover and importantly for our little project, there is a propositional formula $\chi_{s,t}$ of $\{p_n\}_{n\in\omega}$ that corresponds naturally with $C_{s,t}$. More specifically, we let $\chi_{s,t}$ be

$$\bigwedge_{n\in s} p_n \wedge \bigwedge_{n\in t} \neg p_n.$$

A similar formula corresponding to $C_{s,t}$ obviously exists in $\{q_n\}_{n\in\omega}$. This gives us a connection between formulae of propositional logic and our topology. Now recall that every formula φ of propositional logic can be put into disjunctive normal form. This means that there is a finite sequence $\langle s_i, t_i\rangle_{i<n}$ of non-overlapping finite sets of natural numbers such that φ is logically equivalent to

$$\bigvee_{i<n} \chi_{s_i,t_i}.$$

In topological terms, this formula can then be represented by the set

$$\bigcup_{i<n} C_{s_i,t_i} \subseteq 2^\omega.$$

In other words, every formula of propositional logic is naturally represented by a finite union of clopen sets; i.e., it is a clopen subset of 2^ω. This provides the topological link that we require. Finally, we note that a theory in propositional logic based on countably many variables can consist of, at most, a countably infinity of sentences and so is naturally represented by an infinite intersection of clopen sets; i.e., a closed set.

2.2. **Topology recall**[8]**.** We now record three standard topological facts that drive our proof. Recall that a set of reals is perfect if it contains no isolated elements. The definition is similar in 2^ω. A set X is perfect in 2^ω if whenever $f \in X$ and $n \in \omega$, then there is some $g \neq f \in 2^\omega$ such that g and f are the same up to n.

Lemma 3. *Let $X \subseteq 2^\omega$. The following are equivalent:*[9]

(1) *X is perfect; and*
(2) *there is a continuous injection $f : 2^\omega \to 2^\omega$ such that $f[2^\omega] = X$.*

Informally speaking, this lemma allows us to see that any pair of perfect sets are linked by very nice bijections; i.e., they are continuous.

Lemma 4. *Let $f : X \to Y$ where Y is compact, Hausdorff. Then the following are equivalent:*[10]

[8] I'm particularly grateful to Elliot Glazer for his assistance with this section.

[9] This is Lemma 1.22 and Exercise 1.23 in David Marker's notes on Descriptive Set Theory. For a more robust citation source: (1→2) is essentially Corollary 1.A.3 from [Moschovakis, 1980]; and Theorem 6.2 from [Kechris, 2012]. (2→1) is relatively easy. Suppose X is not perfect and toward a contradiction that $f : 2^\omega \to 2^\omega$ is a continuous injection whose range is X. Then we may fix some $f(y)$ that is isolated in X and some $p \in 2^{<\omega}$ such that $f(y)$ is the only point extending p in X. But then there cannot be $q \in 2^{<\omega}$ such that x extends q and $f(z)$ extends p for every z extending p since then f could not be an injection.

[10] This is Exercise 8 in Section 26 of Chapter 5 in [Munkres, 2000].

(1) *f is continuous; and*
(2) *the graph of f is a closed set in $X \times Y$.*

This lemma tells us that the very nice maps have a simple topological characterization; i.e., they are closed sets.

Lemma 5. *If Z is a closed subset of $2^\omega \times 2^\omega$, there is a tree W on $2^{<\omega} \times 2^{<\omega}$ such that*[11] *for all $x, y \in 2^\omega$*

$$\langle x, y\rangle \in Z \;\Leftrightarrow\; \forall n \in \omega \; \langle x \restriction n, y \restriction n\rangle \in W.$$

This lemma tells that closed sets of Cantor space are naturally represented by trees. This will be useful since it is relatively easy to represent a tree using a theory in propositional logic.

2.3. **Implicit interpretation.** We are almost ready for the main theorem, but we still need a way of linking our topological results with definitional equivalence. To this end, it will be helpful to offer an alternative characterization of definitional equivalence that is more amenable to our proof strategy. In a nutshell, are going to make use of Beth's definability theorem to characterize definitional equivalence in such a way that (explicit) definitions are no longer required. The techniques of this section can be generalized to offer a theory of *implicit interpretation* that appears to be just a little stronger than Morita interpretation. We leave the exposition of that theory to a future date and just focus on definitional equivalence here.[12] The following generalization of Beth's definability theorem drives our alternative characterization.

