The Review of Symbolic Logic
Volume 17, Number 3, September 2024

BEYOND LINGUISTIC INTERPRETATION IN THEORY COMPARISON

TOBY MEADOWS
University of California, Irvine

Abstract. This paper assembles a unifying framework encompassing a wide variety of mathematical instruments used to compare different theories. The main theme will be the idea that theory comparison techniques are most easily grasped and organized through the lens of category theory. The paper develops a table of different equivalence relations between theories and then answers many of the questions about how those equivalence relations are themselves related to each other. We show that Morita equivalence fits into this framework and provide answers to questions left open in Barrett and Halvorson [4]. We conclude by setting up a diagram of known relationships and leave open some questions for future work.

§1. Introduction.

> *What's in a name? That which we call a rose*
> *By any other name would smell as sweet;*

Relative interpretation is a powerful tool that allows us to compare different theories using translations. For example, it provides us with reasonable criteria for deciding whether two theories are for some intents and purposes the same. Nonetheless, it is hampered by a number of seemingly draconian limitations. This is perhaps mostly keenly felt in the philosophy of science where theories in physics are generally not articulated using first-order logic, but rather considered as collections of mathematical structures.[1] This has recently led some philosophers of science to turn to more liberal techniques from category theory that—among other things—make no assumption that theories are axiomatizable in first-order logic [11, 24]. Others have thought that these techniques are too liberal and have tried to develop the middle ground [4, 13].

In this paper, we aim to provide a means for organizing these instruments of comparison that arguably turns the traditional picture on its head. Rather than seeing category theoretic techniques as providing a novel and less constrained tool than interpretability that avoids logic; we shall demonstrate that categorical equivalence relations can be understood as the core notion from which many others can be obtained via natural restrictions. Beyond providing a transparent way of understanding the relationships between various techniques, it also provides helpful insights into how these restrictions play out and whether they are appropriate in various contexts. I also

Received: August 3, 2022.
2020 *Mathematics Subject Classification*: Primary 03A05.
Key words and phrases: theoretical equivalence, category theory, set theory, model theory.

[1] See [23] or [10] for good discussions of this issue.

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Association for Symbolic Logic.

doi:10.1017/S1755020323000321

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

believe that the underlying picture provides a helpful guide to those who would like to develop new techniques for theory comparison.

The paper is organized as follows: In Section 2, we describe the category theoretic tools that will be used and the main kind of category considered in this paper: theory categories. In Section 3, we review the elementary theory of relative interpretability and then demonstrate that some common notions of equivalence from relative interpretation are, in fact, instances of categorical equivalence relations in natural restrictions of theory categories. In Sections 4 and 5, we extend these results to accommodate generalizations of interpretability in multi-sorted contexts where new domains can be added to a language. This then brings us to Section 6, where we provide a diagram organizing the content of the paper. The strategy conscious reader may find it helpful to flip ahead to this diagram first in order to get an idea of the lie of the land. This section will also provide some solutions to questions left open in [4] and then finally use the diagram to pose some new questions for future work.

§2. Theory categories.

In this section, we shall recall some elementary definitions that will be used throughout this paper. In particular, we discuss three notions of equivalence between categories. Two of these are standard, while the third is—I believe—novel.

2.1. Category theoretic preliminaries.

We restrict our attention to categories that are small; i.e., where the objects and arrows of the category form a set. We assume that the reader is familiar with the basic axioms of category theory and the definitions of functors and natural transformations.[2]

For our *first equivalence* relation, recall that categories $\mathcal{C}$ and $\mathcal{D}$ are *isomorphic* if there exist functors $F : \mathcal{C} \Rightarrow \mathcal{D}$ and $G : \mathcal{D} \Rightarrow \mathcal{C}$ such that $G \circ F = Id_{\mathcal{C}}$ and $F \circ G = Id_{\mathcal{D}}$ where $Id_{\mathcal{C}}$ and $Id_{\mathcal{D}}$ are the respective identity functors for $\mathcal{C}$ and $\mathcal{D}$. In other words, F and G are such that:

- for all objects A from $\mathcal{C}$, $G(F(A)) = A$;
- for all arrows $h : A \to B$ from $\mathcal{C}$, $G(F(h)) = h$;
- for all objects B from $\mathcal{D}$, $F(G(A)) = A$; and
- for all arrows $i : C \to E$ from $\mathcal{D}$, $F(G(i)) = i$.

Intuitively speaking, the functors take us forth and back returning us to exactly where we started. For our *second equivalence* relation, we first recall the following definition.

Definition 2.1. *A natural transformation η between functors $F, G : \mathcal{C} \Rightarrow \mathcal{D}$ is a* natural isomorphism *if for all objects A in $\mathcal{C}$*

$$\eta_A : F(A) \cong G(A).$$

Let us write $\eta_{\cdot} : F \cong_{nat} G$ to indicate this.

Let $Id_{\mathcal{C}}$ and $Id_{\mathcal{D}}$ be the identity functors for $\mathcal{C}$ and $\mathcal{D}$ respectively. Then recall that categories $\mathcal{C}$ and $\mathcal{D}$ are *equivalent*[3] if there are functors $F : \mathcal{C} \Rightarrow \mathcal{D}$ and $G : \mathcal{D} \Rightarrow \mathcal{C}$ such that $Id_{\mathcal{C}}$ and $G \circ F$ are naturally isomorphic and so are $Id_{\mathcal{D}}$ and $F \circ G$. In other

[2] For details, the reader may wish to consult Section 7.8 of [3].

[3] In contexts where confusion may arise, I will refer to this relation as *categorical equivalence* or *equivalence of categories*.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

words, there is a natural transformation $\eta.$ from $Id_{\mathcal{C}}$ to $G \circ F$ such that for all objects A from $\mathcal{C}$

$$\eta_A : A \cong G(F(A))$$

and a natural transformation $\nu.$ from $Id_{\mathcal{D}}$ to $F \circ G$ such that for all objects B from $\mathcal{D}$

$$\nu_D : B \cong F(G(B)).$$

Or more compactly, $\eta. : Id_{\mathcal{C}} \cong_{nat} G \circ F$ and $\nu. : Id_{\mathcal{D}} \cong_{nat} F \circ G$.

We see that equivalence is a natural weakening of isomorphism in that the functors take us back and forth to an object that is isomorphic to the one we started with but not necessarily identical to it. The fact that the isomorphisms $\eta.$ and $\nu.$ are *natural* ensures a kind of uniformity in the relationship between the isomorphism occurring at the object level. This provides our second equivalence relation. In contrast, our novel, third equivalence relation will abandon this uniformity requirement. We'll describe this in a moment, but first we note a convenient alternative definition of equivalence between categories that will be useful later.

PROPOSITION 2.2. *Categories $\mathcal{C}$ and $\mathcal{D}$ are equivalent if and only if there is a functor $F : \mathcal{C} \Rightarrow \mathcal{D}$ that is full, faithful and essentially surjective.*[4]

For our *third equivalence relation*, we move off the main menu. Informally speaking, my goal is to weaken categorical equivalence in such a way that we go back and forth to an isomorphic object, but we do not demand that the isomorphism is natural or uniform. I note—up front—that in most categorical settings this equivalence relation will be too weak to be of much significance. However, in the setting of categories based on theories it has a natural counterpart in relative interpretability. This is our reason for introducing it here.[5]

DEFINITION 2.3. *We say that categories $\mathcal{C}$ and $\mathcal{D}$ are* objectively equivalent *if there are functors $F : \mathcal{C} \Rightarrow \mathcal{D}$ and $G : \mathcal{D} \Rightarrow \mathcal{C}$ such that:*

(1) $A \cong G(F(A))$ *for all objects A from $\mathcal{C}$; and*
(2) $B \cong F(G(B))$ *for all objects B from $\mathcal{D}$.*

Since functors preserve isomorphism, it is easy to see that this is an equivalence relation on categories. It is similar to a categorical equivalence in that we go back and forth between categories and return to an isomorphic object. But in contrast, we make no demands that this isomorphism be natural: it need not be distributed uniformly over the objects of the category by a natural transformation. Moreover, this equivalence places no constraints at all on what happens to the arrows. This is why I've called it objective.

In the fully general setting where arbitrary categories are permitted, this equivalence is extremely weak. For an example, suppose $\mathcal{C}$ is a category with one object A and just its identity arrow id_A. Then suppose $\mathcal{D}$ is another category with just one object B, but there are two arrows, id_B and $f : B \to B$. Suppose that $f \circ f = id_B$, although the example also works if we suppose $f \circ f = f$. It is the not difficult to see that $\mathcal{C}$ and $\mathcal{D}$ are neither isomorphic nor equivalent as categories. However, they are objectively equivalent. Let

[4] See Proposition 7.25 of [3].

[5] I'm extremely grateful to an anonymous referee for spotting a grievous error in my initial approach to this equivalence relation. Their helpful comments and suggestions have allowed me to simplify things greatly.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

$F : \mathcal{C} \Rightarrow \mathcal{D}$ be such that $F(A) = B$ and $F(id_A) = id_B$; and let $G : \mathcal{D} \Rightarrow \mathcal{C}$ be such that $G(B) = A$, $G(id_B) = id_A$ and $G(f) = id_A$. It can then be seen that Definition 2.3 is satisfied. Of course, this is not a very interesting example. Nonetheless, it should be clear that the sense in which these categories are equivalent is particularly weak. In the next sections, we shall see that in the context of theory categories constrained using definability, this equivalence becomes more interesting.

To close the current section we observe that—in general—these equivalence relations are organized in a strict hierarchy:

PROPOSITION 2.4. *Let $\mathcal{C}$ and $\mathcal{D}$ be categories*:

(1) *If $\mathcal{C}$ and $\mathcal{D}$ are isomorphic, then they are equivalent.*
(2) *If $\mathcal{C}$ and $\mathcal{D}$ are equivalent, then they are objectively equivalent.*

Moreover, there are equivalent categories that are not isomorphic; and objectively equivalent categories that are not equivalent.

Proof. We leave the proof of (1) and (2) to the reader since they essentially follow by definition. For a relatively concrete example of equivalent categories that are not isomorphic, see Awodey's discussion of the categories of finite ordinals and hereditarily finite sets in Section 7.8 of [3]. We provided a pair of categories that are objectively equivalent but not categorically equivalent above and will provide another example in Proposition 2.7. □

2.2. Categories of theories. In the previous section, we described three equivalence relations between categories: isomorphism; equivalence; and objective equivalence. We now aim to apply these tools to makes comparisons between theories. As such, we define a natural category for the purposes of representing theories. To facilitate this, we assume for the rest of the paper that there is an inaccessible cardinal and that Ω is the least of them. This will allow us to treat our theory categories as small when their most obvious definition does not. Thus, unless otherwise stated our background theory for the rest of this paper will be ZFC plus the assumption that an inaccessible cardinal exists.[6]

DEFINITION 2.5. *Let T be a theory. Then $mod(T)$ is the* theory category *consisting of*:

- *objects: models $\mathcal{M}$ of theories T from V_Ω;*[7] *and*
- *arrows: elementary embeddings between those models.*

Given a semantic outlook, the choice of objects for this category is—more or less—forced upon us, but with regard to arrows more options are available. For example one might also consider using: homomorphisms; embeddings or isomorphisms. Some helpful discussion around this is provided in [4], however, for our purposes elementary embeddings provide a very natural fit with relative interpretability. We also note that at various times in the paper, we shall have cause to restrict our attention to various subcategories of $mod(T)$ as in some contexts it will be appropriate to consider particular kinds of models or embeddings. We now recall our equivalence relations as they are instantiated between theory categories.

[6] This approach will not be to everyone's taste and for this reason I will discuss this further in Section 2.2.2.

[7] This is the Ωth level of the cumulative hierarchy of sets.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Definition 2.6. *Suppose there are functors $t : mod(T) \to mod(S)$ and $s : mod(S) \to mod(T)$.*[8] *We say that* t *and* s witness *that $mod(T)$ and $mod(S)$ are*:

- isomorphic *if $s \circ t(\mathcal{M}) = \mathcal{M}$ and $t \circ s(\mathcal{N}) = \mathcal{N}$ when $\mathcal{M} \models T$ and $\mathcal{N} \models S$; and $s \circ t(f) = f$ and $t \circ s(g)$ when* f *and* g *are elementary embeddings between models of T and S respectively*;
- equivalent *if $s \circ t \cong_{nat} 1_{mod(T)}$ and $t \circ s \cong_{nat} 1_{mod(S)}$; and*
- objectively equivalent *if $s \circ t(\mathcal{M}) \cong \mathcal{M}$ and $t \circ s(\mathcal{M}) \cong \mathcal{N}$ when $\mathcal{M} \models T$ and $\mathcal{N} \models S$.*

We say that $mod(T)$ and $mod(S)$ are isomorphic (*respectively*, equivalent *and* objectively equivalent) *if there are pair of functors witnessing this. Below we'll frequently say T and S are, for example, equivalent rather than saying that $mod(T)$ and $mod(S)$ are equivalent.*

2.2.1. Hierarchy? It is then natural to ask whether the hierarchy described in Proposition 2.4 remains strict in context of theory categories. We now show that there are objectively equivalent theories that are not equivalent, but every pair of equivalent theories turns out to be isomorphic.

Proposition 2.7. *There is a pair of theories that are objectively equivalent but not equivalent as categories.*

Proof. Let T be the theory that says there are exactly two objects; and let S be the theory that says that there is exactly one object. To see that they are objectively equivalent, we describe functors $F : mod(T) \Leftrightarrow mod(S) : G$ satisfying Definition 2.3. Since both T and S have exactly one isomorphism class, we may let the action of F on the objects of $\mathcal{C}$ be given by an arbitrary bijection between the models of T and S. Note that for any models $\mathcal{M}$ and $\mathcal{N}$ of S, there is exactly one elementary embedding between them. Thus, given $f : \mathcal{M} \to \mathcal{N}$ in $mod(T)$, we let $F(f)$ be the unique arrow between $F(\mathcal{M})$ and $F(\mathcal{N})$. In the other direction, we let $G(\mathcal{A}) = F^{-1}(\mathcal{A})$ for models $\mathcal{A}$ of S. For arrows, we need to make some decisions. For any arrow $g : \mathcal{A} \to \mathcal{B}$ in $mod(S)$, there will be two arrows between $G(\mathcal{A})$ and $G(\mathcal{B})$ in $mod(T)$. To address this, we simply pick an arrow and its inverse between any two models of T. We then let $G(g)$ be that arrow; and for the arrow $g^{-1} : \mathcal{B} \to \mathcal{A}$, we let $G(g^{-1})$ be its inverse. This makes sense since every elementary embedding between finite models is an isomorphism and thus has an inverse. It can then be seen that F and G are functors that witness an objective equivalence. We leave it to the reader to verify that $mod(T)$ and $mod(S)$ are not equivalent as categories. □

We now show that equivalence between theory categories implies isomorphism. To see this, first observe that in any theory category $mod(T)$ the cardinality of any isomorphism class is Ω. More precisely, for any object $\mathcal{M}$ from $mod(T)$

$$|[M]| = \Omega.$$

To see this, consider, for example, the category of models of Peano arithmetic, $mod(PA)$. It is easy to see that for every ordinal $\alpha < \Omega$, there will be a model of PA

[8] Note that there is no requirement that t or s be in any way definable or determined by translations.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

that includes α in its domain. This means that following lemma suffices to prove this theorem.

Lemma 2.8. *Suppose $\mathcal{C}$ and $\mathcal{D}$ are equivalent as categories and suppose that each of the isomorphism classes of these categories has the same cardinality; i.e., for all objects A from $\mathcal{C}$ and B from $\mathcal{D}$*

$$|[A]| = |[B]|.$$

Then $\mathcal{C}$ and $\mathcal{D}$ are isomorphic.

Proof. Suppose $F : \mathcal{C} \to \mathcal{D}$ is a functor witnessing the categorical equivalence; i.e., F is full, faithful and essentially surjective. We shall use F to define an isomorphism $H : \mathcal{C} \cong \mathcal{D}$. To do this we shall pick a distinguished element from each of the isomorphism classes of $\mathcal{C}$ use the action of F on these elements to generate H's action on the rest of $\mathcal{C}$.

We start by defining two endo-functors and natural transformations. First let $\cdot^* : \mathcal{C} \to \mathcal{C}$ be an endo-functor that takes objects A from $\mathcal{C}$ and returns a distinguished element A^* of $[A]$ such that for all objects A_0, A_1 from $\mathcal{C}$, $A_0^* = A_1^*$ whenever $A_0 \cong A_1$. Then let $\pi_A : A \cong A^*$ be an isomorphism such that $\pi_{A^*} = id_{A^*}$.[9] For arrows $f : A \to B$ from $\mathcal{C}$, let $f^* : A^* \to B^*$ be

$$\pi_B \circ f \circ \pi_A^{-1}.$$

Note that $\pi.$ is a natural isomorphism from the identity functor on $\mathcal{C}$ to $\cdot^*$.

Next we define a similar endo-functor $\cdot^\dagger : \mathcal{D} \to \mathcal{D}$ that is tailored to match up with $\cdot^*$. Note that since F establishes categorical equivalence it is essentially surjective and thus for any object E from $\mathcal{D}$ we may fix an object A from $\mathcal{C}$ such that $F(A) \cong E$. So for all objects E from $\mathcal{D}$, let $E^\dagger = F(A^*)$ where A is some object from $\mathcal{C}$ such that $F(A) \cong E$.[10] Then observe that

$$E^\dagger = F(A^*) \cong F(A) \cong E$$

so $E^\dagger \in [E]$. For each object E from $\mathcal{D}$, let $\rho_E : E \cong E^\dagger$.[11] For an arrow $g : E \to D$ from $\mathcal{D}$, let $g^\dagger : E^\dagger \to D^\dagger$ be

$$\rho_D \circ g \circ \rho_E^{-1}.$$

Note that $\rho.$ is a natural isomorphism between the identity functor on $\mathcal{D}$ and $\cdot^\dagger$.

