7 Semantic Metalogic Redux

In the previous chapter, we covered some of the standard topics in model theory – focusing on those parts we think are of most interest to philosophers. However, the semantic methods of the previous chapter are restricted to the special case of single-sorted logic. In this chapter, we cover semantics for many-sorted logic. But our aim here is not many-sorted logic for its own sake. Indeed, we feel that one first really understands single-sorted logic when one sees it as a special case of many-sorted logic. What's more, even for single-sorted theories, some of the most interesting relations between theories can only be explicated by means of many-sorted methods.

7.1 Structures and Models

For the most part, generalizing semantics to many-sorted logic is straightforward: where a single-sorted structure M has a single domain set, a many-sorted structure has a separate domain set $M(\sigma)$ for each separate sort symbol $\sigma \in \Sigma$. Moreover, if a Σ-formula ϕ has free variables of different sorts, then its extension $M(\phi)$ will be a subset of a Cartesian product of different domains.

DEFINITION 7.1.1 Let Σ be a signature. A **Σ-structure** M is a mapping from Σ to appropriate structures in the category **Sets**. In particular:

- M maps each sort symbol $\sigma \in \Sigma$ to a set $M(\sigma)$.
- M maps each n-ary relation symbol p of sort $\sigma_1 \times \cdots \times \sigma_n$ to a subset $M(p) \subseteq M(\sigma_1) \times \cdots \times M(\sigma_n)$.
- M maps each function symbol f of sort $\sigma_1 \times \cdots \times \sigma_n \rightarrow \sigma_{n+1}$ to a function $M(f) : M(\sigma_1) \times \cdots \times M(\sigma_n) \rightarrow M(\sigma_{n+1})$.

As was the case with single-sorted logic, each Σ-structure M extends to a map, still called M, from Σ-terms to functions, and from Σ-formulas to subsets. In particular:

- For each term t of type $\sigma_1 \times \cdots \times \sigma_n \rightarrow \sigma_{n+1}$,

$$M(t) : M(\sigma_1) \times \cdots \times M(\sigma_n) \rightarrow M(\sigma_{n+1}).$$

- For each formula ϕ of type $\sigma_1 \times \cdots \times \sigma_n$,

$$M(\phi) \subseteq M(\sigma_1) \times \cdots \times M(\sigma_n).$$

DEFINITION 7.1.2 Let M and N be Σ-structures, where Σ has sorts $\sigma_1, \ldots, \sigma_n$. An **elementary embedding** $h : M \to N$ consists of a family $\{h_i \mid \sigma_i \in \Sigma\}$ of functions $h_i : M(\sigma_i) \to N(\sigma_i)$ that preserves the extension of each Σ-formula ϕ.

It's easy to see that the composition of elementary embeddings is an elementary embedding. Thus, for a given theory T, we let $\mathrm{Mod}(T)$ be the category whose objects are models of T and whose arrows are elementary embeddings. Notice that this definition directly generalizes the definition we gave for single-sorted theories; hence, for a single-sorted theory T, there is no ambiguity when we write $\mathrm{Mod}(T)$.

7.2 The Dual Functor to a Translation

Intuitively speaking, a translation $F : T \to T'$ should induce a functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ going in the opposite direction. To see this, recall that models of a theory T' aren't static structures but are more like functors from T' into the category **Sets**. If we think of a model of T' as a functor $M : T' \to \mathbf{Sets}$, then we can precompose this functor with a translation $F : T \to T'$, giving a functor

$$T \overset{F}{\longmapsto} T' \overset{M}{\longmapsto} \mathbf{Sets},$$

i.e., we get a model $F^*M = M \circ F$ of T. However, since M and F are not actually functors, we have to do some work to validate this intuition.