Lemma 6. *[Andréka et al., 2022] Let T be a theory articulated in the $\mathcal{L} \cup \{R_i\}_{i\in I}$ where each R_i is a relation symbol. Now suppose that for any pair of models $\mathcal{M}, \mathcal{N}$ of T such that the reducts $\mathcal{M} \restriction \mathcal{L} = \mathcal{N} \restriction \mathcal{L}$,*[13] *we have $\mathcal{M} = \mathcal{N}$. Then each R_i is definable in T by a formula of $\mathcal{L}$; i.e., there is a formula $\psi_i(\bar{x})$ of $\mathcal{L}$ such that*

$$T \vdash \forall x(R_i\bar{x} \leftrightarrow \psi_i(\bar{x})).$$

The following lemma then provides the missing link for the main theorem. Intuitively, it tells us that we don't need to bother figuring out how the vocabulary of one theory is defined in another. We just need to find a suitable theory that extends them. For simplicity, we suppose that all our theories are articulated in languages that only use relation symbols.

Lemma 7. *Let A and B be theories articulated in the languages $\mathcal{L}_A$ and $\mathcal{L}_B$ respectively where $\mathcal{L}_A \cap \mathcal{L}_B = \emptyset$. Then A and B are definitionally equivalent*[14] *iff there is a theory U articulated in the language $\mathcal{L}_A \cup \mathcal{L}_B$ such that:*

(1) *every model of A, and respectively B, has a unique expansion to a model of U; and*
(2) *every model of U satisfies both A and B.*

[11] See Proposition 2.4 in [Kechris, 2012] or Proposition 2.3 in [Mansfield and Weitkamp, 1985].

[12] I'll note that this work is heavily inspired by the work of Thomas Barrett and Hajnal Andréka.

[13] Here $\mathcal{M} \restriction \mathcal{L}$ is the reduct of the model $\mathcal{M}$ of $\mathcal{L} \cup \{R_i\}_{i\in I}$ to $\mathcal{L}$.

[14] We use essentially the characterization of definitional equivalence from [Barrett and Halvorson, 2016].

($\rightarrow$) Suppose that A and B are definitionally equivalent. Then we may fix definitional extensions A^* and B^* of A and B respectively such that A^* and B^* are the same theory.[15] Let $U = A^* = B^*$. Given a model $\mathcal{M}$ of A there is clearly a unique model satisfying A^* since the new vocabulary was explicitly defined. A similar observation holds for models of B.

($\leftarrow$) Suppose we have a theory U articulated in $\mathcal{L}_A \cup \mathcal{L}_B$ as described above. Let $\mathcal{M}$ and $\mathcal{N}$ be models of U and suppose $\mathcal{M} \upharpoonright \mathcal{L}_A = \mathcal{N} \upharpoonright \mathcal{L}_A$. By the second part of our assumption, these are both models of A and thus, by the first part of our assumption, they have a unique expansion to a model of U. Thus, $\mathcal{M} = \mathcal{N}$. This means that the vocabulary of $\mathcal{L}_B$ is implicitly definable and so using Lemma 6 we see that the vocabulary of $\mathcal{L}_B$ is explicitly definable in U. Let A^* be the theory obtained by adding these explicit definitions to A. A similar argument show that the vocabulary of $\mathcal{L}_A$ is explicitly definable in U and so we may form B^* by adding these definitions to B. It should then be clear that $A^* = B^* = U$ as required.

2.4. **The theorem.**

Finally, we can deliver the main theorem of this paper.

Theorem 8. *Suppose that S^* is a theory articulated in $\mathcal{L}_S$. Then the following are equivalent:*

(1) *X_{S^*} is a perfect set; and*
(2) *T and S^* are definitionally equivalent.*

Proof. (1→2) By Lemma 3, we may fix a continuous injection $f : 2^\omega \rightarrow X_{S^*}$ such that $f[2^\omega] = X_{S^*}$. Moreover, by Lemma 2, we see that the graph of f is closed. By Lemma 5, we may then fix some tree W on $2^{<\omega} \times 2^{<\omega}$ such that

$$f(x) = y \Leftrightarrow \forall n \in \omega \ \langle x \upharpoonright n, y \upharpoonright n \rangle \in W.$$