Our goal now is to define H using $\cdot^*, \cdot^\dagger, \pi.$ and $\rho.$. For all A^* from $\mathcal{C}$, let $H_{A^*} : [A^*] \to [F(A^*)]$ be a bijection such that $H_{A^*}$and F agree on A^*; i.e., $H_{A^*}(A^*) = F(A^*)$. This makes sense given our assumptions about the cardinality of isomorphism classes in $\mathcal{C}$ and $\mathcal{D}$. Now let $H : \mathcal{C} \to \mathcal{D}$ be defined as follows. For objects A from $\mathcal{C}$, we let

$$H(A) = H_{A^*}(A).$$

For arrows $f : A \to B$ from $\mathcal{C}$, let $H(f) : H(A) \to H(B)$ be

$$\rho_{H(B)}^{-1} \circ F(f^*) \circ \rho_{H(A)}.$$

[9] It may be worth noting that the axiom of choice is used here.

[10] Choice is also used here.

[11] And choice is used here.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

To see that this makes sense, the following diagram is helpful.

$$\begin{tikzcd}
B \arrow[r, "\pi_B"] & B^* \arrow[r, mapsto, "F"] & F(B^*) & H(B) \arrow[l, "\rho_{H(B)}"'] \\
A \arrow[u, "f"] \arrow[r, "\pi_A"'] & A^* \arrow[u, "f^*"] \arrow[r, mapsto, "F"] & F(A^*) \arrow[u, "F(f^*)"] & H(A) \arrow[l, "\rho_{H(A)}"] \arrow[u, "H(f)"]
\end{tikzcd}$$

Now we claim that H is a functor and that H witnesses that $\mathcal{C}$ and $\mathcal{D}$ are isomorphic as categories. We leave the reader to verify that H is a functor.

Claim. *H establishes an isomorphism between $\mathcal{C}$ and $\mathcal{D}$.* □

Proof. By design we know H is a bijection on objects, so it suffices to show that H is full and faithful.

To see that H is full, suppose $g : H(A) \to H(B)$ is an arrow from $\mathcal{D}$. It will suffice to show there is some $f : A \to B$ such that $H(f) = g$. To see this first observe that $\rho_{H(B)} \circ g \circ \rho^{-1}_{H(A)}$ is an arrow from $H(A^*)$ to $H(B^*)$. Then since $H(A^*) = F(A^*)$ and $H(B^*) = F(B^*)$ and F is full, we may fix some arrow $i : A^* \to B^*$ from $\mathcal{C}$ such that $F(i) = \rho_{H(B)} \circ g \circ \rho^{-1}_{H(A)}$. We then let $f = \pi^{-1}_B \circ i \circ \pi_A$ and observe that this is an arrow from A to B. We then see chase the diagram observing that

$$\begin{aligned}
H(f) &= \rho^{-1}_{H(B)} \circ F(f^*) \circ \rho_{H(A)} \\
&= \rho^{-1}_{H(B)} \circ F(\pi_B \circ f \circ \pi^{-1}_A) \circ \rho_{H(A)} \\
&= \rho^{-1}_{H(B)} \circ F(i) \circ \rho_{H(A)} \\
&= \rho^{-1}_{H(B)} \circ (\rho_{H(B)} \circ g \circ \rho^{-1}_{H(A)}) \circ \rho_{H(A)} = g.
\end{aligned}$$

To see that H is faithful, suppose $f, g : A \to B$ are such that $H(f) = H(g)$. We claim that $f = g$. To see this first recall that

$$H(f) = \rho^{-1}_{H(B)} \circ F(f^*) \circ \rho_{H(A)} \quad \& \quad H(g) = \rho^{-1}_{H(B)} \circ F(g^*) \circ \rho_{H(A)}$$

and so we see that $F(f^*) = F(g^*)$. And since F is faithful, we see that $f^* = g^*$. Then recall that

$$f^* = \pi_B \circ f \circ \pi^{-1}_A \quad \& \quad g^* = \pi_B \circ g \circ \pi^{-1}_A$$

and so $f = g$ as required. □

Our main claim then follows directly.

Corollary 2.9. *If mod(T) is equivalent to mod(S), then they are also isomorphic.*

Thus, we see that when we restrict our attention to theory categories, the hierarchy we had in Proposition 2.4 is no longer strict. This may cause us to question the value separating them in our framework. I have a couple of things to say about this. First a small point: it is sometimes said in defense of categorical equivalence that it is superior to isomorphism in that it wipes away detail that is not pertinent when it comes to theory comparison. The corollary above tells us that, when it comes to theories, we are—in fact—concerned with isomorphism because it is the same thing as equivalence. Second and more substantively, we shall see in Section 3 that there are natural restrictions of this framework in which this hierarchy question becomes more difficult and interesting.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

2.2.2. The inaccessible cardinal. To conclude this section, we revisit the issue of our assumed inaccessible cardinal Ω. We took this up as it provides a simple way to stay within a standard set theoretic background theory like ZFC while avoiding issues around the size of theory categories. If we had not made such an assumption and instead took the objects of $mod(T)$ to be the all of the models satisfying T, then $mod(T)$ could not be a set and thus, a theory like ZFC would not suffice for our work here.[12] By taking all of the models and elementary embeddings from V_Ω we are able to ensure that theory categories, the functors between them and the natural transformations between them are all sets and thus gloss over this issue. In a moment, we shall note some problems with this approach, but first we observe that this kind of solution emerged in contexts close to the heart of category theory. In particular, it is well-known that V_Ω is essentially a Grothendieck universe [5].

A simple objection to our move is to note that assuming the existence of an inaccessible cardinal goes beyond the resources that—by convention—we take for granted in mathematical contexts; i.e., ZFC. This is correct; ZFC cannot prove that an inaccessible cardinal exists. However, the increase in consistency strength—while not entirely trivial—is very modest in comparison with the kinds of large cardinal generally considered by set theorists today.[13] Beyond this, it is also possible to provide a schematic axiomatization of V_Ω that is equiconsistent with ZFC and which for almost all intents and purposes is just as good as assuming that the existence of an inaccessible cardinal [9].

A deeper worry emerges from the free use of the axiom of choice above. The attentive reader will have noted that in the proof of Lemma 2.8, I apply the choice three times in ways that could not be eliminated if I had used ZFC and demanded that the functors and natural transformations associated with them were definable classes rather than sets. Personally, I do not think this is a problem. If I want to treat functors and natural transformations as mathematical objects about which we may prove theorems, then they should be objects like sets and not metatheoretical substitutes for them. Nonetheless, constructivist attitudes are hardly uncommon and for these people such uses of choice are likely to be distasteful, at the least. Nonetheless, I suspect most constructivists will already have reservations given that I have assumed ZFC as the background theory of this paper. This prompts the question: how much of this work can be done in constructive settings without choice? We leave that question open here.

As to alternatives, there are at least a couple of other options that would also work. For example, we could use the theory GBC.[14] This is a theory that extends ZFC with an extra sort for classes. It then adds a predicative comprehension axiom, a class replacement axiom and the axiom of global choice. While a little more awkward to use than a single sorted theory like ZFC, it will make light work of Lemma 2.8. Another option is to select a smaller structure from which the models and elementary embeddings may be be procured. For example, one might consider $\mathbb{HC}$, the set of hereditarily countable sets. By the downward Löwenheim–Skolem theorem we know

[12] One might attempt to say that the functors associated with theory categories are merely virtual classes; i.e., definable from parameters. We'll discuss this further below.

[13] Moreover, while it is within the realms of possibility that ZFC is consistent but the addition of an inaccessible cardinal is not, it seems that, in such an unlikely event, it would be more likely that they were both be inconsistent together.

[14] See page 35 of [15].

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

that every model of a theory T has an elementary substructure in $\mathbb{HC}$. Thus, it seems that there will be few places where the comparison of theories requires uncountable structures. An example where this does matter is provided in Proposition 3.15, but it is the only example I am aware of. While other examples will certainly be out there, they appear to capture something vanishingly rare and with unexamined philosophical importance. But even taking this into account, there is nothing stopping us from choosing a larger structure where these effects have been removed. It seems very unlikely that anything near an inaccessible cardinal would be required to achieve this.

§3. Relative interpretation.

In this section, we bring relative interpretation into the picture and we examine the relationship between equivalence relations in category theory and relative interpretability. In particular, we'll show that the relations from interpretability are instances of the category theoretic relations. We start by recalling some basic definitions and results about relative interpretation. This exposition is heavily indebted to Visser's [21] and the reader should consult this resource for further details. For convenience, in this section we'll only consider languages containing relation symbols and thus, no function or constant symbols. We shall work in first-order logic and generally follow the notational conventions of model theory as can be found in [12, 17] or [7].

Definition 3.1. *Let T and S be theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. We say that* t *is a* translation *from the sentences of $\mathcal{L}_S$ to those of $\mathcal{L}_T$ if*:

- *there is a formula δ_t of $\mathcal{L}_T$ with at most v_0 free; and*
- *for all relation symbols R of $\mathcal{L}_S$ there is a formula $t(R)$ of $\mathcal{L}_T$ where* n *is the arity of R and $t(R)$ has at most the variables $v_0, \dots, v_{n-1}$ free,*

such that for every formulae φ of $\mathcal{L}_S$, $t(\varphi)$ is defined recursively in such a way that:

- *if $\varphi := R\bar{x}$, then $t(\varphi) = t(R)(\bar{x})$ where $t(R)(\bar{x})$ is the result of replacing the variables $v_0, \dots, v_{n-1}$ with those in $\bar{x}$ while changing bound variables when required to avoid clashes;*[15]
- *if $\varphi := \neg\psi$, then $t(\varphi) = \neg t(\psi)$;*
- *if $\varphi := \psi \wedge \chi$, then $t(\varphi) = t(\psi) \wedge t(\chi)$; and*
- *if $\varphi := \forall x\psi$, then $t(\varphi) = \forall x(\delta_t(x) \to t(\psi))$ where $\delta_t(x)$ is the result of replacing v_0 with* x *and changing bound variables is required.*

We say that T interprets S via t *if for all sentences $\varphi \in \mathcal{L}_S$*[16]

$$S \vdash \varphi \;\Rightarrow\; T \vdash t(\varphi).$$

Informally speaking, we see that when T interprets S, T is able to *simulate* the behavior of S by proving everything that S can through the lens of the translation. Indeed, this simulation metaphor can be take further by observing that when T

[15] For example, if $t(R)(v_0)$ is $\exists x\ x = v_0$, then $t(R)(x)$ is $\exists y\ y = x$ for some sensibly chosen variable y.

[16] To save some space and help with visual clarity, we shall frequently abuse notation and write $\varphi \in \mathcal{L}_S$ when strictly $\mathcal{L}_S$ is just the underlying non-logical vocabulary, not its formulae. This should not cause any confusion, but in places where it might we shall take care to remark upon it.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

interprets S via t, this gives rise to a function t^* from the models of T to the models of S.

Theorem 3.2. *Suppose T interprets S via* t. *Then* t *determines a function*

$$t^* : mod(T) \to mod(S)$$

such that for all $\mathcal{M} \models T$, $\bar{m} \in M^{lh(\bar{m})}$ and $\varphi \in \mathcal{L}_S$

$$\mathcal{M} \models t(\varphi)(\bar{m}) \;\Leftrightarrow\; t^*(\mathcal{M}) \models \varphi(\bar{m})$$

when for all $i < lh(\bar{m})$, $\mathcal{M} \models \delta_t(m_i)$.

Proof (*Sketch only*). Let $t^*(\mathcal{M})$ be the model of $\mathcal{L}_S$ whose domain

$$t^*(M) = \{x \in M \mid \mathcal{M} \models \delta_t(x)\}$$

and whose interpretation of relation symbols R from $\mathcal{L}_S$ is such that

$$R^{t^*(\mathcal{M})} = \{\bar{x} \in M^n \mid \mathcal{M} \models t(R)(\bar{x})\}.$$

We then prove the equivalence claim (i.e., the $\Leftrightarrow$ claim) by induction on the complexity of formulae. Finally, we show that $t^*(\mathcal{M}) \in mod(S)$ by observing that the interpretation ensures that $\mathcal{M} \models t(\varphi)$ for all $\varphi \in S$ using the equivalence. □

For reasons that will become clear in the next section, we call t^* a *mod-functor*.[17] Note that in the example above t takes us from the language of S to the language of T, while t^* takes us in the other direction from models of T to models of S. To avoid clutter I'll just write t instead of t^* from now on. It should not cause any confusion.

Remark 3.3. *Note that we are only considering* identity-preserving translations *here. Thus, we are not making use of products or quotients. This is arguably a divergence from orthodoxy, but it will be important when we come to Section* 5 *and we'll discuss this issue further there. See* [18] *for an excellent discussion of this.*

We now recall three standard notions of equivalence between theories defined using relative interpretability. For the second of these we require a preliminary definition.

Definition 3.4. *For a theory T in language $\mathcal{L}_T$, we say that a formula $\psi(x, y)$ of $\mathcal{L}_T$* defines a function over T *if*

$$T \vdash \forall x \exists! y \psi(x, y).$$

We say that $\psi(x, y)$ defines an isomorphism *over T if T proves that ψ is a bijection and that for all relation symbols $R \in \mathcal{L}_T$ with arity $n + 1$, T proves that $\forall x_0 \ldots \forall x_n \forall y_0 \ldots \forall y_n$ if $\bigwedge_{i \leq n} \psi(x_i, y_i)$ then*

$$R(x_0, \ldots, x_n) \leftrightarrow R(y_0, \ldots, y_n).$$

We are now in a position to describe three standard notions of equivalence between theories.[18]

[17] Strictly, we are yet to describe the action of t^* on arrows so we do not have a functor yet. This will be described in Lemma 3.6.

[18] See [21] for more details.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

DEFINITION 3.5. *Suppose T and S are theories in the languages $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. Suppose that we have mod-functors $t : mod(T) \to mod(S)$ and $s : mod(S) \to mod(T)$. Then we say that* s *and* t witness *that*:

(1) *T and S are* definitionally equivalent *if*:
- $\mathcal{M} = s \circ t(\mathcal{M})$ *for all models $\mathcal{M}$ of T; and*
- $\mathcal{N} = t \circ s(\mathcal{N})$ *for all modes $\mathcal{N}$ of S.*

(2) *T and S are* strictly bi-interpretable *if there exist definable isomorphisms $\sigma_t(x, y)$ and $\sigma_s(x, y)$ over T and S respectively such that*:
- *$\sigma_t(x, y)$ witnesses that $\mathcal{M} \cong s \circ t(\mathcal{M})$ for all models $\mathcal{M}$ of T; and*
- *$\sigma_s(x, y)$ witnesses that $\mathcal{N} \cong t \circ s(\mathcal{N})$ for all models $\mathcal{N}$ of T.*

(3) *T and S are* iso-congruent *if*:
- $\mathcal{M} \cong s \circ t(\mathcal{M})$ *for all models $\mathcal{M}$ of T; and*
- $\mathcal{N} \cong t \circ s(\mathcal{N})$ *for all modes $\mathcal{N}$ of S.*

We then say T and S are definitionally equivalent (*respectively,* strictly bi-interpretable *and* iso-congruent) *if there are a pair of mod-functors witnessing that this property holds.*

Before we move back to category theory, some remarks about these definitions are warranted.[19] One reason for this is that there is no real consensus around the terminology regarding interpretability. For example, it is common to hear people confuse bi-interpretability with the much weaker relationship of mutual interpretability. Of course, once we have provided precise definitions of these terms the ambiguity is removed. However, there are also different ways of stating these definitions and different terminological traditions within the field that can lead to further confusion.

Many authors aim to provide syntactic definitions of the equivalence relations of Definition 3.5. So rather than using mod-functors to define these equivalences they simply use the translations. For example, Barrett and Halvorson describe definitional equivalence between theories by saying that such theories can be definitionally expanded to become the same theory [4]. It turns out that these two approaches are—in essence—equivalent.[20] In this paper, I've opted to take a more model-based approach for three reasons. First, this is almost always the easiest way to understand how an interpretation works and, in general, most theorems are more easily proven from this perspective.[21] When trying to provide an interpretation, the natural way to think of this is that we are trying to define an internal model. Second, and as we shall discuss in the next section, the model theoretic perspective seems to provide a clearer analogy between equivalence relations based on interpretability and those used in category theory. And finally, I do not believe that a syntactic counterpart can be provided for iso-congruence.

Many authors also take up a more liberal approach to definability in describing these relations. For example, Button and Walsh make use of quotient structures in their definition of bi-interpretability [6]. And similarly, Visser and Friedman make

[19] I'm grateful to a referee for suggesting the addition of this section.

[20] See Proposition 7.1 in the Appendix for a quick proof of this. Note, however, that the equivalence requires that the languages of the respective theories are disjoint. See [16] for a detailed discussion of this issue. For an example of natural equivalence relations that is similar but weaker than definitional equivalence, see [2].

[21] The only case—that I'm aware of—where the syntactic approach is superior is in the proofs of Section 5 in [22].

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

use of multi-dimensional interpretations in [22]. Using these approaches, objects from an interpreted theory can be represented by equivalence classes of sequences of objects from the theory that is providing the interpretation. By contrast, in this paper we have restricted our attention to one-dimensional and identity preserving interpretations where the objects of the interpreted theory are simply represented by objects of the theory doing the interpreting. For this reason we use the term *strict bi-interpretability* instead of ordinary bi-interpretability. This means that the definitions above are stronger and thus, more difficult to satisfy. Our main reason for doing this is that the work done by multi-dimensional interpretations and quotients can almost always be recovered by the use of Morita interpretation.[22] Moreover, a primary goal of this paper is to clarify the relationship between equivalences based on Morita approaches and those of category theory. We might say that Morita interpretation is a competitor to the approaches used by Button and Walsh [6]. While our goal here is not to adjudicate that debate, I think the reader will see in Section 5 that there is certain efficiency to the Morita technique that warrants further investigation.

3.1. Back to category theory. We are now ready to compare the equivalence notions of category with those from relative interpretability. We start by observing that functions between models of one theory and another that are determined by interpretations give rise to functors.