DEFINITION 7.2.1 Let $F : T \to T'$ be a fixed translation. Given an arbitrary model M of T', we define a Σ-structure F^*M as follows:

- For a sort symbol σ of Σ, first consider the set

  $$M(F(\sigma)) = M(F(\sigma)_1) \times \cdots \times M(F(\sigma)_n),$$

  and its subset $M(D_\sigma)$, where D_σ is any one of the domain formulas that F associates with σ. (These domain formulas are all equivalent.) Since F is a translation and M is a model of T', $M(E_\sigma)$ is an equivalence relation on $M(D_\sigma)$. Let $q : M(D_\sigma) \to Y$ be the corresponding quotient map, and let $(F^*M)(\sigma) = Y$.
- For a relation symbol $p : \sigma_1, \ldots, \sigma_n$ of Σ, we define

  $$(F^*M)(p) = (q_1 \times \cdots \times q_n)(M(Fp)),$$

  where $q_i : M(D_{\sigma_i}) \to Y_i$ is the corresponding projection.
- For a function symbol $f : \sigma_1, \ldots, \sigma_n \to \sigma_{n+1}$ of Σ, we define $(F^*M)(f)$ be the function with graph

  $$(q_1 \times \cdots \times q_n \times q_{n+1})(M(Ff)).$$

NOTE 7.2.2 Suppose that $F : T \to T'$ is a translation, and let $\phi(x)$ be a Σ-formula. Then $F(\phi)(\vec{x})$ is compatible with the relation $E(\vec{x}, \vec{y})$ in the following sense: T' implies that if $F(\phi)(\vec{x})$ and $E(\vec{x}, \vec{y})$, then $F(\phi)(\vec{y})$. It follows from this that in any model M of T', the subset $A \equiv M(F(\phi)(\vec{x}))$ of $M(D)$ is compatible with the equivalence relation

$M(E(\vec{x},\vec{y}))$. That is, if $q : M(D) \to Y$ is the quotient map induced by $M(E(\vec{x},\vec{y}))$, then $q^{-1}(q(A)) = A$.

PROPOSITION 7.2.3 *Suppose that $F : T \to T'$ is a translation, and that ϕ is a Σ-formula. Then $(F^*M)(\phi)$ is the image of $M(F(\phi))$ under the corresponding quotient map q.*

Proof To be precise, we prove that the result holds for each Σ-formula ϕ, and context $x_1, \ldots, x_n$ for ϕ. Once a context $x_1, \ldots, x_n$ for ϕ is fixed, we also fix the corresponding context $\vec{x}_1, \ldots, \vec{x}_n$ for $F(\phi)$. Moreover, for a Σ-structure N, we take $N(\phi)$ to mean the extension of ϕ relative to the context $x_1, \ldots, x_n$.

The base case follows immediately from the definition of F^*M. As for inductive cases, we will treat $\wedge$ and $\exists y$ and leave the others to the reader. We simplify notation as follows: let $N = F^*M$, and for each Σ-formula ϕ, let $\phi^* = F(\phi)$.

- Suppose that the result is true for ϕ_1 and ϕ_2, in any of their contexts. Let $x_1, \ldots, x_n$ be a context for $\phi_1 \wedge \phi_2$, hence also for ϕ_1 and ϕ_2. Let $D = D(\vec{x}_1) \wedge \cdots \wedge D(\vec{x}_n)$ be the conjunction of domain formulas for $x_1, \ldots, x_n$; let $E = E(\vec{x}_1, \vec{y}_1) \wedge \cdots \wedge E(\vec{x}_n, \vec{y}_n)$ be the conjunction of the corresponding equivalence relations; and let $q : M(D) \to Y$ be the quotient map determined by $M(E)$. If we let $A_i = M(\phi_i^*) \subseteq M(D)$, then the inductive hypothesis asserts that $N(\phi_i) = q(A_i)$. Since $(\phi_1 \wedge \phi_2)^* = \phi_1^* \wedge \phi_2^*$, it follows that $M((\phi_1 \wedge \phi_2)^*) = A_1 \cap A_2$. Thus,

$$\begin{aligned} q(M(\phi^*)) &= q(A_1 \cap A_2) \\ &= q(A_1) \cap q(A_2) \\ &= N(\phi_1) \cap N(\phi_2) \\ &= N(\phi). \end{aligned}$$