Next we observe that this tree can be represented as a theory of propositional logic in the language with atoms $\{p_i, q_i\}_{i \in \omega}$. Given a pair $\langle t, s \rangle \in 2^{<\omega} \times 2^{<\omega}$ with $lh(t) = lh(s) = n$, we let $\psi_{t,s}$ be

$$\begin{gathered} \bigwedge \{p_i \mid i < n \wedge t(n) = 1\} \\ \wedge \bigwedge \{\neg p_i \mid i < n \wedge t(n) = 0\} \\ \bigwedge \{q_i \mid i < n \wedge s(n) = 1\} \\ \wedge \bigwedge \{\neg q_i \mid i < n \wedge s(n) = 0\}. \end{gathered}$$

This gives us a way of representing a potential element $\langle t, s \rangle$ of the tree W in propositional logic. We then put this together by letting U^* be the propositional theory whose axioms are $\psi_{t,s}$ where $\langle t, s \rangle \in W$. And finally, we let U be the obvious first order theory in $\mathcal{L}_T \cup \mathcal{L}_S$ that corresponds to U^*. Given a model $\mathcal{M}$ of U, let x_T be the characteristic function of the set $\{n \in \omega \mid \mathcal{M} \models P_n a\}$ and x_S be the characteristic function for the set $\{n \in \omega \mid \mathcal{M} \models Q_n b\}$. It

[15] Recall that the standard definition of a theory is a set of sentences closed under consequence.

can then be seen that for all models $\mathcal{M}$ of $\mathcal{L}_T \cup \mathcal{L}_S$

$$\mathcal{M} \models U \ \Leftrightarrow f(x_T) = x_S.$$

From here, it is easy to see that whenever $\mathcal{M}, \mathcal{N}$ satisfy U, the clauses of Lemma 7 are satisfied and so T and S are definitionally equivalent.

(2→1) Since T and S^* are definitionally equivalent, we may use Lemma 7 to fix a theory U in $\mathcal{L}_T \cup \mathcal{L}_S$ satisfying conditions (1) and (2) described there. Recalling that $X_T = 2^\omega$, it can then be seen from this that there is a function $f : 2^\omega \to 2^\omega$ such that $f[2^\omega] = X_S$. Moreover, since f is described by what is effectively a theory in propositional logic, we see that its graph must be closed. But then Lemma 2 tells us that f is continuous; and finally Lemma 3 tells us that X_S is perfect. □

The fact that Halvorson's T and S are not definitionally equivalent is essentially a corollary of our theorem.

Corollary 9. *T and S are not definitionally equivalent.*

Proof. X_S is not a perfect set, since the sequence $\langle 1, 1, ... \rangle$ is isolated in X_S. □

But as promised, our theorem also delivers some surprising equivalences between theories similar to those in the Halvorson example. Very generally, we have the following:

Corollary 10. *Suppose A and B are (consistent) theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ that both include the statement that there is exactly one object. Then suppose that the set of reals X_A and X_B associated with A and B are both perfect. Then A and B are definitionally equivalent.*

Proof. Using Theorem 8, we see that A and B are both definitionally equivalent to T and so they are definitionally equivalent to each other. □

For a more specific and perhaps more striking example we have the following:

Proposition 11. *If A and B are finite, consistent theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ that both say there is exactly one object, then they are definitionally equivalent.*

Proof. Since A and B are both finite, they are both represented by finite theories in propositional logic; i.e., sentences of propositional logic. Since each sentence of propositional logic can be converted into disjunctive normal form, we see that each of these sets is a finite union of clopen basis elements. It is then easy to see that such sets have no isolated elements. And since they are perfect, we see by Corollary 10 that A and B are definitionally equivalent. □

In other words, this tells us that any pair of finite theories in propositional logic can be understood as being definitionally equivalent.

3. Discussion

What should we make of this? I'd like to offer a contrived example that I believe will help us better understand the odd tension between mathematics and intuition that the Halvorson example seems to provoke. To this end, suppose that every year at the end of our harvest, we make an offering to our gods with our finest produce and afterwards celebrate with a massive feast. Everyone sleeps long and deep after the feast and each year, we wake to find that the offering has either been taken or left in its place. Naturally, this leads our community to questions and speculation. What are the gods trying to tell us? What will happen next year? And the year after that? Two members of our community, Terry and Shirley, take it upon themselves to prophesize the future plans of the gods with regard to our offerings. Let's take it that Terry's prophesy is empty. He makes no commitments at all and just says that anything can happen. Shirley, on the other hand, does make a little conditional prophesy. She proposes that if the gods take their offering this year, then they'll continue taking the offering forever after. It's not difficult to see that this little story corresponds with the theories of the Halvorson example. Terry can be understood as offering the theory, T, while Shirley offers S.