LEMMA 3.6. *If* $t : mod(T) \to mod(S)$ *is derived from an interpretation, then* t *determines a functor.*

Proof. t takes models of T and returns models of S. To obtain a genuine functor we need to also explain the action of t on arrows in the $mod(T)$ category. Given an elementary embedding $j : \mathcal{M} \to \mathcal{N}$, we let $t(j)$ be the restriction of j to $t(M)$. To see $t(j) : t(\mathcal{M}) \to t(\mathcal{N})$ is an elementary embedding consider a sequence $\bar{m}$ from $t(M)$ and a formula $\varphi(\bar{x})$ from $\mathcal{L}_S$. Then we see that

$$\begin{aligned} t(\mathcal{M}) \models \varphi(\bar{m}) &\Leftrightarrow \mathcal{M} \models t(\varphi)(\bar{m}) \\ &\Leftrightarrow \mathcal{N} \models t(\varphi)(j\bar{m}) \\ &\Leftrightarrow t(\mathcal{N}) \models \varphi(j\bar{m}). \end{aligned}$$

It is then easy to see that t preserves the identity arrow and composition. □

It is worth nothing that this is a place where we make essential use of our requirement that arrows in a theory category are elementary embeddings.[23]

This tells us that interpretations determine functors and thus we can see the relative interpretations as special cases of functors between theory categories. However in

[22] For an excellent discussion of how Morita interpretation relates to interpretations based on quotient structure approaches, see [18].

[23] To see this suppose that $f : \mathcal{M} \to \mathcal{N}$ is not an elementary embedding where $\mathcal{M}$ is a model of ZFC^- (i.e., ZFC without the powerset axiom). Then we may fix some $\varphi(x)$ in the language of arithmetic and some sequence $m \in M^{<\omega}$ such that

$$\mathcal{M} \models \varphi(m) \,\&\, \mathcal{N} \models \neg\varphi(f(m)).$$

We use the fact that $\mathcal{M}$ is a model of ZFC^- to ensure that $\varphi(x)$ only requires one free variable. Let t then be the interpretation that merely restricts the domain by letting $\delta_t(x)$ be $\varphi(x)$. Then $f \restriction t(M)$ is not a function from $t(M)$ to $t(N)$ so it cannot be an elementary embedding.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

general, many functors will not be determined by interpretations. For example any functor that takes a model and returns a model with greater cardinality cannot be determined by an interpretation. We address this by showing that the background framework for our interpretative equivalences can be seen as a very natural restriction of the framework we used for our equivalences between theory categories. To make this idea clearer, we shall first abstract away a formal framework in which the equivalences of Section 2 can be understood. First recall that a *2-category* is a generalization of the ordinary notion of category obtained by adding an extra sort of arrows and governing their interaction with objects and ordinary arrows with new axioms.[24] In this section, we shall call; ordinary arrows, 1-morphisms; and the new arrows, 2-morphisms. From our point of view the key point is that 2-morphisms can be seen as arrows between 1-morphisms, while 1-morphisms are arrows between objects. We can then define 1-isomorphisms and 2-isomorphisms in the obvious way. We shall say that two (small) categories are 1-equivalent, if there are functors going back and forth between them that are 2-isomorphic; in other words, they are categorically equivalent. The classic example of a 2-category is obtained by taking: every small category $\mathcal{C}$ as an *object*; functors F between categories as *1-morphisms*; and natural transformations η. between those functors as *2-morphisms*.[25] Let us call this 2-$\mathbb{CAT}$. This motivates the following definition.

Definition 3.7. *Let* $\mathbb{TH}$ *be the* 2*-category of theory categories with*:

- *objects*: *categories* $mod(T)$ *for some theory* T;
- 1*-morphisms*: *functors between those categories*; *and*
- 2*-morphisms*: *natural transformations between the functors.*

Let's call this the *theory framework*. It is then easy to see that isomorphism and categorical equivalence can be articulated using $\mathbb{TH}$. More specifically, we see that:

Proposition 3.8. *Let* $mod(T)$ *and* $mod(S)$ *be theory categories. Then*:

(1) $mod(T)$ *is categorically isomorphic to* $mod(S)$ *if they are* 1*-isomorphic in* $\mathbb{TH}$.
(2) $mod(T)$ *is categorically equivalent to* $mod(S)$ *if they are* 1*-equivalent in* $\mathbb{TH}$.

With this in hand, we can now define a restriction of the theory framework that fits perfectly with our definitions from interpretability.

Definition 3.9. *Let* $\mathbb{TH}_{Def}$ *be the sub-*2*-category of* $\mathbb{TH}$ *where*:

- 1*-morphisms are the mod-functors determined by interpretations*; *and*
- 2*-morphisms are the natural transformations that are given by definable functions.*[26]

Let us call this the *definable theory framework*. We the observe that the equivalence relations given by interpretability fit very naturally here.

[24] A precise definition of a 2-category can be found in [14], although we'll make no use of this here.

[25] There are of course problems around size here. So let us assume for our purposes that a small category is one whose objects objects all come from V_Ω and which is such that for any pair of objects the arrows between them form a set in V_Ω.

[26] In the sense of Definition 3.4.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

THEOREM 3.10. *Let T and S be theories in $\mathcal{L}_T$ and $\mathcal{L}_S$, respectively. Then*:

(1) *T and S are definitionally equivalent iff $mod(T)$ and $mod(S)$ are 1-isomorphic in $\mathbb{TH}_{def}$.*
(2) *T and S are strictly bi-interpretable iff $mod(T)$ and $mod(S)$ are 1-equivalent in $\mathbb{TH}_{def}$.*
(3) *T and S are iso-congruent iff $mod(T)$ and $mod(S)$ are objectively equivalent as witnessed by functors from $\mathbb{TH}_{def}$.*

Proof. (1) and (3) are immediate. (2) Suppose that T and S are strictly bi-interpretable. Then let $t : mod(T) \to mod(S)$ and $s : mod(S) \to mod(T)$ be such that there exist formulae $\psi_t(x, y) \in \mathcal{L}_T$ and $\psi_s(x, y) \in \mathcal{L}_S$ defining isomorphisms over T and S respectively that witness the strict bi-interpretation. For a model $\mathcal{M}$ of T, let $\tau_{\mathcal{M}} : \mathcal{M} \cong s \circ t(\mathcal{M})$ be the isomorphism defined by $\psi_t(x, y)$ in $\mathcal{M}$. Similarly, for models $\mathcal{N}$ of S, let $\sigma_{\mathcal{N}} : \mathcal{N} \cong t \circ s(\mathcal{N})$. This satisfies the back and forth condition for bi-interpretation, it suffices to show that $\tau.$ and $\sigma.$ are natural transformations. We just prove this for $\tau.$ as the proof for $\sigma.$ is similar. Thus, it will suffice to show that the following diagram commutes.

$$\begin{tikzcd}
s \circ t(\mathcal{M}) \arrow[r, "s \circ t(j)"] & s \circ t(\mathcal{N}) \\
\mathcal{M} \arrow[u, "\sigma_{\mathcal{M}}"] \arrow[r, "j"'] & \mathcal{N} \arrow[u, "\sigma_{\mathcal{N}}"']
\end{tikzcd}$$

To see this first observe that for $x \in s \circ t(M)$, $j(x) = s \circ t(j)(x)$, since $s \circ t(j)$ is just a restriction of j. Now suppose that $x \in M$. Then we see that

$$s \circ t(j)(\sigma_{\mathcal{M}}(x)) = j(\sigma_{\mathcal{M}}(x)) = \sigma_{\mathcal{N}}(j(x))$$

as required. □

I'd like to suggest that this result motivates a somewhat subtle conceptual turning of the tables. While relative interpretability has had a long history in mathematical logic, categorical equivalence relations between theory categories are relatively new on the scene. As such, it seems natural to think of categorical approaches as generalizations of the core notions from interpretability. The results above suggest a different story. The categorical equivalence relations can be seen as the prototype from which the interpretative picture can be derived by a natural restriction on the theory framework. We shall put this idea to work in Section 5, by considering a further natural restriction of the theory framework.

3.1.1. Hierarchy?

As was the case with the categorical equivalence relations, it is natural to ask whether the equivalences given in Definition 3.5 are arranged in a strict hierarchy. I have only been able to provide partial answers to these questions. First we observe.

PROPOSITION 3.11. (1) *If T and S are definitionally equivalent, then they are strictly bi-interpretable.*
(2) *If T and S are strictly bi-interpretable, then they are iso-congruent.*

We now consider whether the top two positions collapse.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

PROBLEM 3.12. *Is there a pair of theories that are strictly bi-interpretable but not definitionally equivalent*?

We note first that there is a beautiful result from Visser and Friedman [22] that could appear to establish that there is such a pair. However, it turns out that this is not the case since the interpretations used there are not identity-preserving. More specifically, they make use of a quotient interpretation.[27] Nonetheless, the main theorem of that paper gives an indication of how rare such an example would be.

THEOREM 3.13 [22]. *If T and S are sequential*[28] *theories that are strictly bi-interpretable, then T and S are definitionally equivalent.*

The result is established using an internalization of the Cantor–Bernstein theorem. This tells us that if we are interested in comparing theories that can provide a foundation for mathematics, then they will certainly be sequential; and thus, whenever they are strictly bi-interpretable they are definitionally equivalent. This brings us to the lower two rungs on the ladder.

PROBLEM 3.14. *Is there a pair of theories that are iso-congruent but not strictly bi-interpretable*?

The following provides a partial answer.

PROPOSITION 3.15. *If we restrict theory categories to countable models, there is a pair of theories that are iso-congruent but not strictly bi-interpretable.*

Proof. Let D be the theory in the language $\mathcal{L}_D = \{<, d_n\}_{n\in\omega}$ which says that $<$ is a dense linear order without end points and that $d_n < d_{n+1}$ for all $n \in \omega$. Let B be a theory in the language $\mathcal{L}_B = \{\prec, b_n\}_{n\in\omega}$ that $\prec$ is dense linear order with no top point but with a bottom point that is d_0 and that $b_n < b_{n+1}$ for all $n \in \omega$.[29]

First we show that iso-congruence. Let $t : mod(D) \to mod(B)$ discard everything below d_0 and preserve everything else. More precisely, let

$$\begin{aligned} \delta_t(x) &:= d_0 < x \vee x = d_0 \\ t(x \prec y) &:= x < y \\ t(d_n) &:= d_n,\ \forall n \in \omega. \end{aligned}$$

Let $s : mod(B) \to mod(D)$ discard the bottom element b_0 and then shift each d_n up to the next remaining constant b_{n+1}. More precisely, let

$$\begin{aligned} \delta_s(x) &:= x \neq b_0 \\ \delta_s(x < y) &:= x \prec y \\ t(d_n) &:= b_{n+1},\ \forall n \in \omega. \end{aligned}$$

Let $\mathcal{A}$ be a model of D. We claim that $s \circ t(\mathcal{A}) \cong \mathcal{A}$. To see this note that $s \circ t(\mathcal{A})$ is the submodel of $\mathcal{A}$ obtained by removing those elements $\leq d_0^{\mathcal{A}}$. Observe $\mathcal{A}$ and $s \circ t(\mathcal{A})$

[27] See Remark 3.3. We give a version of their example using Morita techniques below in Section 5.2.2.

[28] This means it interprets AS (see Section 5.2.2). In fact, the weaker notion of a conceptual theory suffices.

[29] Of course, we could have used the same language for both, but this tends to make things more confusing.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

can be seen as infinite sequences of countable dense linear orders without endpoints demarcated by their interpretations of the $\langle d_n \rangle_{n\in\omega}$ sequence. Since these orders are countable, it can then be seen, via Cantor, that every such linear order on the sequence of $\mathcal{A}$ is isomorphic to every such linear order on the sequence in $s \circ t(\mathcal{A})$. Thus $\mathcal{A}$ and $s \circ t(\mathcal{A})$ are clearly isomorphic. Similarly in the other direction, given a model $\mathcal{B}$ of B, we see that $t \circ s(\mathcal{B})$ is the submodel of $\mathcal{B}$ obtained by removing from the domain those elements below $b_0^{\mathcal{B}}$ and so it can be seen that $t \circ s(\mathcal{B})$ and $\mathcal{B}$ are isomorphic as required.

Remark *Note, however, that these interpretations do not witness a strict bi-interpretation. This is because the isomorphism between $\mathcal{A}$ and $s \circ t(\mathcal{A})$ cannot be defined in $\mathcal{A}$. To get some insight into this note that a formula $\sigma(x, y)$ representing such an isomorphism would need to be such that D entailed $\sigma(c_n, c_{n+1})$ for every $n \in \omega$.* □

To see that D and B are not strictly bi-interpretable, suppose toward a contradiction that they. Then fix $f : mod(D) \to mod(B)$, $g : mod(B) \to mod(D)$ and formulae $\varphi_D(x, y)$ and $\varphi_B(x, y)$ from $\mathcal{L}_D$ and $\mathcal{L}_B$ respectively where:

- $\mathcal{A} \cong g \circ f(\mathcal{A})$ for all models $\mathcal{A}$ of D; and
- $\mathcal{B} \cong f \circ g(\mathcal{B})$ for all models $\mathcal{B}$ of B

and the relevant isomorphisms are defined by φ_D and φ_B respectively.

Now we observe that B and D are are too weak to define many functions.

Claim *D cannot define any non-trivial isomorphisms.*

Proof. Let $\mathcal{A}$ be a model of D. Then the only definable elements of $\mathcal{A}$ are those $d_n^{\mathcal{A}}$ for $n \in \omega$. This means that the only bijections that can be defined in $\mathcal{A}$ are those that permute a finite some finite subsets of $\{d_n^{\mathcal{A}}\}_{n\in\omega}$. But any bijection that gives such a finite permutation is not an isomorphism since it will break the required ordering on the d_n sequence. □

This means that if B and D are strictly bi-interpretable, then they are definitionally equivalent and so, in particular, D must be able to interpret B in such a way that the domain is preserved. Thus, the following claim suffices to establish the proposition.

Claim *D cannot interpret B with a domain preserving translation.*

Proof. See Appendix. □

Note that proof illustrates a quite draconian limitation of interpretative techniques. We are blocked by what in this context seem like arbitrary restrictions on our ability to provide interpretations. This provides some motivation for the more general relations like categorical equivalence and Morita equivalence. We'd like to free ourselves of at least some of these bonds.

§4. Multi-sorted interpretation.

In this section and the next, we develop some of the space between categorical equivalences and their restricted interpretative cousins. This will be achieved by letting theories define new domains (or sorts) over which they may then quantify. The approach taken here is essentially from Barrett and Halvorson [4]. However, we shall aim to generalize their framework by defining a notion of *Morita interpretation* from which Morita equivalence can then be derived. This will allow us to demonstrate that Morita equivalence is also a natural restriction of categorical

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

equivalence relations. In order to develop this approach, we first recall some basic facts about multi-sorted languages and interpretations between them.

4.1. Multi-sorted languages. Our exposition here is very similar to Barrett and Halvorson [4] and we'll mainly aim to note divergences rather than provide a full exposition. A multi-sorted language $\mathcal{L}$ will consist of sorts, relations symbols and function symbols. We use $\sigma, \tau, \sigma_0, \tau_0, \sigma_1, \tau_1, \ldots$ for sorts. Relation symbols have an arity of the form $\sigma_0 \times \cdots \times \sigma_n$, where $\sigma_0, \ldots, \sigma_n$ are sorts from $\mathcal{L}$. Function symbols f have an arity of the form $\sigma_0 \times \cdots \times \sigma_n \to \sigma$.[30] For each sort $\sigma \in \mathcal{L}$ we shall have countably many variable symbols labeled with their sort $v_0^\sigma, v_1^\sigma, \ldots$.[31] And we shall use x^σ, y^σ as metavariables of sort σ.

A model $\mathcal{M}$ for $\mathcal{L}$, will be a structure with a domain M_σ for each sort σ from $\mathcal{L}$. Moreover for each relation symbol R from $\mathcal{L}$ with arity $\sigma_0 \times \cdots \times \sigma_n$, the interpretation of R in $\mathcal{M}$, abbreviated $R^{\mathcal{M}}$, will be a subset of $M_{\sigma_0} \times \cdots \times M_{\sigma_n}$; and for each function symbol with arity $\sigma_0 \times \cdots \times \sigma_n \to \sigma$, $f^{\mathcal{M}}$ will be a function from $M_{\sigma_0} \times \cdots \times M_{\sigma_n}$ to M_σ. The terms and formulae of $\mathcal{L}$ can then be defined inductively in the obvious way as are the term denotation and satisfaction relations.[32] We'll use φ, ψ, χ as metavariables for formulae and $s_0, \ldots, s_n$ as metavariables for terms.

4.2. Interpretation. We now define interpretation in the context of multi-sorted languages. This is a little more fiddly than Definition 3.1 since our work with Morita extensions pushes us to be very explicit in our treatment of function symbols. To facilitate this, we first translate formulae of a language $\mathcal{L}$ into a normal form that ensures that the only atomic formulae in which function symbols occur are of the form $x^\sigma = f(x_0^{\sigma_0}, \ldots, x_n^{\sigma_n})$; i.e., we can never have an atomic formula of the form $x^\sigma = f(g(x^{\sigma_0})$. Let us call this *function normal form*. Once a formula is in this form, the usual translation process works smoothly. The following proposition establishes that nothing is lost if we just use formulae in function normal form.

PROPOSITION 4.1. *Let $\mathcal{L}$ be a multi-sorted language. For any formula $\varphi(x_0^{\sigma_0}, \ldots, x_n^{\sigma_n})$ from $\mathcal{L}$ there is a logically equivalent formula $\varphi^\dagger(x_0^{\sigma_0}, \ldots, x_n^{\sigma_n})$ that is in function normal form.*

We leave it to the reader to establish this, however, an example could be helpful to illustrate the effect. For simplicity, suppose $\mathcal{L}$ has just one sort and suppose $\varphi := \exists x(R(x, f(x)) \wedge g(x) = f(x))$ is a formula of $\mathcal{L}$. Then we see that φ is equivalence to the following formula in function normal form:

$$\exists x(\forall y(y = f(x) \to Rxy) \wedge \forall z(z = g(x) \wedge y = f(x))).$$

Given the proposition above, let us assume without loss of generality for the remainder of this section that we only deal with formulae in function normal form. We can now provide our definition of an interpretation between multi-sorted languages.

[30] Note that $\sigma_0 \times \cdots \times \sigma_n$ and $\sigma_0 \times \cdots \times \sigma_n \to \sigma$ are not, in general, sorts of $\mathcal{L}$ themselves.

[31] Barrett and Halvorson [4] label their quantifies rather than their variables. This makes no substantive difference, although perhaps makes the mechanism of translation a little more transparent.

[32] See Section 2 of [4].