  The second equation follows from the preceding note; the third equation follows from the induction hypothesis; and the final equation by the fact that N is a Σ-structure.
- Suppose that the result is true for ϕ. That is,

$$N(\phi) = (q_1 \times q_2)(M(\phi^*)),$$

  where $q_1 : M(D_1) \to Y_1$ and $q_2 : M(D_2) \to Y_2$ are the quotient maps. We show that the result is also true for $\exists y\phi$. Consider the commuting diagram:

$$\begin{array}{ccc} M(D_1) \times M(D_2) & \xrightarrow{\pi} & M(D_2) \\ \downarrow{\scriptstyle q_1 \times q_2} & & \downarrow{\scriptstyle q_2} \\ Y_1 \times Y_2 & \xrightarrow{\pi} & Y_2 \end{array}$$

  where π is the projection onto the second coordinate. By definition,

$$M((\exists y\phi)^*) = M(\exists \vec{y}(\phi^*)) = \pi^*(M(\phi^*)).$$

Hence,

$$\begin{aligned} N(\exists y\phi) &= \pi(N(\phi)) = \pi((q_1 \times q_2)(M(\phi^*))) \\ &= q_2(\pi^*(M(\phi^*))) = q_2(M((\exists y\phi)^*)). \end{aligned}$$

Thus, the result also holds for $\exists y\phi$. □

PROPOSITION 7.2.4 *Let $F : T \to T'$ be a translation. If M is a model of T', then F^*M is a model of T.*

Proof Let ϕ be a Σ-formula in context $x_1, \ldots, x_n$ such that $T \vdash \phi$. Since $F : T \to T'$ is a translation, $T' \vdash F(\phi)$. If M is a model of T', then

$$M(F(\phi)) \;=\; M(\vec{x}_1, \ldots, \vec{x}_n).$$

By the previous proposition, $(F^*M)(\phi)$ is the image of $M(F(\phi))$ under the quotient map $q : M(D(\vec{x}_1, \ldots, \vec{x}_n)) \to Y$ induced by the equivalence relation $M(E)$, where

$$E \;=\; E(\vec{x}_1, \vec{y}_1) \wedge \cdots \wedge E(\vec{x}_n, \vec{y}_n).$$

Thus, $(F^*M)(\phi) = Y = (F^*M)(x_1, \ldots, x_n)$, and F^*M is a model of T. □

Thus, if $F : T \to T'$ is a translation, then F gives rise to a mapping F^* from the objects of $\mathrm{Mod}(T')$ to the objects of $\mathrm{Mod}(T)$. We now define F^* on the arrows of $\mathrm{Mod}(T')$. Let M and N be models of T', and let $h : M \to N$ be an elementary embedding. Recall that h consists of family $\{h_\sigma \mid \sigma \in \Sigma'\}$ of functions $h_i : M(\sigma) \to N(\sigma)$ that preserves the extension of each Σ'-formula ψ. Now let σ be a sort of Σ, and let $\vec{x}$ be a sequence of Σ'-variables of sort $F(\sigma) = \sigma_1, \ldots, \sigma_n$. Consider the following diagram:

$$\begin{array}{ccc} M(D_\sigma) & \xrightarrow{\;h\;} & N(D_\sigma) \\ \downarrow{\scriptstyle q_M} & & \downarrow{\scriptstyle q_N} \\ (F^*M)(\sigma) & \xrightarrow{\;F^*h\;} & (F^*N)(\sigma) \end{array}$$

where q_M and q_N are the quotient maps induced by $M(E)$ and $N(E)$, respectively, and h is the restriction of $h_1 \times \cdots \times h_n$ to $M(D_\sigma)$, which is well defined since h preserves the extensions of Σ'-formulas. Moreover, if $\langle \vec{a}, \vec{b} \rangle \in M(E)$, then $\langle h(\vec{a}), h(\vec{b}) \rangle \in N(E)$. Thus, h determines a unique function $F^*h : (F^*M)_\sigma \to (F^*N)_\sigma$ such that the previous diagram commutes.