Now it seems obvious that Terry and Shirley are offering two noticeably different theories about the future behavior of our gods. Moreover, the fact that T and S are not definitionally equivalent seems to support this intuition. But let's complicate the picture and suppose that another member of our community, Botao, offers a new prophesy which says that for the next one hundred years the gods will not take our offering. This clearly corresponds to a theory B in $\mathcal{L}_S$ which says that there is only one object, b, and that $\neg Q_n b$ holds for all $n < 100$. Now it also seems obvious that Terry and Botao are offering different prophesies. Terry isn't saying anything and Botao is committing to a sequence of events that extend beyond almost all of our lives. Nonetheless, Corollary 10 tells us that the associated theories T and B are definitionally equivalent. Indeed in this particular case, it is not difficult to see that the required translations can be procured without recourse to the very general machinery described above.

So what is going on? Given the clash between formalism and intuition, it seems obvious that we should blame definitional equivalence. Just as we took the equivalence of the categories associated with T and S as reason to be suspicious of categorical equivalence, we might take the definitional equivalence T and B as reason to be suspicious of definitional equivalence. There is something to this line of thought, but without some care we can easily take it too far. The key point is that there is nothing wrong with definitional equivalence, but it is very easy to be misled into thinking definitional equivalence tells us more than it actually does.

Returning to Terry and Botao, I think it is obvious that our intuitions are correct. Terry and Botao are not offering the same prophesy. And yet, it is also obvious that their prophesies are definitionally equivalent. However, the main driver behind our intuition is the fact that Terry and Botao are offering theories in an *interpreted language*. The words they are uttering already have a specific predetermined meaning. When Botao says that the gods will take the offering next year, his theory, B, will proven wrong if they do not, while Terry's empty prophesy, T, will remain intact. The very fact that B could be wrong while T is right means that B and T are not different versions the same theory.[16] The key to the puzzle then is the fact that definitional equivalence has absolutely nothing to say about interpreted languages. Definitional equivalence speaks about T and B as purely syntactic items: strings of symbols with no antecedent meaning. Put like that, one might be tempted think that definitional equivalence is just some mathematical artifact with little or no real world significance. This, I think, would be a mistake. The problem is not that definitional equivalence is irrelevant, it is that we've been tempted to squeeze more out of the mathematics than was there in the first place. We need to offer a better explanation of what it means when two theories are definitionally equivalent.

Here is what I think a serviceable informal explanation of definitional equivalence might look like. When theories U and W are definitionally equivalent, we learn – not that they are different versions of the same theory – but rather that: through translation we could use theory U (understood syntactically) to do whatever representational work for which we might use W; and similarly, we might use W to, so to speak, stand in for U. Of course, there's a lot to unpack, but this is merely intended as a helpful gloss. Now it could happen that we were already using U and W for some real world purpose. As such, we should regard the languages $\mathcal{L}_U$ and $\mathcal{L}_W$ of U and W as interpreted. But there is absolutely no guarantee that the translations used in establishing the definitional equivalence will respect the meanings of those interpreted languages. Indeed, this is exactly what happens in the comparison of Terry and Botao's prophesies. While T and B are definitionally equivalent, the translations that witness this do not preserve the meaning of the expressions in the underlying interpreted languages. Indeed in this case, any (non-trivial) translation will break the underlying interpretations, since Terry and Botao are speaking essentially the same interpreted language.[17] So while it is true that you could use T (understood syntactically) to do the representational work done by B, the translation used to achieve this will change the meaning of T.[18]

Where does this leave us? I don't think we've found any reason to doubt the value of definitional equivalence as an instrument for understanding when two theories are equivalent. That said, we did show that definitional equivalence does not deliver a simple sufficient condition for saying that two theories are merely different versions of each other. Sometimes we need to

[16] Note that we can make an analogous observation distinguishing T and S. If the gods take our offerings this year and then fail to take them in some later year, Shirley's prophesy will have been refuted, while Terry's remains vacuously correct. Thus again, we have good reason to think that T and S are not the same theory.

[17] Strictly, we should say there is a fixed translation between the languages that preserves the underlying interpretation of the vocabularies of $\mathcal{L}_T$ and $\mathcal{L}_S$. More specifically, a and b translate to each other, and so do P_n and Q_n for all $n \in \omega$.