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Definition 4.2. *Let T and S be theories in multi-sorted languages $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. We say that a function* t *is a* translation *from the sentences of $\mathcal{L}_S$ to those of $\mathcal{L}_T$ if: for all sorts $\sigma \in \mathcal{L}_S$, $t(\sigma)$ is a sort in $\mathcal{L}_T$; and*

- *for all sorts $\sigma \in \mathcal{L}_S$, there is a formula $\delta_{t,\sigma}$ of $\mathcal{L}_T$ with at most $v_0^{t(\sigma)}$ free,*
- *for all relation symbols R of $\mathcal{L}_T$ of arity $\sigma_0 \times \cdots \times \sigma_n$, there is a formula $t(R)$ with at most the variables $v_0^{t(\sigma_0)}, \dots, v_n^{t(\sigma_n)}$ free, and*
- *for all function symbols* g *of $\mathcal{L}_T$ of arity $\sigma_0 \times \cdots \times \sigma_n \to \sigma_{n+1}$, there is a formula $t(g)$ with at most the variables $v_0^{t(\sigma_0)}, \dots, v_{n+1}^{t(\sigma_{n+1})}$ free,*

such that for formulae φ of $\mathcal{L}_S$ (in function normal form), $t(\varphi)$ is defined recursively as follows:

- *if $\varphi := Rx_0^{\sigma_0} \dots x_n^{\sigma_n}$ where $R \in \mathcal{L}_S$ is a relation symbol of arity $\sigma_0 \times \cdots \times \sigma_n$, then*
$$t(\varphi) = t(R)(x_0^{t(\sigma_0)}, \dots, x_n^{t(\sigma_n)}),$$
- *if $\varphi := (x_{n+1}^{\sigma_{n+1}} = g(x_0^{\sigma_0}, \dots, x_n^{\sigma_n}))$ where $g \in \mathcal{L}_S$ is a function symbol of arity $\sigma_0 \times \cdots \times \sigma_n \to \sigma$, then*
$$t(\varphi) = t(g)(x_0^{t(\sigma_0)}, \dots, x_{n+1}^{t(\sigma_{n+1})}),$$
where $t(g)$ is a formula of $\mathcal{L}_T$ such that
$$T \vdash \forall x_0^{t(\sigma_0)} \dots \forall x_n^{t(\sigma_n)} \exists!\, x^{\sigma_{n+1}}\ t(g)(x_0^{t(\sigma_0)}, \dots, x_{n+1}^{t(\sigma_{n+1})}),$$
- *$t(\neg\varphi) = \neg t(\varphi)$;*
- *$t(\varphi \wedge \psi) = t(\varphi) \wedge t(\psi)$; and*
- *$t(\forall x^\sigma \varphi) = \forall x^{t(\sigma)}(\delta_{t,\sigma}(x^{t(\sigma)}) \to t(\varphi)$,*

where $t(R)(x_0^{\sigma_0}, \dots, x_n^{\sigma_n})$ is the result of replacing each $v_i^{t(\sigma_i)}$ with $x_i^{t(\sigma_i)}$ while changing bound variables when required to avoid clashes, and analogous remarks apply to $t(g)(x_0^{t(\sigma_0)}, \dots, x_{n+1}^{t(\sigma_{n+1})})$ and $\delta_{t,\sigma}(x^{t(\sigma)})$.

We say that T interprets S *via a translation* t, *if for all sentences $\varphi \in \mathcal{L}_S$ we have*
$$S \vdash \varphi \ \Rightarrow\ T \vdash t(\varphi).$$

Then as with ordinary interpretation, we see that an interpretation gives rise to a function from models of T to models of S. The following example shows how a single-sorted theory ZFC can be used to interpret a multi-sorted theory.

Example 4.3. *ACA_0 is a theory in a language with two sorts: ob for natural numbers; and cl for classes of numbers. It can be axiomatized by extending PA with: a comprehension schema that says that any formula that avoids quantification over class variables determines a class of natural numbers; and an induction axiom saying that every class has a least element. A detailed description can be found in* [20], *however, the basic idea is that we take a model of arithmetic and a new sort for subsets of the natural numbers. In contrast ZFC is a theory with the single sort of sets. ZFC can be used to interpret ACA_0 with an interpretation* t *such that $t(ob) = t(cl) = sets$. So both sorts of ACA_0 are sent to the only sort for ZFC, sets. We then define number domain with a formula $\delta_{t,ob}(x)$ that says $x \in \omega$; and the class domain with a formula $\delta_{t,cl}(x)$ that says $x \subseteq \omega$. We leave the rest of the interpretation to the reader.*

The following theorem generalizes our earlier Theorem 3.2 linking interpretations to mod-functors.

THEOREM 4.4. *If T interprets S via $t : \mathcal{L}_S \to \mathcal{L}_T$, then* t *determines a function*

$$t^* : mod(T) \to mod(S)$$

such that for all models $\mathcal{M}$ of T, formulae $\varphi(x_0^{\sigma_0}, \dots, x_n^{\sigma_n})$, we have

$$\mathcal{M} \models t(\varphi)(m_0, \dots, m_n) \;\Leftrightarrow\; t^*(\mathcal{M}) \models \varphi(m_0, \dots, m_n)$$

when for all $i \leq n$, $\mathcal{M} \models \delta_{t,\sigma_i}(m_i)$.

We omit the proof, which is similar to that of Theorem 3.2. As above, we shall just write t^* instead of t below. We can now define analogues of the interpretative equivalences described in Section 3. Everything works much the same, however, we shall take a little care to explain how isomorphisms and elementary equivalences work in this setting.

DEFINITION 4.5. *Let us say that $f : \mathcal{M} \to \mathcal{N}$ is a* homomorphism *if $f : \prod_{\sigma \in \mathcal{L}}(M_\sigma \to N_\sigma)$ is such that: for all relation symbols R of arity $\sigma_0 \times \cdots \times \sigma_n$ and $m_0, \dots, m_n$ from $M_{\sigma_0}, \dots, M_{\sigma_n}$ respectively,*[33]

$$\mathcal{M} \models Rm_0 \dots m_n \;\Rightarrow\; \mathcal{N} \models Rf_{\sigma_0}(m_0) \dots f_{\sigma_n}(m_n),$$

and for all function symbols g *of arity $\sigma_0 \times \cdots \times \sigma_n \to \sigma$ and $m_0, \dots, m_n$ from $M_{\sigma_0}, \dots, M_{\sigma_n}$ respectively,*

$$g^{\mathcal{M}}(m_0, \dots, m_n) = g^{\mathcal{N}}(f_{\sigma_0}(m_0), \dots, f_{\sigma_n}(m_n)).$$

We say that $f : \mathcal{M} \to \mathcal{N}$ is an embedding *if* f *is a homomorphism and for all relation symbols (including identity) R of arity $\sigma_0 \times \cdots \times \sigma_n$ and $m_0, \dots, m_n$ from $M_{\sigma_0}, \dots, M_{\sigma_n}$ respectively,*

$$\mathcal{M} \models Rm_0 \dots m_n \;\Leftrightarrow\; \mathcal{N} \models Rf_{\sigma_0}(m_0) \dots f_{\sigma_n}(m_n).$$

We say that $f : \mathcal{M} \to \mathcal{N}$ is an isomorphism *if there is some homomorphism $g : \mathcal{N} \to \mathcal{M}$ such that*

$$g \circ f = id_{\mathcal{M}} \text{ and } f \circ g = id_{\mathcal{N}}.$$

We say that $f : \mathcal{M} \to \mathcal{N}$ is an elementary embedding *if for all formulae $\varphi(x_{\sigma_0}, \dots, x_{\sigma_n})$ of $\mathcal{L}$ and $m_0, \dots, m_n$ from $M_{\sigma_0}, \dots, M_{\sigma_n}$ respectively,*

$$\mathcal{M} \models \varphi(m_0 \dots m_n) \;\Leftrightarrow\; \mathcal{N} \models \varphi(f_{\sigma_0}(m_0) \dots f_{\sigma_n}(m_n)).$$

For a theory T in a multi-sorted language $\mathcal{L}$, we now let $mod_{mult}(T)$ be the category whose objects are models of T and whose arrows are elementary embeddings between them. The following proposition establishes that interpretations also determine functors in the multi-sorted setting. The proof is similar to that of Lemma 3.6.

PROPOSITION 4.6. *If $t : mod_{mult}(T) \to mod_{mult}(S)$ is derived from an interpretation, then* t *determines a functor.*

[33] Note that f takes a sort σ and returns a function $f_\sigma : M_\sigma \to N_\sigma$.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Then in preparation for the more general notion of strict bi-interpretability, we define what it means for an isomorphism to be definable over some theory.

Definition 4.7. *Suppose $\mathcal{L}$ is a multi-sorted language with sorts $\{\sigma_i\}_{i\in I}$. Let T be an $\mathcal{L}$-theory. We say that $\{\psi_i(x,y)\}_{i\in I}$* defines a function *over T if for all models $\mathcal{N}$ and $\mathcal{M}$ of T where $\mathcal{N}$ is a submodel of $\mathcal{M}$ there is some $f : \mathcal{M} \cong \mathcal{N}$ such that for all $i \in I$ and all $m_0, m_1 \in M_{\sigma_i}$*

$$f_{\sigma_i}(m_0) = m_1 \Leftrightarrow \mathcal{M} \models \psi_i(m_0, m_1).$$

Finally, we are can define the restricted framework suitable for multi-sorted interpretation.

Definition 4.8. *Let $\mathbb{TH}_{mult}$ be the 2-category with:*

- *objects: categories $mod_{mult}(T)$ where T is a theory in a multi-sorted language $\mathcal{L}$;*
- *1-morphisms: functors between those categories; and*
- *2-morphisms: natural transformations between those functors.*

Let $\mathbb{TH}_{mult-def}$ be the sub-2-category of $\mathbb{TH}_{mult}$ where: the 1-morphisms are given by mod-functors given by interpretations; and the 2-morphism are given by definable functions.

We may then define the standard equivalence relations in much same way as we did above. This time we take the category theoretic equivalences as the prototype from which the ordinary definitions are obtained.

Definition 4.9. *Suppose T and S are theories in the multi-sorted languages $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. We say that:*

- T and S are definitionally equivalent *if they are 1-isomorphic in $\mathbb{TH}_{mult-def}$;*
- T and S are strictly bi-interpretable *if they are 1-equivalent in $\mathbb{TH}_{mult-def}$; and*
- T and S are iso-congruent *if $mod_{mult-def}(T)$ and $mod_{mult-def}(S)$ are objectively equivalent as witnessed by functors from $\mathbb{TH}_{mult-def}$.*

Thus we have the natural generalization of the theory of interpretability to multi-sorted languages.

§5. Morita interpretation.

5.1. Morita extension. The framework reviewed in the previous section allows us to compare theories articulated in multi-sorted languages. However, there are many occasions where we might want to compare, say, a single-sorted theory with a multi-sorted theory but the tools above do not—in general—allow for this.[34] For example, we might want to compare the axiomatization of category theory with sorts for arrows and objects with its axiomatization using only arrows. In this section, we describe a framework developed by Barrett and Halvorson [4] for exactly this purpose. We have two main goals in this section. First, we aim to generalize their approach so that we not only get a new notion of equivalence, but further a new notion of interpretation: Morita interpretation. Second, we use this to show that there is another natural restriction of $\mathbb{TH}$ that gives us Morita equivalence and more. I'm going to depart a little from

[34] For an instance where it does work, recall Example 4.3.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

the presentation of Barrett and Halvorson [4], although most of the deviations are minor. The definition below describes a way of taking a theory T articulated in some language $\mathcal{L}$ and adding a new sort σ corresponding to one of four common type constructions.

Definition 5.1. *Let T be a theory in a (possibly) multi-sorted language $\mathcal{L}$. We say that T^* is a* pure Morita successor *of T in $\mathcal{L}^* \supsetneq \mathcal{L}$ where $\sigma \in \mathcal{L}^* \backslash \mathcal{L}$ is the new sort if one of the following holds*:

- (*Product*) There exist sorts $\sigma_0, \sigma_1 \in \mathcal{L}$, $\mathcal{L}^* = \mathcal{L} \cup \{\sigma, \pi_0, \pi_1\}$ where $\pi_0 : \sigma \to \sigma_0$ and $\pi_1 : \sigma \to \sigma_1$; and T^* is T extended by
$$\forall x_0^{\sigma_0} \forall x_1^{\sigma_1} \exists! \, y^\sigma (\pi_0(x_0^{\sigma_0}) = y^\sigma \wedge \pi_1(x_1^{\sigma_1}) = y^\sigma).$$
- (*Co-product*) There exist sorts $\sigma_0, \sigma_1 \in \mathcal{L}$, $\mathcal{L}^* = \mathcal{L} \cup \{\sigma, \pi_0, \pi_1\}$ where $\pi_0 : \sigma_0 \to \sigma$ and $\pi_1 : \sigma_1 \to \sigma$; and T^* is T extended by
$$\forall y^\sigma (\exists x_0^{\sigma_0} \, \pi_0(x_0^{\sigma_0}) = y^\sigma \vee \exists x_1^{\sigma_1} \, \pi_1(x_1^{\sigma_1}) = y^\sigma) \wedge$$
$$\forall x_0^{\sigma_0} \forall x_1^{\sigma_1} (\pi_0(x_0^{\sigma_0}) \neq \pi_1(x_1^{\sigma_1})).$$
- (*Subsort*) There exists a sort $\sigma_0 \in \mathcal{L}$ and a formula $\varphi(x^{\sigma_0})$ from $\mathcal{L}$; $\mathcal{L}^* = \mathcal{L} \cup \{\sigma, \pi\}$ where $\pi : \sigma \to \sigma_0$; and T^* is T extended by
$$\forall x^{\sigma_0} (\varphi(x^{\sigma_0}) \leftrightarrow \exists z^\sigma \, \pi(z^\sigma) = x^{\sigma_0})) \wedge$$
$$\forall x_0^\sigma, x_1^\sigma (\pi(x_0^\sigma) = \pi(x_1^\sigma) \to x_0^\sigma = x_1^\sigma).$$
- (*Quotient*) There exists sort σ_0 and a formula $\varphi(x^{\sigma_0}, y^{\sigma_0})$ in $\mathcal{L}$ and T proves $\varphi(x^{\sigma_0}, y^{\sigma_0})$ represents an equivalence relation; $\mathcal{L}^* = \mathcal{L} \cup \{\sigma, \pi\}$ where $\pi : \sigma_0 \to \sigma$; and T^* is T extended by
$$\forall x_0^{\sigma_0}, x_1^{\sigma_0} (\pi(x_0^{\sigma_0}) = \pi(x_1^{\sigma_0}) \leftrightarrow \varphi(x_0^{\sigma_0}, x_1^{\sigma_0})) \wedge$$
$$\forall y^\sigma \exists x^{\sigma_0} (\pi(x^{\sigma_0}) = y^\sigma).$$

Let T^+ be a theory in a language $\mathcal{L}^+$ extending $\mathcal{L}^*$ with possibly new relation and function symbols. We say that T^+ is a *mixed Morita successor* of T if T^+ interprets T^* in the sense of Definition 4.2.[35]

Thus, we are allowed to add new sorts corresponding to products, co-products, subsorts and quotients. For example, if I am working in the PA in the language of arithmetic, I could add a subsort corresponding to the prime numbers. The distinction between pure and mixed Morita successors is introduced for a technical reason. Barrett and Halvorson [4] define a *Morita extension* to be what we have called a mixed Morita successor. However, we shall do things a little differently by only using pure Morita successors and then using a multi-sorted interpretation at the end. We demonstrate below that this makes no substantive difference. However, the approach taken here makes for an easier comparison with the categorical approach above. We shall generally omit the "pure" and call pure Morita successors, "Morita successors," unless confusion could arise.

[35] This means that a mixed Morita extension can add new relation symbols, in addition to new sorts (and their accompanying functions).

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Definition 5.2. *We say that T^* is a* pure (and respectively mixed) Morita expansion *of T if T^* is the theory resulting after taking finitely many pure (mixed) Morita successors of T.*[36]

We then note that when T^* is a Morita expansion of T any model of T can be expanded to obtain a model of T^* that is unique up to isomorphism.

Theorem 5.3 [4]. *Let T be a theory articulated in a language $\mathcal{L}$. If T^* is a mixed (or pure) Morita successor of T, then for any model $\mathcal{M}$ of T, there is model $\mathcal{M}^*$ of T^* whose reduct to $\mathcal{L}$ is $\mathcal{M}$. Moreover, any two such models are isomorphic.*

From this it follows that T^* is model-theoretic conservative extension of T. Thus, we might say that whatever is added by a Morita expansion comes at a relatively low cost.

5.2. Morita interpretation and equivalence(s). Using Morita expansions, we can now describe a notion of Morita interpretation that is implicit in Barrett and Halvorson [4]. The main work of this section will be establishing that these interpretations can be understood as giving rise to functors in $\mathbb{TH}_{mult-def}$.

Definition 5.4. *Let us say that* T Morita interprets S, *if T has a Morita expansion T^+ that interprets S. Let us say that T and S are* mutually Morita interpretable *if T and S have Morita expansions T^+ and S^+ that interpret each other.*

Observe that unlike ordinary relative interpretation, Morita interpretation has two components. First we make a Morita expansion, then we make the (usually multi-sorted) interpretation. We'd like now to define something like a mod-functor for Morita interpretation, however, the initial Morita expansion poses a problem: since we are adding new domains, there will be many different ways to populate them. Thus a Morita interpretation does not determine a particular functor but rather a family of them.

Definition 5.5. *Let T and S be theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. Suppose T Morita interprets S via the Morita expansion T^+ in $\mathcal{L}_{T^+}$ and the translation $t : \mathcal{L}_S \to \mathcal{L}_{T^+}$. Let us say that a functor $t^\dagger : mod(T) \to mod(S)$ is* compatible *with* t, *if for all models $\mathcal{M}$ of T, there is some model $\mathcal{M}'$ of T^+ such that:*

- *for all sorts σ from the ground language $\mathcal{L}_T$, $(M^+)_\sigma = M_\sigma$; and*
- *$t^\dagger(\mathcal{M}) = t^*(\mathcal{M}')$.*[37]

Note that $t^\dagger$ takes models of T while t^* takes models of T^+. The idea here is that we restrict our attention to functors $t^\dagger$ that yield models that could have been obtained by taking a pure Morita extension and then applying the mod-functor t^*. Notice, however, that while we have a definition of compatible functors, we are yet to show that any such functors exist. To establish this, we first define what we call the *canonical*

[36] I'm using the term "expansion" rather than "extension" for a couple of reasons. First, I want to make a clear distinction between the concepts defined here and those defined in Barrett and Halvorson [4]. Second, model theorists tend to use "expansion" to refer to the process of adding vocabulary and "extension" for augmenting the domain of a model [12]. A Morita expansion is primarily an addition of vocabulary, so "expansion" seems apropos. That said, Morita expansions also involve adding new domains, so either name could work.