Now let ϕ be a Σ-formula. Then $a \in (F^*M)(\phi)$ iff $a = q_M(b)$, for some $b \in M(F(\phi))$. Moreover, $b \in M(F(\phi))$ iff $h(b) \in N(F(\phi))$ iff $q_N(h(b)) \in (F^*N)(\phi)$. This shows that $a \in (F^*M)(\phi)$ iff $(F^*h)(a) \in (F^*N)(\phi)$. Therefore, F^*h is an elementary embedding.

It is easy to see that F^* preserves composition of elementary embeddings, as well as identity morphisms of models. Therefore, F^* is a functor from $\mathrm{Mod}(T')$ to $\mathrm{Mod}(T)$.

DISCUSSION 7.2.5 It is tempting to think that any functor $G : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ corresponds to some potentially interesting relationship between the theories T and T'. However, functors of the form $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$, where $F : T \to T'$ is a translation, seem to be singled out by the fact that they preserve important theoretical structures. First, the functor F^* is "definable" in the sense that the resulting model F^*M is defined in terms of the model M, and in a uniform fashion. That is, the "same definition" of the new model works, regardless of which model we plug into F^*. (For more on the notion of definable functors, see Hudetz [2018a].) Second, in the case of propositional theories, a functor $G : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is simply a function from the Stone space X' of T' to the Stone space X of T, and functors of the form F^* are precisely the continuous functions.

We now have the tools we need to do some work with the notion of Morita equivalence. We'll first show how similar Morita equivalence is to its poorer cousin, definitional equivalence. In particular, we'll show that Morita equivalent theories have equivalent categories of models.

7.3 Morita Equivalence Implies Categorical Equivalence

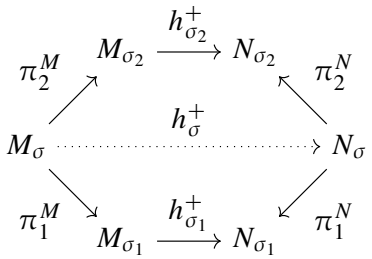
As with a definitional extension, a Morita extension T^+ should "say nothing more" than the original theory T. We will make this idea precise by proving three results about the relationship between $\mathrm{Mod}(T^+)$ and $\mathrm{Mod}(T)$. First, the models of T^+ are "determined" by the models of T.

THEOREM 7.3.1 (Barrett) *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. If T^+ is a Morita extension of T to Σ^+, then every model M of T has a unique expansion (up to isomorphism) M^+ that is a model of T^+.*

Before proving this result, we introduce some notation and prove a lemma. Suppose that a Σ^+-theory T^+ is a Morita extension of a Σ-theory T. Let M and N be models of T^+ with $h : M|_\Sigma \to N|_\Sigma$ an elementary embedding between the Σ-structures $M|_\Sigma$ and $N|_\Sigma$. The elementary embedding h naturally induces a map $h^+ : M \to N$ between the Σ^+-structures M and N.

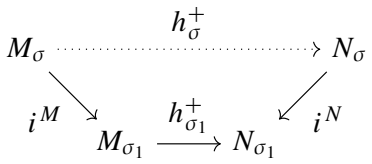
We know that h is a family of maps $h_\sigma : M_\sigma \to N_\sigma$ for each sort $\sigma \in \Sigma$. (Here we have used M_σ for the domain $M(\sigma)$ that M assigns to the sort symbol σ, and similarly for N_σ.) In order to describe h^+, we need to describe the map $h^+_\sigma : M_\sigma \to N_\sigma$ for each sort $\sigma \in \Sigma^+$. If $\sigma \in \Sigma$, we simply let $h^+_\sigma = h_\sigma$. On the other hand, when $\sigma \in \Sigma^+ \backslash \Sigma$, there are four cases to consider. We describe h^+_σ, in the cases where the theory T^+ defines σ as a product sort or a subsort. The coproduct and quotient sort cases are described analogously.

First, suppose that T^+ defines σ as a product sort. Let $\pi_1, \pi_2 \in \Sigma^+$ be the projections of arity $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$ with $\sigma_1, \sigma_2 \in \Sigma$. The definition of the function h^+_σ is suggested by the following diagram.