[18] Note that an analogous observation applies to the comparison of T and S. Categorical equivalence is not so helpful here, but our earlier observation that T and S are definitionally equivalent in the infinitary logic, $\mathcal{L}_{\omega_1\omega}$, tells us that the infinitary translations witnessing this cannot preserve the meanings of the words used by Terry and Shirley.

be attentive to the question of whether a theory is articulated in an interpreted language.[19] Given that definitional equivalence, properly understood, has been saved from the chopping block, this gives us some reason to revisit categorical equivalence. As we noted at the beginning of this paper, the fact that T and S have equivalent theory categories is often thought of as a mark against categorical equivalence as an indicator of genuine equivalence. I think the examples above indicate that categorical equivalence also deserves something of a reprieve. Like definitional equivalence, categorical equivalence has nothing to say about interpreted theories. Theory categories are also defined with a syntactic understanding of a theory with no attention paid to predetermined interpretations of the underlying language of the theory. As such and with the help of a couple of footnotes above, we see that the problem presented by T and S is analogous to that of T and B. So what then is a better explanation of what it means to say two theories have equivalent categories? I would contend that the explanation is much the same as that we gave for definitional equivalence. To draw out the difference, we'd need to dig deeper into details. But the key point for this paper, is that the fact that T and S are categorically equivalent when T and S are intuitively inequivalent is better explained by our failures to observe when we are thinking about interpreted languages than it is by fact that definitional equivalence is more difficult to obtain than categorical equivalence.

So what's the answer then? Are T and S equivalent or not? I'm inclined to say that the question is somewhat under-determined. If we are thinking of T and S as purely syntactic theories, then I think their categorical equivalence tells that what we might call their *representational capacities* are, at least to some degree, equivalent. But if we were thinking of T and S as theories articulated in interpreted languages, as in the discussion above, then they may be obviously inequivalent. Given that the Halvorson example is usually presented with T and S as theories in uninterpreted languages, it seems perhaps a little more natural to say that they *are* equivalent. Since we started with pure theories, this seems like a reasonable answer. Of course, the value of this response depends on what we want to do with T and S. As we've seen above, this answer may not be that helpful if we are considering theories in languages whose terms have predetermined meanings.

As a final remark, I'd like to note that although definitional equivalence is unable to detect the effects of interpreted languages on theories, this should not be taken as evidence that formal methods are not useful here. The problem is rather that the standard apparatus of translation and interpretation was not designed to take such issues into account. We leave the generalization and development of a framework that is sensitive to the predetermined meaning of its vocabulary as a challenge and target for future work.

[19] I think it is plausible that we generally don't need to be sensitive to this with mathematical theories, but we leave this discussion for another article.

References

H. Andréka, J. Madarász, I. Németi, and G. Székely. Testing definitional equivalence of theories via automorphism groups, 2022.

Steve Awodey. *Category Theory.* Clarendon Press, Oxford, 2006.

Thomas William Barrett and Hans Halvorson. Morita equivalencce. *The Review of Symbolic Logic*, 9(3):556–582, 2016.

Thomas William Barrett and Hans Halvorson. Extension, translation, and the cantor-bernstein property. manuscript.

Hans Halvorson. What scientific theories could not be. *Philosophy of Science*, 79(2):183–206, 2012. doi: 10.1086/664745.

Laurenz Hudetz. The semantic view of theories and higher-order languages. *Synthese*, 196 (3):1131–1149, 2017.

Laurenz Hudetz. Definable categorical equivalence. *Philosophy of Science*, 86(1):47–75, 2019.

A. Kechris. *Classical Descriptive Set Theory.* Graduate Texts in Mathematics. Springer New York, 2012.

R. Mansfield and G. Weitkamp. *Recursive aspects of descriptive set theory.* Oxford logic guides. Oxford University Press, 1985.

Eleanor March. On some examples from first-order logic as motivation for categorical equivalence of kpms, 2024.

Y.N. Moschovakis. *Descriptive Set Theory.* North Holland, 1980.

J.R. Munkres. *Topology.* Featured Titles for Topology. Prentice Hall, Incorporated, 2000.

James Owen Weatherall. Why not categorical equivalence? In *Hajnal Andréka and István Németi on Unity of Science: From Computing to Relativity Theory Through Algebraic Logic*, pages 427–451. Springer Verlag, 2021.