[37] $t^* : mod(T^+) \to mod(S)$ is the function defined in Theorem 4.4.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Morita expansion of a model $\mathcal{M}$ of T. The basic idea is to use the obvious operations on the original sort domains to define the new sort domains.

Definition 5.6. *Suppose $T^+ \subseteq \mathcal{L}_{T^+}$ is a (pure) Morita successor of $T \subseteq \mathcal{L}_T$ and let $\mathcal{M}$ be a model of T. We let the* canonical (Morita) expansion of $\mathcal{M}$, *denoted $\mathcal{M}^+$, be a model of $\mathcal{L}_{T^+}$ defined as follows. For sorts σ from $\mathcal{L}_T$, we let—as required—$(M^+)_\sigma = M_\sigma$. If σ is the new sort in $\mathcal{L}_{T^+} \backslash \mathcal{L}_T$ defined from sorts σ_0, σ_1 from $\mathcal{L}_T$, we define M^+_σ based on how σ is formed:*

- *if σ is a product formed from σ_0, σ_1, let*
$$M^+_\sigma = M_{\sigma_0} \times M_{\sigma_1},$$
$\pi_0^{\mathcal{M}^+}(\langle m_0, m_1 \rangle) = m_0$ *and* $\pi_1^{\mathcal{M}^+}(\langle m_0, m_1 \rangle) = m_1$;
- *if σ is a co-product sort formed from σ_0, σ_1, let*
$$M^+_\sigma = (M_{\sigma_0} \times \{0\}) \cup (M_{\sigma_1} \times \{1\}),$$
$\pi_0^{\mathcal{M}^+}(m) = \langle m, 0 \rangle$ *and* $\pi_1^{\mathcal{M}^+}(m) = \langle m, 1 \rangle$;
- *if σ is a subsort of σ_0 formed using $\varphi(x^{\sigma_0})$, let*
$$M_\sigma = \{m \mid m \in M_{\sigma_0} \ \wedge \ \mathcal{M} \models \varphi(m)\},$$
and $\pi^{\mathcal{M}^+}(m) = m$;
- *if σ is a quotient on σ_0 formed using $\varphi(x^{\sigma_0}, y^{\sigma_0})$, let*
$$M^+_\sigma = \{[m]_\varphi \mid m \in M^k_{\sigma_0}\},$$
and $\pi^{\mathcal{M}^+}(m) = [m]_\varphi$, *where* $[m]_\varphi = \{m_0 \in M_\sigma \mid \mathcal{M} \models \varphi(m, m_0)\}$.

If $T^* \subseteq \mathcal{L}_{T^*}$ is a (pure) Morita expansion of $T \subseteq \mathcal{L}_T$, then the canonical Morita expansion of $\mathcal{M}$, denoted $\mathcal{M}^*$ is formed by taking successive canonical expansions of $\mathcal{M}$ corresponding to the new sorts that are added in the Morita expansion.

We now aim to prove that functors compatible with interpretations exist. To do this, we first recall a technical theorem from [4] that will be used further during this paper. In order for this theorem to make sense, we first need to explain the notion of a *code*. A detailed discussion of this is provided after Example 4.5 in [4], but we'll content ourselves here with a brief overview. Suppose that $T^+ \subseteq \mathcal{L}_{T^+}$ is a Morita successor of $T \subseteq \mathcal{L}_T$ and that σ is the new sort that was added. Despite being a new sort, we can still understand the behavior of objects of sort σ using objects from sorts from the original language $\mathcal{L}_T$. For example, if σ is formed as a product of sorts τ_0 and τ_1, using projection functions π_0 and π_1, then we see that any x^σ is determined by a unique pair $z_0^{\tau_0}$ and $z_0^{\tau_1}$. More precisely, we see that

$$\pi_0(x^\sigma) = z_0^{\tau_0} \wedge \pi_1(x^\sigma) = z_1^{\tau_1}$$

and we call such a formula a *code* for variables of sort σ. A code provides a kind of bridge between the old sorts of $\mathcal{L}_T$ and the new sorts of $\mathcal{L}_{T^+}$. Similarly, a code for each of the other three sort types can also be defined. The following theorem then shows how the behavior of the new sorts can be understood from the perspective of the original language.

Theorem 5.7 [4]. *Suppose $T^+ \subseteq \mathcal{L}_{T^+}$ is a Morita successor of $T \subseteq \mathcal{L}_T$. Suppose $\varphi(x_0^\sigma, \ldots, x_m^\sigma, y_0^{\sigma_0}, \ldots, y_n^{\sigma_n})$ where $\sigma \in \mathcal{L}_{T^+} \backslash \mathcal{L}_T$ and $\sigma_0, \ldots, \sigma_m \in \mathcal{L}_T$. Suppose that for*

$i \leq m$, $\xi_i(x^\sigma, z_0^{\tau_0}, z_1^{\tau_1})$ is a code for variables of sort σ where τ_0 and τ_1 are sorts from $\mathcal{L}_T$. Then there is a formula $\varphi^*(z_{0,0}^{\tau_0}, z_{0,1}^{\tau_1}, \dots, z_m^{\tau_0}, z_{m,1}^{\tau_1}, y_0^{\sigma_0}, \dots, y_n^{\sigma_n})$ of $\mathcal{L}_T$ such that T^+ proves that for all $x_0^\sigma, \dots, x_m^\sigma, y_0^{\sigma_0}, \dots, y_n^{\sigma_n}, z_{0,0}^{\tau_0}, z_{0,1}^{\tau_1}, \dots, z_m^{\tau_0}, z_{m,1}^{\tau_1}$ if $\bigwedge_{i \leq m} \xi_i(x_i^\sigma, z_{i,o}^{\tau_0}, z_{i,1}^{\tau_1})$, then

$$\varphi(x_0^\sigma, \dots, x_m^\sigma, y_0^{\sigma_0}, \dots, y_n^{\sigma_n}) \leftrightarrow \varphi^*(z_{0,0}^{\tau_0}, z_{0,1}^{\tau_1}, \dots, z_m^{\tau_0}, z_{m,1}^{\tau_1}, y_0^{\sigma_0}, \dots, y_n^{\sigma_n}).$$

Informally speaking, we are using the code formulae ξ_i to form a bridge between the formula φ of $\mathcal{L}_{T^+}$ and an equivalent formula φ^* of the original language $\mathcal{L}_T$.[38] A little more formally, we are replacing all the variables of sort σ with pairs of variables of sort τ_0 and τ_1 and then modifying the formula φ to accommodate this change. Given the appropriate coding formulae, we end up with a formula φ^* from the original language, $\mathcal{L}_T$, that is equivalent to φ according to T^+. The upshot of this is that we see that T is capable of simulating what happens in T^+. Thus, it might be argued that the expansion is—in some sense—harmless. We then use this result in the following lemma to establish the existence of compatible functors for Morita expansions.

Lemma 5.8. *If $T^+ \subseteq \mathcal{L}_{T^+}$ is a Morita successor of $T \subseteq \mathcal{L}$ and $j : \mathcal{M} \to \mathcal{N}$ is an elementary embedding between models of T, then there is an elementary embedding $j^+ : \mathcal{M}^+ \to \mathcal{N}^+$ such that for all sorts σ from $\mathcal{L}$, $j_\sigma^+ = j_\sigma$.*

Proof. First we define $j^+ : \mathcal{M}^+ \to \mathcal{N}^+$ where $j_\sigma^+ = j_\sigma$ for all σ from $\mathcal{L}$ and then we show that j^+ is an elementary embedding. If σ is a sort from $\mathcal{L}$ we let $j_\sigma^+ = j_\sigma$. So suppose σ is the new sort from $\mathcal{L}_{T^+} \backslash \mathcal{L}_T$ defined from sorts σ_0, σ_1 from $\mathcal{L}_T$. We we define j_σ^+ depending on how it was formed:

- if σ is a product formed from σ_0, σ_1, and $\langle m_0, m_1 \rangle \in M_\sigma^+$
$$j_\sigma^+(\langle m_0, m_1 \rangle) = \langle j_{\sigma_0}^+(m_0), j_{\sigma_1}^+(m_1) \rangle;$$
- if σ is a co-product sort formed from σ_0, σ_1 and $\langle m, i \rangle \in M_\sigma^+$
$$j_\sigma^+(\langle m, i \rangle) = \begin{cases} \langle j_{\sigma_0}^+(m), i \rangle, & \text{if } i = 0, \\ \langle j_{\sigma_1}^+(m), i \rangle, & \text{if } i = 1; \end{cases}$$
- if σ is a subsort of σ_0 formed using $\varphi(x^{\sigma_0})$ and $m \in M_\sigma^+$
$$j_\sigma^+(m) = j_{\sigma_0}^+(m); \text{ and}$$
- if σ is a quotient on σ_0 formed using $\varphi(x^{\sigma_0}, y^{\sigma_0})$, let
$$j_\sigma^+([m]_\varphi) = [j_{\sigma_0}^+(m)]_\varphi.$$

We then claim that j^+ gives an elementary embedding. Let $\varphi(x_0^\sigma, \dots, x_m^\sigma, y_0^{\sigma_0}, \dots, y_k^{\sigma_k})$ be a formula of $\mathcal{L}_{T^+}$ and $m_0, \dots, m_n, a_0, \dots, a_k$ where $m_0, \dots, m_n$ are from M_σ and $a_0, \dots, a_k$ are from $M_{\sigma_0}, \dots, M_{\sigma_k}$ respectively. We claim that

$$\mathcal{M}^+ \models \varphi(m_0, \dots, m_n, a_0, \dots, a_k) \Leftrightarrow \mathcal{N}^+ \models \varphi(j_\sigma^+(m_0), \dots j_\sigma^+(m_n), j_{\sigma_0}^+(a_0), \dots, j_{\sigma_k}^+(a_k)).$$

We'll just do the case when σ is a co-product sort formed from sorts τ_0, τ_1 from $\mathcal{L}_T$. Then $M_\sigma = (M_{\tau_0} \times \{0\}) \cup (M_{\tau_1} \times \{1\})$. For each $i \leq n$, let $\xi_i(x_i^\sigma, z_{i,0}^{\tau_0}, z_{i,1}^{\tau_1})$ be: the

[38] Note that we may need more than one code formula for a particular sort in order to deal with the case of co-products. This emerges in the proof of Lemma 5.8 and is addressed in more detail in the final part of the proof of Theorem 4.6 in [4].

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

formula $\pi_0(z^{\tau_0}_{i,0}) = x^\sigma_i$ if $m_i = \langle m^*_i, 0\rangle$ for some $m^*_i \in M_{\tau_0}$; and $\pi_1(z^{\tau_1}_{i,1}) = x^\sigma_i$ otherwise. Each of these formulae ξ_i is a code for the sort σ. We then see that for all $i \leq n$, we may fix $c_{i,0} \in M_{\tau_0}$ and $c_{i,1} \in M_{\tau_1}$ such that $\mathcal{M}^+ \models \xi_i(m_i, c_{i,0}, c_{i,1})$. Thus we see that

$$\mathcal{M}^+ \models \bigwedge_{i \leq n} \xi_i(m_i, c_{i,0}, c_{i,1}).$$

Moreover, using the definition of j^+ and it can be seen that $\mathcal{N}^+ \models \bigwedge_{i \leq n} \xi_i(j^+_\sigma(m_i), j_{\tau_0}(c_{i,0}), j_{\tau_1}(c_{i,1}))$.

By the previous lemma we may fix a formula $\varphi^*(z^{\tau_0}_{0,0}, z^{\tau_1}_{0,1}, \dots, z^{\tau_0}_{n,0}, z^{\tau_1}_{n,1}, y^{\sigma_0}_0, \dots, y^{\sigma_k}_k)$ such that T^+ proves that for all $x^\sigma_0, \dots, x^\sigma_m$ and $z^{\tau_0}_{0,0}, z^{\tau_1}_{0,1}, \dots, z^{\tau_0}_{n,0}, z^{\tau_1}_{n,1}, y^{\sigma_0}_0, \dots, y^{\sigma_k}_k$ if $\bigwedge_{i \leq n} \xi_i(x^\sigma_i, z^{\tau_0}_{i,0}, z^{\tau_1}_{i,1})$ then

$$\varphi(x^\sigma_0, \dots, x^\sigma_m, y^{\sigma_0}_0, \dots, y^{\sigma_k}_k) \leftrightarrow \varphi^*(z^{\tau_0}_{0,0}, z^{\tau_1}_{0,1}, \dots, z^{\tau_0}_{n,0}, z^{\tau_1}_{n,1}, y^{\sigma_0}_0, \dots, y^{\sigma_k}_k).$$

Finally, we put this together to see that

$$\begin{aligned} &\mathcal{M}^+ \models \varphi(m_0, \dots, m_n, a_0, \dots, a_k) \\ \Leftrightarrow &\mathcal{M} \models \varphi^*(c_{0,0}, c_{0,1}, \dots, c_{n,0}, c_{n,1}, a_0, \dots, a_k) \\ \Leftrightarrow &\mathcal{N} \models \varphi^*(j_{\tau_0}(c_{0,0}), j_{\tau_1}(c_{0,1}), \dots, j_{\tau_0}(c_{n,0}), j_{\tau_1}(c_{n,1}), j_{\sigma_0}(a_0), \dots, j_{\sigma_k}(a_k)) \\ \Leftrightarrow &\mathcal{N}^+ \models \varphi(j^+_\sigma(m_0), \dots j^+_\sigma(m_n), j^+_{\sigma_0}(a_0), \dots, j^+_{\sigma_k}(a_k)) \end{aligned}$$

as required. □

It is easy to see that Lemma 4.6 gives us the successor case in a proof by induction establish the result above holds for Morita expansions more generally. Finally, we can use this to show that Morita interpretations are compatible with functors between theory categories.

THEOREM 5.9. *Let T and S be theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. Suppose T Morita interprets S via the Morita expansion T^* in $\mathcal{L}_{T^*}$ and translation $t : \mathcal{L}_S \to \mathcal{L}_{T^*}$. Then there is a functor $t^\dagger$ that is compatible with* t.

Proof. First we define $t^\dagger : mod_{mult}(T) \to mod_{mult}(S)$. Give $\mathcal{M}$ a model of T, we let $t^\dagger(\mathcal{M}) = t(\mathcal{M}^*)$ where: $\mathcal{M}^*$ is the canonical expansion of $\mathcal{M}$ to a model of T^* described above; and $t : mod_{mult}(T^*) \to mod_{mult}(S)$. This describes the action of $t^\dagger$ on the objects of $mod_{mult}(T)$. For arrows, suppose that $j : \mathcal{M} \to \mathcal{N}$ is an elementary embedding between models of T. By Lemma 5.8, we may fix $j^* : \mathcal{M}^* \to \mathcal{N}^*$. And by Proposition 4.6 we see that $t(j^*) : t(\mathcal{M}^*) \to t(\mathcal{N}^*)$ is an elementary embedding such that $t(j^*)$ restricted to domains from T is the same as j. Thus we let $t^\dagger(j) = t(j^*)$. It is then easy to see that this function preserves identity arrows and composition. Thus $t^\dagger$ is a function compatible with t as required. □

This puts us in position, analogous to that in Section 3, to define a framework in which Morita interpretation can be understood as another natural restriction of the theory framework based in category theory.

DEFINITION 5.10. *Let $\mathbb{TH}_{Mor}$ be the subcategory of $\mathbb{TH}_{mult}$ where:*

- 1-*morphisms are mod-functors compatible with Morita interpretations*; *and*
- 2-*morphisms are natural transformations given by definitions over the relevant theory.*

And then we may define our standard equivalences in this restricted framework.

Definition 5.11. *Let T and S be theories in $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. Then*:

(1) T and S are definitionally Morita equivalent *if $mod(T)$ and $mod(S)$ are* 1-*isomorphic in* $\mathbb{TH}_{Mor}$;
(2) T and S are bi-Morita-equivalent *if $mod(T)$ and $mod(S)$ are* 1-*equivalent in* $\mathbb{TH}_{Mor}$; *and*
(3) T and S are Morita iso-congruent *if $mod(T)$ and $mod(S)$ are objectively equivalent as witnessed by functors from* $\mathbb{TH}_{Mor}$.

Once again, we see that the category theoretic framework provides a clean housing of a notion of interpretation. However, we must note that Barrett and Halvorson [4] define a different notion, Morita equivalence, which raises the question: where does their notion Morita equivalence fit above? The answer is that it is the same as definitional Morita equivalence and we shall now prove this. As mentioned above, the key issue is whether we use pure or mixed Morita successors. Let us recall the definition of what we shall call standard Morita equivalence.

Definition 5.12 [4]. *Let T and S be theories in multi-sorted language $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively. We say that T and S are* standard Morita equivalent *if T and S have mixed Morita expansions $T^* \subseteq \mathcal{L}_{T^*}$ and $S^* \subseteq \mathcal{L}_{S^*}$ that are definitionally equivalent.*

It is easy to see that T and S are definitionally Morita equivalent if T and S have pure Morita expansions T^* and S^* such that they are definitionally equivalent. One might thus worry that by using a succession of mixed Morita expansions rather than pure ones, we may be able to interpret more and thus obtain more standard Morita equivalences than definitional Morita equivalences. The following lemma show that this worry is misplaced.

Lemma 5.13. *If T^* is a mixed Morita expansion of T, then T^* can be obtained by a single interpretation from a pure Morita expansion T^+ of T.*

Proof. We proceed by induction the Morita successors of T that culminate in T^*. We suppose for our induction hypothesis that we have a sequence $T = T_0^*, T_1^*, \dots, T_n^*$ of mixed Morita successors of T in languages $\mathcal{L} = \mathcal{L}_0^*, \dots, \mathcal{L}_n^*$; and for each $i \leq n$, there is a Morita expansion T_i^+ in $\mathcal{L}_i^+$ such that an interpretation of T_i^+ that is T_i^*. Let T_{n+1}^* be a mixed Morita expansion of T_n^* in the language $\mathcal{L}_{n+1}^*$. We claim that there is a Morita expansion of T_n^+ such that T_{n+1}^* can be obtained by interpretation. Thus, it suffices to show that for the new sort σ added to $\mathcal{L}_{n+1}^*$, this sort can also added as a Morita expansion of T_n^+. First, we note that $\mathcal{L}_n^*$ and $\mathcal{L}_n^+$ must have the same sorts. Thus, it is easy to see that any product or coproduct sort added to $\mathcal{L}_n^*$ could also be added to $\mathcal{L}_n^+$.

Things are a little more difficult with subsorts and quotient sorts since they are defined using a formula of $\mathcal{L}_n^*$ that may contain more relation and function symbols than $\mathcal{L}_n^+$. We focus on the quotient sort case as it is more complex. We suppose that σ is a quotient sort formed from the sort σ_0 from $\mathcal{L}_n^*$ formed using the formula $\varphi(x^{\sigma_0}, y^{\sigma_0})$ from $\mathcal{L}_n^*$. It then suffices to show that there is a formula $\varphi^*(x^{\sigma_0}, y^{\sigma_0})$ such that

$$T_n^* \vdash \forall x^{\sigma_0} \forall y^{\sigma_0} (\varphi(x^{\sigma_0}, y^{\sigma_0}) \leftrightarrow \varphi^*(x^{\sigma_0}, y^{\sigma_0})).$$

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

This follows directly from Theorem 5.7. Thus, T_n^+ is also able to define σ and so we may form T_{n+1}^+ by adding the quotient sort σ using the formula $\varphi^*(x^{\sigma_0}, y^{\sigma_0})$. Then any new relation symbols involving σ can be procured by interpretation as required. □

Putting this together, we see that whenever we have a series of mixed Morita interpretations witnessing standard Morita equivalence, those interpretations can be recovered using a pure Morita expansion following by an interpretation. Thus, we get the following.

COROLLARY 5.14. *T and S are definitionally Morita equivalent if and only if they are standard Morita equivalent.*

Thus, we now see that the equivalence relation of [4] fits very comfortably into the framework provided by this paper. Let us now look at some examples of definitional Morita equivalence.

5.2.1. A simple example. Our first example establishes a well-known kind of redundancy in multi-sorted approaches to theories.

PROPOSITION 5.15. *Any two-sorted theory T in a language with just relation symbols is Morita equivalent to a single sorted theory.*

The proof below is easily adapted to accommodate theories with any finite number of sorts and languages that use function and constant symbols. However, it mostly just makes the notation harder to read and the underlying concepts more difficult to discern.

Proof. Suppose T is articulated in the language $\mathcal{L}_T$ with sorts σ_0, σ_1 and a set $\{R_i\}_{i \in I}$ of relation symbols. Let S be be the theory articulated in the language $\mathcal{L}_S$ that has one sort σ; one-place relation symbols $P_0, \dots, P_n$; and the same set of relation symbols. Before we describe the content of S, it is convenient to define a translation function $s : \mathcal{L}_T \to \mathcal{L}_S$ that works by letting:

- $s(\sigma_i) = \sigma$ for all $i < 2$;
- $s(R_i) = R_i$ for all $i \in I$; and
- $\delta_{s,\sigma_i} = P_i$ for all $i < 2$.

The idea here is that we let the 1-place relations symbols P_i play the role of the sorts σ_i by restricting quantification to P_i where we were once quantifying within σ_i. We then let S be the pointwise image of T via s. Thus, by its definition we see that S interprets T.

We now define the Morita interpretations. First we Morita interpret S in T by taking successive Morita expansions that add subsorts σ_0 and σ_1 defined by the formulae saying P_0x and P_1x respectively. Call the result T^+. We then let the interpretation that follows be the identity. In the other direction we obtain a Morita interpretation of T in S by taking a Morita successor S^+ that adds a coproduct sort σ that combines σ_0 and σ_1 and letting the interpretation again be trivial. Thus, both T^+ and S^+ are theories that both have the same sorts σ, σ_0 and σ_1. Moreover, it is easily seen that these theories are logically equivalent. □

Thus, we have established a sense in which it doesn't matter whether we work in multi-sorted theories or restrict our attention to single-sorted theories. Of course, this doesn't mean that it can't be extremely convenient to work in a multi-sorted theory, but this kind of difference is beyond the scope of the tools developed here.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

5.2.2. A more interesting example. We now employ this framework on an example from Visser and Friedman [22] that we mentioned above. In the context of interpretations that are not identity preserving (and thus, not that of this paper), this example was used to provide a pair of theories that were (non-strictly) bi-interpretable but not definitionally equivalent. In the context of this paper, it makes for a pleasing example of Morita equivalence.

Let AS be the theory in the language $\mathcal{L}_{AS}$ with a single sort ob and a two place relation symbol $\in$. Let it have the following two axioms:

- $\exists x \forall y\ y \notin x$; and
- $\forall x \forall y \exists z \forall w(w \in z \leftrightarrow w \in x \vee w = y)$.

Let ACF be the theory in the language $\mathcal{L}_{ACF}$ with two sorts ob and cl where we use lower case variables for the ob sort and upper case variables for the cl sort. Let the non-logical vocabulary consist of: a relation symbol ε of arity $ob \times cl$; and a function symbol F of sort $ob \to cl$. Let it be axiomatized by the following axioms:

- $\exists X \forall y\ \neg y \varepsilon X$;
- $\forall X \forall y \exists Z \forall w(w \varepsilon Z \leftrightarrow w \varepsilon X \vee w = y)$;
- $\forall X \forall Y(\forall z(z \varepsilon X \leftrightarrow z \varepsilon Y) \to X = Y)$;
- $\forall X \exists y\ F(y) = X$.

PROPOSITION 5.16. *AS is Morita equivalent to ACF.*

Proof. We let AS^+ be a Morita expansion of AS with a quotient sort on ob given by the formula $\varphi(x, y) := \forall z(z \in x \leftrightarrow z \in y)$. Call this new sort σ and let $\pi : ob \to \sigma$ be the associated function. Use upper case letters for σ-variables.

It suffices to show that AS^+ is definitionally equivalent to ACF. Let $t : \mathcal{L}_{ACF} \to \mathcal{L}_{AS^+}$ be such that $t(ob) = ob$ and $t(cl) = \sigma$. Let

$$\begin{aligned} t(x = y) &:= x = y \\ t(X = Y) &:= X = Y \\ t(x \varepsilon Y) &:= \exists z(\pi(z) = Y \wedge x \in z) \\ t(F(x) = Y) &:= \pi(x) = y. \end{aligned}$$

Let $s : \mathcal{L}_{AS^+} \to \mathcal{L}_{ACF}$ be such that $s(ob) = ob$ and $s(\sigma) = cl$. Let

$$\begin{aligned} s(x = y) &:= x = y \\ s(X = Y) &:= X = Y \\ s(x \in y) &:= x \varepsilon F(y) \\ s(\pi(x) = Y) &:= F(x) = Y. \end{aligned}$$

It's easy to see that this gives a definitional equivalence. □

§6. A tableau of interpretability. In this final section, we take some stock and put the work of the preceding sections into a Hasse diagram that is intended to give a clearer idea of the logical landscape. An arrow from one vertex to another indicates that whenever we have two theories satisfying the root position, then those theories also satisfy the relationship of the target position.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

$$\begin{tikzcd}
\text{Definitional Equivalence} \arrow[r] \arrow[d] & \text{Morita Equivalence} \arrow[r] \arrow[d] & \text{Categorical Isomorphism} \arrow[d, leftrightarrow, "\text{Cor 2.9}"] \\
\text{Bi-interpretable} \arrow[r] \arrow[d] & \text{Morita Bi-interpretable} \arrow[r] \arrow[d] & \text{Categorical Equivalence} \arrow[d] \\
\text{Iso-congruence} \arrow[r] \arrow[d] & \text{Morita-congruence} \arrow[r] \arrow[d] & \text{Objective Equivalence} \arrow[d] \\
\text{Mutually Interpretable} \arrow[r] & \text{Morita Mutual} \arrow[r] & \text{Category Mutual}
\end{tikzcd}$$

We include a bottom row in this table to accommodate mutual interpretability and its Morita and categorical counterparts. This will be useful for understanding of the results below that demonstrate that some of these arrows cannot be reversed. Each of the downward implications follow essentially by their definitions. The arrows from positions in the leftmost column to those in the middle column follow since the single sorted cases are clearly instances of their Morita counterparts. The arrows from position in the middle column to those in the rightmost column follow from Theorem 5.9. The only reversal we have obtained was established in Corollary 2.9.

The diagram arguably above provides a richer picture than was available in Barrett and Halvorson [4], which focuses on what we might now think of as the top row. Thus, when they come to consider the question of whether arrows reverse they are only focused on the two arrows in the top row of the diagram. Nonetheless, the proofs of their non-reversal claims deliver strengthened claims in current framework. We sketch these results below, using the following definition and lemma.

Definition 6.1. *Let us say that T is* rigid *if no model of $\mathcal{M}$ of T has an embedding $f : \mathcal{M} \to \mathcal{M}$ that is not the identity function on each of its sort domains.*

Any theory that implies that there is only one object will be rigid since any function between a model of that theory and itself must be the identity. Note also that when $f : \mathcal{M} \to \mathcal{N}$ is an embedding f is an isomorphism between $\mathcal{M}$ and the obvious structure formed from its range. This entails that the composition of functors witnessing strict bi-interpretability or iso-congruence will be an embedding that is also an endomorphism. The following lemma entails us that embeddings between models of some theory have can be uniquely lifted to embeddings of models of its Morita expansions.

Lemma 6.2. *Suppose T^+ is a Morita successor T, $\mathcal{M}$ is a model of T and $f : \mathcal{M} \to \mathcal{M}$ is an embedding. Then for any model $\mathcal{M}^+$ of T^+ extending $\mathcal{M}$ there is a unique embedding $f^+ : \mathcal{M}^+ \to \mathcal{M}^+$ that extends* f.

Proof. The proof is routine so we just verify the result for the case of product sorts. Suppose $f : \mathcal{M} \to \mathcal{M}$ is an embedding. Suppose σ is a product sort formed from sorts σ_0 and σ_1 from $\mathcal{L}_T$, the language of T. Let $f^+ : \mathcal{M}^+ \to \mathcal{M}^+$ be such that $f^+_\tau = f_\tau$ for all sorts τ from $\mathcal{L}_T$. Then for $m \in M_\sigma$ let

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

$$f_\sigma^+(x) = y$$

when: $\pi_0(x) = x_0$, $\pi_1(x) = x_1$, $\pi_0(y) = y_0$, $\pi_1(y) = y_1$ are satisfied in $\mathcal{M}^+$; and $f_{\sigma_0}(x_0) = y_0$ and $f_{\sigma_1}(x_1) = y_1$. Now toward uniqueness suppose $g : \mathcal{M}^+ \to \mathcal{M}^+$ is an embedding that extends f but $g \neq f^+$. Then there must be some x in M_σ such that $f_\sigma^+(x) \neq g_\sigma(x)$. But then g cannot be an embedding. To see this note that we must have either

$$\mathcal{M}^+ \models \pi_0(g_\sigma(x)) \neq g_{\sigma_o}(x_0) \text{ or } \mathcal{M}^+ \models \pi_1(g_\sigma(x)) \neq g_{\sigma_o}(x_1)$$

when we already know that $\mathcal{M}^+ \models \pi_0(x) = x_0$ and $\mathcal{M}^+ \models \pi_1(x) = x_1$. □

Corollary 6.3. *If T is rigid and T^+ is a Morita expansion of T, then $T^+$is also rigid.*

Proof. Let $\mathcal{M}^+$ be a model of T^+ and $\mathcal{M}$ be the retract of $\mathcal{M}^+$back to the language $\mathcal{L}_T$ of T. Let $f^+ : \mathcal{M}^+ \to \mathcal{M}^+$ be an embedding and let $f : \mathcal{M} \to \mathcal{M}$ be the restriction of f back to sorts in $\mathcal{L}_T$. Since T is rigid, we see that f must be the identity. Then using Lemma 5.8, we see that there is a unique embedding extending the identity function and so this must be f^+. Since the identity function is clearly an embedding from $\mathcal{M}^+$to itself, we see that f^+ must be the identity on its sort domains. □

Here then are the results from Barrett and Halvorson [4] establishing that arrows cannot be reversed. We include brief sketches of the proofs of the generalizations of these results to the current framework, however, the underlying machinery remains the same.

Theorem 6.4.

(1) (*Essentially* [4]) *There are first-order theories that are Morita equivalent but not mutually interpretable.*
(2) (*Essentially* [4]) *There are first-order theories that are categorically isomorphic, but not Morita iso-congruent.*

Proof. (1) Let T be the theory saying there is exactly one object in the empty language. Let S be the theory in the language $\mathcal{L}_S = \{R, a, b\}$ where R is a 2-place relation symbol and a, b are constant symbols. Let S say that Rab and that for no x, y do we have Rxy unless $x = a$ and $y = b$. It is easily seen that T cannot interpret S since an interpretation $t : mod(T) \to mod(S)$ would need to give a submodel of a model of T that contains two objects. On the other hand, it can be seen that T will be Morita equivalent to any theory satisfying the conditions set out for S. For T to Morita interpret S, we take a Morita successor T^+ taking the disjoint union of the domain with itself. Then we use the associated embedding functions σ_0, σ_1 to define a relation between $\sigma_0(x)$ and $\sigma_1(x)$ for the only object x. For S to interpret T we take a subsort that picks out one of the objects from the domain. Call this S^+. It can then be seen that these interpretations give us definitional equivalence between $T^+$and S^+ and thus Morita equivalence between T and S.

(2) Let T be the theory that says there is exactly one object in the language $\mathcal{L} = \{P_n, a\}_{n\in\omega}$ where each P_n is a one-place relation symbol and a is a constant symbol. Let S be the theory in $\mathcal{L}$ which extends T with the axioms

$$P_0 a \to P_n a$$

for all $n \in \omega$. Barrett and Halvorson show that these theories are equivalent as categories but not Morita equivalent. Since they are equivalent, we see by Corollary 2.9

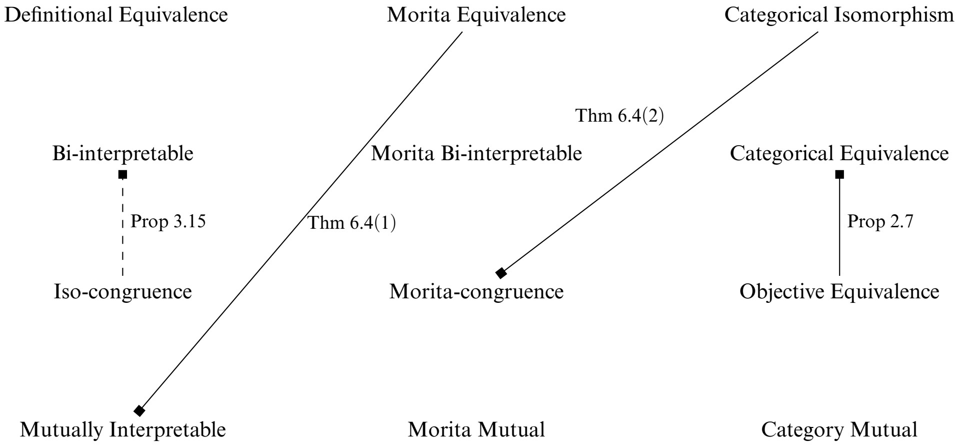
https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

that $mod(T)$ and $mod(S)$ are also isomorphic as categories. To see that T and S are not Morita iso-congruent, we suppose toward a contradiction that there are. Now fix Morita expansions T^+ and S^+ of T and S respectively and interpretations $t : mod(T^+) \leftrightarrow mod(S^+) : s$ witnessing the iso-congruence. Thus we have:

- $s \circ t(\mathcal{M}) \cong \mathcal{M}$ for all models $\mathcal{M}$ of T^+; and
- $t \circ s(\mathcal{N}) \cong \mathcal{N}$ for all models $\mathcal{N}$ of S^+.

Note that T and S are both rigid. This is because both theories demand that there is only one object and thus, any function between a model of T (respectively S) and itself will be the identity. Let $\mathcal{M}$ be a model of $T^+$and let $\mathcal{M}^-$ be the reduct of $\mathcal{M}$ back to the language $\mathcal{L}$. Then note that since the only embedding $f : \mathcal{M}^- \to \mathcal{M}^-$ is the identity, we see by Corollary 6.3 that any embedding between $\mathcal{M}$ and itself witnessing that $\mathcal{M} \cong s \circ t(\mathcal{M})$ must also be the identity. Similarly, any automorphism between $\mathcal{N}$ and $t \circ s(\mathcal{N})$ for a model $\mathcal{N}$ of S must also be the identity. Thus, T^+ and S^+ are definitionally equivalent; and so T and S are Morita equivalent, which we know cannot be the case. □

The table below highlights the failures of reversal that we have described so far. The *square* at the end of the line between two points is intended to indicate that one cannot always infer that theories satisfying the kind of equivalence in the root position also satisfy the equivalence described in the *square* position. Note that since there is a pair of theories that are Morita equivalent but not mutually interpretable entails that this pair of theories is also not iso-congruent, strictly bi-interpretable or definitionally equivalent. We have used a dashed line between iso-congruence and strict bi-interpretability to indicate that this is a partial result that only holds when we restrict our attention to countable models.



6.1. Open questions. We now explore the failure of reversals in the table a little further by addressing the following open questions posed by Barrett and Halvorson [4].[39]

[39] A third open question was also posed by Barrett and Halvorson concerning whether Theorem 6.4(2) can be established using theories with finite vocabularies. We established some partial

Problem 6.5 [4].

(1) *Suppose T is a single sorted theory that is Morita equivalent to PA (respectively ZF). Then is T definitionally equivalent to PA (respectively ZF)?*

(2) *For sufficiently strong single-sorted theories, does Morita equivalence imply definitional equivalence?*

These problems focus on whether the differences between Morita equivalence and definitional equivalence collapse when we consider theories that are sufficiently strong. In other words, they ask whether the relationships along the top of the table reverse. We answer these two questions below.

Problem 6.5(2). We answer this question negatively by providing a theory that is Morita equivalent to ZF but is not definitionally equivalent to ZF. Let ZF_{twin} be ZF with the axiom of extensionality replaced by axiom stating that for any set there is exactly one other set with the same members. More specifically,

$$\forall x \exists! x^* (x \neq x^* \wedge \forall w (w \in x \leftrightarrow w \in x^*)).$$

Thus, we have a strange modification of ZF in which every set has a twin. Note that ZF_{twin} is unable to define an element of any of its models since there will always be a twin. For this reason, we also demand that our axiomatization of ZF includes the axiom of collection instead of replacement. In ZF this makes no difference but in contexts where extensionality fails it is important since genuine class functions are difficult to find.[40]

Theorem 6.6. (1) *ZF_{twin} is Morita equivalent to ZF.*
(2) *ZF_{twin} is not definitionally equivalent to ZF.*

Proof. (1) To Morita interpret ZF_{twin} in ZF add the subsort with domain X defined by transfinite recursion such that

$$\begin{aligned} X_0 &= \{\langle 0, \emptyset\rangle, \langle 1, \emptyset\rangle\} \\ X_{\alpha+1} &= \{\langle 0, y\rangle \mid y \in \mathcal{P}(X_\alpha)\} \cup \\ &\quad \{\langle 1, y\rangle \mid y \in \mathcal{P}(X_\alpha)\} \cup X_\alpha \\ X_\lambda &= \bigcup_{\alpha<\lambda} X_\alpha \text{ for limit } \lambda \end{aligned}$$

and $X = \bigcup_{\alpha \in Ord} X_\alpha$. We then define a relation $\in_X$ on this sort which is such that for $\langle i, x\rangle, \langle j, y\rangle \in X$

$$\langle i, x\rangle \in_X \langle j, y\rangle \;\Leftrightarrow\; x \in y.$$

The idea is to build in the twins by tagging sets with the natural numbers 0 and 1. This is clearly a model of ZF_{twin}. Call this Morita expansion ZF^+. To Morita interpret ZF in ZF_{twin} we define a quotient sort following [19]. In ZF_{twin}, let us say that a set x is *extensional* if whenever y and z have the same members, then $y \in x$ iff $y \in z$. We $\varphi_1(x)$ say that there is a transitive set Y such that $x \in Y$ and every element of Y is

results in this regard by comparing a restricted class of models of ZFC and GBN. However a simpler and complete solution has now been provided by Andréka et al. [1] using techniques based on automorphisms.

[40] For some pathological results in this regard, see [19].

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

extensional. Then let $\varphi_0(x, y)$ say that x and y have the same members. It can then be seen that the quotient sort given by φ_0 over φ_1 gives a model of ZF. Call the expansion ZF^+_{twin}. One can then see that ZF^+_{twin} and ZF^+ are definitionally equivalent and so ZF_{twin} and ZF are Morita equivalent.

(2) We claim that ZF_{twin} cannot even interpret ZF. Suppose toward a contradiction that there is an interpretation $t : \mathcal{L}_\in \to \mathcal{L}_\in$ such that for all $\varphi \in \mathcal{L}_\in$

$$ZF \vdash \varphi \Rightarrow ZF_{twin} \vdash t(\varphi).$$

Now consider the sentence ψ that says there is a unique set that has no members; i.e., $\exists!\, x\forall y\ y \notin x$. t must translate this to a sentence of the form

$$\exists!\, x(\delta_t(x) \wedge \forall y\ \neg t(\in)(x, y)).$$

But this would mean that ZF_{twin} was able to define an empty set which is impossible. □

We also note that this example can be generalized to hold for PA and not just ZF. To see this we first recall a standard result. Let ZF_{fin} be ZF with the axiom of infinity removed and its negation added. Moreover, suppose we use the axiom of set schema of set induction rather than Foundation.[41]

FACT 6.7. *PA and ZF_{fin} are definitionally equivalent.*[42]

It is then easy to see that the proof above made no use of the axiom of infinity, thus there is a theory $ZF_{fin,twin}$ that is Morita equivalent but not definitionally equivalent with ZF_{fin}. Fact 6.7, then entails that PA is also Morita equivalent but not definitionally equivalent with $ZFC_{fin,twin}$.

Problem 6.5(3). We answer the question affirmatively by using a natural condition—*Morita completeness*—suggested by Barrett and Halvorson [4]—that entails that definitional equivalence follows from Morita equivalence. The idea here is that some theories have sufficient expressive strength that they can replicate the effects of Morita expansion by defining objects internally without the need to add new sorts. We shall say that such a theory is *Morita complete*. The following definition is intended to capture this idea.

DEFINITION 6.8. *Let us say that a theory T in a single sorted language $\mathcal{L}_T$ is* Morita complete *if each of the following hold*:

(1) *For all $\varphi_0(x), \varphi_1(x) \in \mathcal{L}_T$ there exist formulae, $\varphi_{0\times 1}(x), \pi_0(x, y)$ and $\pi_1(x, y)$ of $\mathcal{L}_T$ such that T proves π_0 and π_1 represent functions and T proves*

$$\forall x\forall y(\varphi_0(x) \wedge \varphi_1(y) \to \exists!\, z(\varphi_{0\times 1}(z) \wedge \pi_0(z, x) \wedge \pi_1(z, y)).$$

(2) *For all $\varphi_0(x), \varphi_1(x) \in \mathcal{L}_T$ there exist formulae, $\varphi_{0+1}(x), \sigma_0(x, y)$ and $\sigma_1(x, y)$ of $\mathcal{L}_T$ such that T proves σ_0 and σ_1 represent functions and*

[41] Alternatively, we can—equivalently—add an axiom stating that for every set x the transitive closure of x is a set. Without this, the theory is not strictly bi-interpretable with PA. See [8] for a detailed discussion of this issue.

[42] It is well-known that these theories are strictly bi-interpretable. Theorem 3.13 from [22] then gives us definitional equivalence.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

$$T \vdash \forall z(\varphi_{0+1}(z) \rightarrow \exists x((\varphi_0(x) \wedge \sigma_0(x,z)) \vee (\varphi_1(x) \wedge \sigma_1(x,z))) \wedge$$
$$\forall x \forall y \forall z(\varphi_0(x) \wedge \varphi_1(y) \rightarrow \neg(\sigma_0(x,z) \wedge \sigma_1(y,z))).$$

(3) *For all $\varphi_0(x)$ and $\varphi_1(x,y)$ from $\mathcal{L}_T$ if T proves that $\varphi_1$is an equivalence relation on φ_0, then there exist formulae $\varphi_{0/1}(x)$ and $\rho(x,y)$ such that T proves that ρ represents a function and*

$$T \vdash \forall x \forall y \forall z(\varphi_0(x) \wedge \varphi_0(y) \rightarrow (\rho(x,z) \wedge \rho(y,z) \leftrightarrow \varphi_1(x,y))) \wedge$$
$$\forall y(\varphi_0(y) \rightarrow \exists! x(\varphi_{0/1}(x) \wedge \rho(y,x))).$$

Note that we did not include a clause for subsorts above. This is because any theory can replicate the effect of subsorts. Then we observe that ZF is an example of a Morita complete theory.

Proposition 6.9. *ZF is Morita complete.*

Proof. It will suffice to show that (1)–(3) of Definition 6.8 hold. (1) Suppose we have formulae $\varphi_0(x)$ and $\varphi_1(x)$ from the language of set theory. We let $\varphi_{0\times1}(x)$ be the formula $\exists y, z(\varphi_0(y) \wedge \varphi_1(z) \wedge x = \langle x, y\rangle$. Then let $\pi_0(x,y)$ say that there is some z such that $x = \langle y, z\rangle$, and let $\pi_1(x,z)$ say that there is some y such that $x = \langle y, z\rangle$.

(2) Suppose $\varphi_0(x)$ and $\varphi_1(x)$ are formulae of set theory. Let $\varphi_{0+1}(x)$ say that there is some z, i such that $x = \langle z, i\rangle$ and either: $i = 0$ and $\varphi_0(z)$; or $i = 1$ and $\varphi_1(z)$. Let $\sigma_0(x,y)$ say that $y = \langle x, 0\rangle$ and let $\sigma_1(x,y)$ say that $y = \langle x, 1\rangle$.

(3) Suppose that ZF proves that $\varphi_1(x,y)$ describes an equivalent relation on $\varphi_0(x)$. Let $\rho(x,y)$ say that y is the set of those z of least rank such that $\varphi_1(x,z)$.[43] Let $\varphi_{0/1}(x)$ say that there is some z such that $\varphi_0(z)$ and $\rho(z,y)$. □

To establish the result, we prove the following lemma, which is essentially a special case of Lemma 5.7 of [18].

Lemma 6.10. *If T is Morita complete in a single sorted language $\mathcal{L}_T$ and T^+ is a Morita successor of T, then T^+ and T are strictly bi-interpretable.*

Proof. Given that this is essentially a special case of a more general result, we just do the case for products. Suppose T^+ is formed by adding a product sort σ based on subsorts σ_0 and σ_1 from $\mathcal{L}_T$ defined by formulae $\varphi_0(x)$ and $\varphi_1(x)$ respectively. Let $\rho_0 : \sigma \rightarrow \sigma_0$ and $\rho_1 : \sigma \rightarrow \sigma_1$ be the associated projection functions. Then since T is Morita complete we may fix formulae $\varphi_{0\times1}(x)$, $\pi_0(x,y)$ and $\pi_1(x,y)$ from $\mathcal{L}_T$ such that T proves that π_0 and π_1 represent functions and that

$$\forall x \forall y(\varphi_0(x) \wedge \varphi_1(y) \rightarrow \exists! z(\varphi_{0\times1}(z) \wedge \pi_0(z,x) \wedge \pi_1(z,y)).$$

We then define an interpretation $t : \mathcal{L}_{T^+} \rightarrow \mathcal{L}_T$ such that $t(\sigma)$ is sent to the only sort in $\mathcal{L}$ and $\delta_{t,\sigma}(x)$ is $\varphi_{0\times1}(x)$. We then let $t(\rho_0) = \pi_0$ and $t(\rho_1) = \pi_1$. In the other direction we let $t^+ : \mathcal{L}_T \rightarrow \mathcal{L}_{T^+}$ be the trivial interpretation.

Given a model $\mathcal{M}$ of T, it is easy to see that $t^+ \circ t(\mathcal{M}) = \mathcal{M}$. Essentially, t just adds the new sort and then t^+ discards it. From the other direction, suppose $\mathcal{N}^+$ is a model of T^+. We define an isomorphism between $\mathcal{N}^+$ and $t \circ t^+(\mathcal{M})$ with a formula $\psi(x,y)$ of $\mathcal{L}_{T^+}$ as follows. For the single sort of $\mathcal{L}_T$, we just use the identity. For the

[43] This is commonly known as Scott's trick.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

new product sort, we let $\psi(x, y)$ be

$$\exists x_0 \exists x_1 (\rho_0(x) = x_0 \wedge \rho_1(x) = x_1 \wedge \pi_0(y, x_0) \wedge \pi_1(y, x_1)).$$

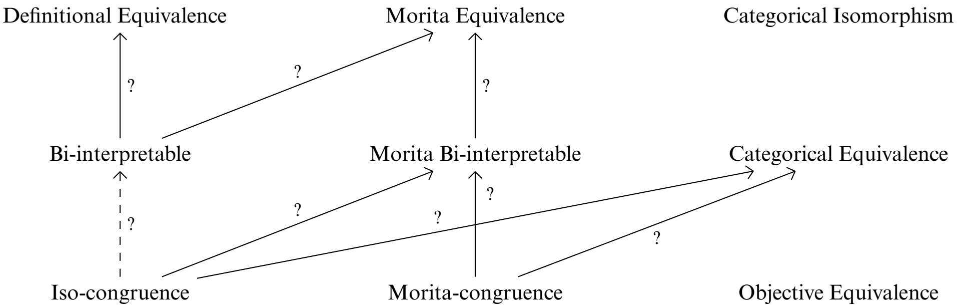
It can then be seen that this defines an isomorphism. □

THEOREM 6.11. *If T and S are sequential Morita-complete theories that are Morita equivalent, then T and S are definitionally equivalent.*

Proof. Suppose T and S are Morita equivalent, Morita complete theories. Suppose that T^* and S^* are the Morita expansions witnessing this. Given that T and S are Morita complete we see by induction on Lemma 6.10 that T and T^* are strictly bi-interpretable and so are S and S^*. Thus since S^* are and T^* are definitionally equivalent, we see that T and S are strictly bi-interpretable. So by Theorem 3.13, we see that T and S are definitionally equivalent. □

Thus, we see that for sufficiently strong theories, the distinction between Morita equivalence and definitional equivalence collapses.

6.1.1. Some remaining questions. While we've come to understand a lot of the relationships between equivalences in our table a number of questions remain. We highlight them in the following table.



§7. Conclusion. In this paper, we have explored the relationship between a number of instruments suitable for establishing equivalence between theories. We have investigated the world beyond relative interpretation and the space between it and natural equivalence relations defined using category theory. In particular, we have developed a general framework that suggests equivalence relations stronger than those in category theory should be thought of as natural restrictions of category theoretic prototypes. We have then provided some classification of this space, answered some open questions and posed a few more.

Appendix.

An equivalent formulation of definitional equivalence.

PROPOSITION 7.1. *Let T and S be articulated in the languages $\mathcal{L}_T$ and $\mathcal{L}_S$ respectively where $\mathcal{L}_S$ and $\mathcal{L}_T$ share no vocabulary. The following are equivalent*:

(1) *T and S are definitionally equivalent*; *and*
(2) *There exist definitional expansions T^+ and S^+ of T and S to $\mathcal{L}_T \cup \mathcal{L}_S$ such that T^+ and S^+ are the same theory.*

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

Before we start the proof, we first make a quick and convenient notational convention. Given a model $\mathcal{M}$ of $\mathcal{L}$ and some sublanguage $\mathcal{L}^*$ of $\mathcal{L}$, we denote the reduct of $\mathcal{M}$ down to $\mathcal{L}^*$ by $\mathcal{M}|\mathcal{L}^*$.

Proof. To make things simpler, suppose that $\mathcal{L}_T$ and $\mathcal{L}_S$ each just contain one relation symbol P and R respectively.

(1→2) Fix mod-functors $t : mod(T) \leftrightarrow mod(S) : s$ witnessing that T and S are definitionally equivalent. It is easy to see that neither δ_t and δ_s alter the domain. Let T^+ be the extension of T in $\mathcal{L}_T \cup \mathcal{L}_S$ be the following axiom:

$$\forall \bar{x}(R\bar{x} \leftrightarrow t(R)(\bar{x})).$$

Similarly, let S^+ be the extension of S in $\mathcal{L}_S \cup \mathcal{L}_T$ with the axiom:

$$\forall \bar{y}(P\bar{y} \leftrightarrow s(P)(\bar{y})).$$

It suffices to then show that T^+ and S^+ have the same models. To see this, we start by letting $\mathcal{M}$ be a model of T^+ and show that is is also a model of S^*. First note that $\mathcal{M}|\mathcal{L}_T$ is a model of T and that $\mathcal{M}|\mathcal{L}_S = t(\mathcal{M}|\mathcal{L}_T)$.

We can then expand $\mathcal{M}|\mathcal{L}_S$ into a model $\mathcal{M}^*$ of S^+ by letting $P^{\mathcal{M}^*}$ be the set of $\bar{m}$ from M such that $(\mathcal{M}|\mathcal{L}_S) \models s(P)(\bar{m})$. Then it can be seen that

$$(\mathcal{M}^*|\mathcal{L}_T) = s(\mathcal{M}^*|\mathcal{L}_S) = s \circ t(\mathcal{M}|\mathcal{L}_T) = \mathcal{M}|\mathcal{L}_T.$$

Putting this together with the fact that $\mathcal{M}^*|\mathcal{L}_S = \mathcal{M}|\mathcal{L}_S$ we see that $\mathcal{M}^* = \mathcal{M}$ and so $\mathcal{M}$ is a model of S^+ as required. A similar argument establishes that every model of S^+ is a model of T^+.

(2→1) Given that T^+ is a definitional expansion of T to $\mathcal{L}_T \cup \mathcal{L}_S$, T^+ contains an axiom of the form

$$\forall \bar{x}(R\bar{x} \leftrightarrow \varphi(\bar{x}))$$

for some formula $\varphi(\bar{x})$ of $\mathcal{L}_T$. We let t be the translation where $t(R)(\bar{x})$ is $\varphi(\bar{x})$ and $\delta_t(x)$ is $x = x$. Similarly, S^+ contains an axiom of the form

$$\forall \bar{y}(P\bar{y} \leftrightarrow \psi(\bar{y}))$$

for some formula $\psi(\bar{y})$ of $\mathcal{L}_S$. We then let s be the translation where $s(P)(\bar{y})$ is $\psi(\bar{y})$ and $\delta_s(y)$ is $y = y$.

We claim that the mod-functors associated with these translations witness that S and T are definitionally equivalent. To see this, let $\mathcal{M}$ be a model of T and let $\mathcal{M}^*$ be the model of T^+ that expands $\mathcal{M}$. Since T^+ and S^+ are the same theory, we see that $\mathcal{M}^*$ is also a model of S^+ and so $\mathcal{M}^*|\mathcal{L}_S$ is a model of S. This tells us that t witnesses that T interprets S. Moreover, it is also clear from our definition of t that $(\mathcal{M}^*|\mathcal{L}_S) = t(\mathcal{M})$. And similarly by our definition of s, we see that $(\mathcal{M}^*|\mathcal{L}_T) = s(\mathcal{M}^*|\mathcal{L}_S)$. Putting this together, we see that

$$\mathcal{M} = \mathcal{M}^*|\mathcal{L}_T = s(\mathcal{M}^*|\mathcal{L}_S) = s \circ t(\mathcal{M}).$$

A similar argument establishes that s witnesses that S interprets T and that $t \circ s(\mathcal{N}) = \mathcal{N}$ for all models $\mathcal{N}$ of S. □

Interpreting theories of dense linear orders. In this section we complete the proof of the remaining claim within the proof of Proposition 3.15. This occurs in Theorem 7.6. We make heavy use of Marker [17]. In particular, we use Theorems 2.4.1 and 3.1.3

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

which establish that $\aleph_0$-categoricity and quantifier elimination hold for theory of dense linear orders. Recall the theories used in the proof of Proposition 3.15:

- Let D be the theory in the language $\mathcal{L}_D = \{<, d_n\}_{n\in\omega}$ which says that $<$ is a dense linear order without end points and that $d_n < d_{n+1}$ for all $n \in \omega$.
- Let B be a theory in the language $\mathcal{L}_B = \{\prec, b_n\}_{n\in\omega}$ that $\prec$ is dense linear order with no top point but with a bottom point that is b_0 and that $b_n < b_{n+1}$ for all $n \in \omega$.

First, we note that unlike the ordinary theory of dense linear orders, D and B are not $\aleph_0$-categorical. They are, however, very close to being so.

Proposition 7.2. *D and B have three models up to isomorphism.*

Proof. We describe three models $\mathcal{D}_0, \mathcal{D}_1$ and $\mathcal{D}_1$ of D each using the rationals $\mathbb{Q}$ under their natural ordering:

(1) $\mathcal{D}_0$ is such that for all $n \in \omega$, $d_n = n$;
(2) $\mathcal{D}_1$ is such that for all $n \in \omega$, $d_n = \frac{1}{2^n}$; and
(3) $\mathcal{D}_2$ is such that for all $n \in \omega$, $d_n = f(n)$ where $f : \omega \to \mathbb{Q}$ is an increasing sequence of rationals converging to $\sqrt{2}$.

The key point here is that the axiomatization of D doesn't specify whether the sequence of constant symbols is cofinal in the ordering. Indeed this is not expressible in this language since we can't internally quantify over the constant symbols. The three possibilities then are that: the constants are cofinal; they are not cofinal and converge to an element of the domain; they are not cofinal and they converge to a hole. Three analogous models are available for B. In both cases, the $\aleph_0$-categoricity of countable dense linear orders ensure that every model of D (or B) will be isomorphic to one of these three models. □

The following result provides a helpful limit on the kinds of relations that are definable in B.

Proposition 7.3. *B has quantifier elimination.*

Proof. The proof here is almost identical to that of Theorem 3.1.3 in [17], so we shall mostly focus on the required changes.

We aim to show that for any formula φ from $\mathcal{L}_B$ with at most n-free variables there is a quantifier free formula ψ from $\mathcal{L}_B$ such that:

$$B \models \forall \bar{x}(\varphi(\bar{x}) \leftrightarrow \psi(\bar{x})).$$

Let φ be in $\mathcal{L}_B$ and note that it just uses a finite set of constant symbols, say $\bar{e} = \{e_0, \dots, e_m\}$, from $\{b_n\}_{n\in\omega}$. Without loss of generality, we'll assume that $e_0 = b_0$ and that $m > 1$. Then note that the obvious restriction of B to the constant symbols from $\bar{e}$ is $\aleph_0$-categorical.

For $\sigma : \{\langle i, j\rangle \mid i < j \leq n\} \to 3$ let $\chi_\sigma(x_0, \dots, x_n)$ be the formula:

$$\bigwedge_{\sigma(\langle i,j\rangle)=0} x_i = x_j \wedge \bigwedge_{\sigma(\langle i,j\rangle)=1} x_i < x_j \wedge \bigwedge_{\sigma(\langle i,j\rangle)=2} x_i > x_j.$$

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

For $\tau : (n+1) \to 2m+2$, let $\delta_\tau(x_0, \dots, x_n)$ be the formula:

$$\bigwedge_{j \leq m} \bigwedge_{\tau(i)=2j} (x_i = e_j) \wedge$$
$$\bigwedge_{j < m} \bigwedge_{\tau(i)=2j+1} (e_j < x_i < e_{j+1}) \wedge$$
$$\bigwedge_{\tau(i)=2m+1} e_m < x_i.$$

Roughly following Marker, we call such pairs $\langle \sigma, \tau \rangle$ *sign conditions* for $\bar{e}$. The idea is that χ_σ captures the configuration of $x_0, \dots, x_n$ with respect to $<$; and for each x_i, $\delta_\tau(x_0, \dots, x_n)$ records whether x_i is identical to one of the constants e_j or within one of the intervals that lie between them or the one that is above them all.

Let $\mathbb{Q}^+$ be the structure of the positive rational numbers with 0. Let $\mathcal{Q} = \langle \mathbb{Q}^+, e_0^{\mathcal{Q}}, \dots, e_m^{\mathcal{Q}} \rangle$ be the expansion of $\mathbb{Q}^+$ where $e_i^{\mathcal{Q}} = i$ for $i \leq m$. Note that $\mathcal{Q}$ is a model of B.

Let Λ_φ be the set of pairs $\langle \sigma, \tau \rangle$ such that there is some $\bar{a} \in \mathbb{Q}$ where $\mathcal{Q} \models \chi_\sigma(\bar{a}) \wedge \delta_\tau(\bar{a}) \wedge \varphi(\bar{a})$. There are two cases to consider.

If $\Lambda_\varphi = \emptyset$. Then $\mathcal{Q} \models \forall \bar{x} \neg \varphi(\bar{x})$. Thus we may let ψ be $x_1 \neq x_1$.

If $\Lambda_\varphi \neq \emptyset$, let

$$\psi(\bar{x}) = \bigvee_{\langle \sigma, \tau \rangle \in \Lambda_\varphi} (\chi_\sigma(\bar{x}) \wedge \delta_\tau(\bar{x})).$$

Clearly we have $\mathcal{Q} \models \varphi(\bar{x}) \to \psi(\bar{x})$. In the other direction, suppose $\bar{b} \in \mathbb{Q}$ and $\mathcal{Q} \models \psi(\bar{x})$. Then we may fix $\langle \sigma, \tau \rangle \in \Lambda_\varphi$ such that $\mathcal{Q} \models \chi_\sigma(\bar{b}) \wedge \delta_\tau(\bar{b})$. By the definition of Λ_φ we may also fix $\bar{a} \in \mathbb{Q}$ such that $\mathcal{Q} \models \chi_\sigma(\bar{a}) \wedge \delta_\tau(\bar{a}) \wedge \varphi(\bar{a})$.

It is then easy to see that whenever $\mathcal{Q} \models \chi_\sigma(\bar{a}) \wedge \delta_\tau(\bar{a})$ and $\mathcal{Q} \models \chi_\sigma(\bar{b}) \wedge \delta_\tau(\bar{b})$, then there is an automorphism f on $\mathcal{Q}$ such that $f(\bar{a}) = \bar{b}$. Thus, $\mathcal{Q} \models \varphi(\bar{b})$ and so $\mathcal{Q} \models \varphi(\bar{b}) \leftrightarrow \psi(\bar{b})$ as required. □

We now describe a condition on particular subsets of sign conditions which are intended to give us formulae for every definable linear order on the entire domain.

We shall use the name *cell* to denote both intervals and constants and—somewhat awkwardly—we'll say that y is in a cell to mean y is either in an interval or identical to a constant. We shall then index the cells using the natural ordering of the cells. This entails that constants are assigned even indices while intervals are assigned odd indices.

$[e_0]$	(e_0, e_1)	$[e_1]$	...	(e_m, ∞)
0	1	2		$2m+1$

Definition 7.4. *Let X be a set of sign conditions for a formula $\varphi(x_0, x_1)$ with constant symbols $\{e_0, \dots, e_m\}$ from $\{d_n\}_{n \in \omega}$ where $e_0 = b_0$ and $m > 1$. We say that X is* proto-linear *if*:

(1) *The cells are totally ordered, so the following conditions are satisfied*:
  (a)*(Transitivity) If $\langle \sigma_0, \tau_0 \rangle, \langle \sigma_1, \tau_1 \rangle \in X$ are such that $\tau_0(0) = j_0$, $\tau_0(1) = \tau_1(0) = j_1$ and $\tau_1(1) = j_2$, then there is some $\langle \sigma, \tau \rangle \in X$ such that $\tau(0) = j_0$ and $\tau(1) = j_1$.*

(b)(*Asymmetry*) *There are no* $\langle \sigma_0, \tau_0 \rangle, \langle \sigma_1, \tau_1 \rangle \in X$ *are such that* $\tau_0(0) = j_0$, $\tau_0(1) = \tau_1(0) = j_1$ *and* $\tau_1(1) = j_0$.

(c)(*Connectedness*) *Every way in which* x_0 *and* x_1 *could be in different cells is addressed by some* $\langle \sigma, \tau \rangle \in X$. *More formally, for all* $j_0 < j_1 \leq 2m + 2$, *there is some* $\langle \sigma, \tau \rangle \in X$ *such that either*: $\tau(0) = j_0$ *and* $\tau(1) = j_1$; *or for some* $i \in 2$, $\tau(i) = j_0$, $\tau(1 - i) = j_1$.

(2) *The points inside any cell are totally ordered. In particular, every interval is ordered by exactly one of either* $>$ *or* $<$. *More formally, for all* $2j + 1 \leq 2m + 2$, *there is some* $\langle \sigma, \tau \rangle \in X$ *such that* $\tau(0) = \tau(1) = 2j + 2$ *and* $\sigma(\langle 0, 1 \rangle) \in \{1, 2\}$, *and if* $\langle \sigma^*, \tau \rangle \in X$, *then* $\sigma^* = \sigma$.

Now we establish that pro-linearity captures what is intended.

Lemma 7.5. *Let* $\varphi(x_0, x_1)$ *be a formula of* $\mathcal{L}_D$. *The following are equivalent*:

(1) $\varphi(x_0, x_1)$ *defines a linear order of the domain of every model of* D; *and*
(2) X_φ *is proto-linear.*

Proof. Let X_φ be the set of sign conditions $\langle \sigma, \tau \rangle$ such that

$$\langle \mathbb{Q}, \bar{e} \rangle \models \chi_\sigma(a_0, a_1) \wedge \delta_\tau(a_0, a_1) \wedge \varphi(a_0, a_1).$$

From the proof of Proposition 7.3, we know that

$$D \vdash \forall x(\varphi(x_0, x_1) \leftrightarrow \bigvee_{\langle \sigma, \tau \rangle \in X_\varphi} (\chi_\sigma(x_0, x_1) \wedge \delta_\tau(x_0, x_1)).$$

(1→2) Contraposing suppose X_φ is not proto-linear. Suppose (1a) fails. Then we get a failure of transitivity. Suppose (1b) fails. Then we may fix $\langle \sigma_0, \tau_0 \rangle, \langle \sigma_1, \tau_1 \rangle \in X_\varphi$ such that $\tau_0(0) = \tau_1(1)$ and $\tau_1(0) = \tau_0(1)$. Fix a in the $\tau_0(0)^{th}$ cell and b from the $\tau_0(1)^{th}$ cell. Then we have both $a \lhd b$ and $b \lhd a$ so we don't have asymmetry. Suppose (1c) fails. Then we get a failure of connectedness. Suppose (2) fails. Then we fail to order a cell.

(2→1) Suppose X_φ is proto-linear. We claim that $\varphi(x_0, x_1)$ does defines a linear order of the domain in any model of D. Without loss of generality, we may assume $\mathcal{Q}$ is such a model.

Let $\lhd = \{\langle q_0, q_1 \rangle \in \mathbb{Q}^+ \mid \mathcal{Q} \models \varphi(q_0, q_1)$.

Then each of the following must hold:

(1) $\lhd$ is transitive;
(2) $\lhd$ is asymmetric; and
(3) $\lhd$ is connected.

(1) Suppose $a \lhd b$ and $b \lhd c$ for some $a, b, c \in \mathbb{Q}^+$. Then (2a) ensures that $a \lhd c$. (2) Suppose $a \lhd b$. Then (2b) ensures that $b \not\lhd a$. (3) Suppose $a, b \in \mathbb{Q}^+$. Suppose a is in the j_0^{th} cell and b is in the j_1^{th} cell. Then by condition (1c) of Definition 7.4 ensures that they are related to each other. Suppose a, b are both in the same cell. Then they must be in an interval and thus connected by (2). □

Theorem 7.6. *D cannot be interpreted in B with a domain preserving interpretation.*

Proof. Suppose that B interprets D with a domain preserving interpretation $t : mod(B) \to mod(D)$ where $t(<)$ is some formula of $\mathcal{L}_B$ using constant symbols among $\{e_0, \ldots, e_m\}$ from $\{b_n\}_{n \in \omega}$ where $e_0 = b_0$. Then Lemma 7.5, tells us that $t(<)$

is equivalent to $\bigvee X_{t(<)}$ and since $t(<)$ defines a linear ordering, we see that $X_{t(<)}$ is proto-linear.

Then it can be seen that $t(<)$ can be construed as a permutation s of the cells where the original ordering within some cells may have been inverted. Thus, we start with cells

$$c_0^1 + c_1^1 + \cdots + c_{2m+1}^1$$

and obtain a new order

$$c_{s(0)}^{i_0} + c_{s(1)}^{i_1} + \cdots + c_{s(2m+1)}^{i_{2m+1}},$$

where $i_l \in \{-1, 1\}$ for $l \leq 2m + 1$.

We claim that if this ordering has no endpoints, it cannot be dense. To see this note that the ordering between the points within the intervals that are cells is not important since these intervals have neither top nor bottom points. Thus we can think of the original ordering being represented by a sequence of length $2m + 1$

$$1010 \ldots 0,$$

where 1's represent constants and 0's represent cells. It is then obvious that there is no permutation of this sequence which has 0's at either end and no adjacent 1's. But the only way for there to be no endpoints is to have 0's at either end and having adjacent 1's violates density. □

BIBLIOGRAPHY

[1] Andréka, H., Madarász, J., Németi, I., & Székely, G. (2023). Testing definitional equivalence of theories via automorphism groups. *The Review of Symbolic Logic*, 1–22. https://doi.org/10.1017/S1755020323000242

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

[3] Awodey, S. (2006). *Category Theory*. Oxford: Clarendon Press.

[4] Barrett, T. W., & Halvorson, H. (2016). Morita equivalence. *Review of Symbolic Logic*, **9**(3), 556–582. https://doi.org/10.1017/S1755020316000186

[5] Bourbaki, N. (1972). Univers. In Artin, M., Grothendieck, A., and Verdier, J.-L., editors. *Séminaire de Géométrie Algébrique du Bois Marie - 1963-64 - Théorie Des Topos et Cohomologie étale Des schémas - (SGA 4) - Vol. 1 (Lecture Notes in Mathematics 269)*. Berlin: Springer, pp. 185–217.

[6] Button, T., & Walsh, S. P. (2018). *Philosophy and Model Theory*. Oxford: Oxford University Press.

[7] Chang, C. C., & Keisler, H. J. (1973). *Model Theory*. Amsterdam: North-Holland.

[8] Enayat, A., Schmerl, J. H., & Visser, A. (2011). ω-models of finite set theory. In Kennedy, J. and Kossak, R., editors. *Set Theory, Arithmetic, and Foundations of Mathematics: Theorems, Philosophies*. Lecture Notes in Logic. Cambridge: Cambridge University Press, pp. 43–65.

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press

[9] Feferman, S., & Kreisel, G. (1969). Set-theoretical foundations of category theory. In *Reports of the Midwest Category Seminar III*, Lecture Notes in Mathematics, 106. Berlin: Springer, pp. 201–247. https://doi.org/10.1007/BFb0059148.

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

[11] Halvorson, H., & Tsementzis, D. (2018). Categories of scientific theories. In Landry, E., editor. *Categories for the Working Philosopher*. Oxford: Oxford University Press, pp. 402–429.

[12] Hodges, W. (1997). *A Shorter Model Theory*. Cambridge: CUP.

[13] Hudetz, L. (2019). Definable categorical equivalence. *Philosophy of Science*, **86**(1), 47–75.

[14] Kelly, G. M., & Street, R. (1974). Review of the elements of 2-categories. In Kelly, G. M., editor. *Category Seminar*. Berlin: Springer, pp. 75–103.

[15] Kunen, K. (2006). *Set Theory: An Introduction to Independence Proofs*. Sydney: Elsevier.

[16] Lefever, K., & Székely, G. (2019). On generalization of definitional equivalence to non-disjoint languages. *Journal of Philosophical Logic*, **48**(4), 709–729.

[17] Marker, D. (2002). *Model Theory: And Introduction*. New York: Springer.

[18] McEldowney, P. A. (2020). On Morita equivalence and interpretability. *Review of Symbolic Logic*, **13**(2), 388–415

[19] Scott, D. (1961). More on the axiom of extensionality. In *Essays on the Foundations of Mathematics*. Jerusalem: Magnes Press, pp. 115–131.

[20] Simpson, S. G. (1999). *Subsystems of Second Order Arithmetic*. Berlin: Springer.

[21] Visser, A. (2006). Categories of theories and interpretations. In Enayat, A., Kalantari, I., and Moniri, M., editors. *Logic in Tehran Proceedings of the Workshop and Conference on Logic, Algebra and Arithmetic*, held October 18–22, 2003, Vol. 26. Wellesley: ASL, pp. 284–341.

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

[23] Weatherall, J. O. (2019). Part 1: Theoretical equivalence in physics. *Philosophy Compass*, **14**(5), e12592.

[24] ———. (2021). Why not categorical equivalence? In Aladova, E., Barceló, P., van Benthem, J., Berger, G., Dannert, K. M., Dewar, N., Diaconescu, R., Düntsch, I., Dzik, W., Kurd-Misto, M. E., Formica, G., Friend, M., Goldblatt, R., Gottlob, G., Grädel, E., Hirsch, R., Hodkinson, I., Jackson, M., Jipsen, P., Maddux, R. D., Manchak, J. B., Orłowska, E., Pieris, A., Plotkin, B., Plotkin, T., Pratt, V. R., Pratt-Hartmann, I., Ahmed, T. S., Weatherall, J. O., Westerståhl, D., Wimberley, J., Wójtowicz, K., and Christian, W., editors. *Hajnal Andréka and István Németi on Unity of Science: From Computing to Relativity Theory Through Algebraic Logic*. Cham: Springer, pp. 427–451.

DEPARTMENT OF LOGIC AND PHILOSOPHY OF SCIENCE
UNIVERSITY OF CALIFORNIA, IRVINE
IRVINE, CA 92617 USA
E-mail: meadowst@uci.edu

https://doi.org/10.1017/S1755020323000321 Published online by Cambridge University Press