Let $m \in M_\sigma$. We define $h^+_\sigma(m)$ to be the unique $n \in N_\sigma$ that satisfies both $\pi^N_1(n) = h^+_{\sigma_1} \circ \pi^M_1(m)$ and $\pi^N_2(n) = h^+_{\sigma_2} \circ \pi^M_2(m)$. We know that such an n exists and is unique because N is a model of T^+ and T^+ defines the symbols σ, π_1, and π_2 to be a product sort. One can verify that this definition of h^+_σ makes the preceding diagram commute.

Suppose, on the other hand, that T^+ defines σ as the subsort of "elements of sort σ_1 that are ϕ." Let $i \in \Sigma^+$ be the inclusion map of arity $\sigma \to \sigma_1$ with $\sigma_1 \in \Sigma$. As before, the definition of h^+_σ is suggested by the following diagram.



Let $m \in M_\sigma$. We see that following implications hold:

$$M \vDash \phi[i^M(m)] \Rightarrow M|_\Sigma \vDash \phi[i^M(m)]$$
$$\Rightarrow N|_\Sigma \vDash \phi[h^+_{\sigma_1}(i^M(m))] \Rightarrow N \vDash \phi[h^+_{\sigma_1}(i^M(m))]$$

The first and third implications hold since $\phi(x)$ is a Σ-formula, and the second holds because $h_{\sigma_1} = h^+_{\sigma_1}$ and h is an elementary embedding. T^+ defines the symbols i and σ as a subsort and M is a model of T^+, so it must be that $M \vDash \phi[i^M(m)]$. By the preceding implications, we see that $N \vDash \phi[h^+_{\sigma_1}(i^M(m))]$. Since N is also a model of T^+, there is a unique $n \in N_\sigma$ that satisfies $i^N(n) = h^+_{\sigma_1}(i^M(m))$. We define $h^+_\sigma(m) = n$. This definition of h^+_σ again makes the diagram commute.

When T^+ defines σ as a coproduct sort or a quotient sort one describes the map h^+_σ analogously. We leave it to the reader to work out the details of these cases.

For the purposes of proving Theorem 7.3.1, we need the following simple lemma about this map h^+.

LEMMA 7.3.2 *If $h : M|_\Sigma \to N|_\Sigma$ is an isomorphism, then $h^+ : M \to N$ is an isomorphism.*

Proof We know that $h_\sigma : M_\sigma \to N_\sigma$ is a bijection for each $\sigma \in \Sigma$. Using this fact and the definition of h^+, one can verify that $h^+_\sigma : M_\sigma \to N_\sigma$ is a bijection for each sort $\sigma \in \Sigma^+$. So h^+ is a family of bijections. And furthermore, the commutativity of the preceding diagrams implies that h^+ preserves any function symbols that are used to define new sorts.

It only remains to check that h^+ preserves predicates, functions, and constants that have arities and sorts in Σ. Since $h : M|_\Sigma \to N|_\Sigma$ is a isomorphism, we know that h^+ preserves the symbols in Σ. So let $p \in \Sigma^+\backslash\Sigma$ be a predicate symbol of arity $\sigma_1 \times \ldots \times \sigma_n$ with $\sigma_1, \ldots, \sigma_n \in \Sigma$. There must be a Σ-formula $\phi(x_1, \ldots, x_n)$ such that $T^+ \vDash \forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n (p(x_1, \ldots, x_n) \leftrightarrow \phi(x_1, \ldots, x_n))$. We know that $h : M|_\Sigma \to N|_\Sigma$ is an elementary embedding, so in particular it preserves the formula $\phi(x_1, \ldots, x_n)$. This implies that $(m_1, \ldots, m_n) \in p^M$ if and only if $(h_{\sigma_1}(m_1), \ldots, h_{\sigma_n}(m_n)) \in p^N$. Since $h^+_{\sigma_i} = h_{\sigma_i}$ for each $i = 1, \ldots, n$, it must be that h^+ also preserves the predicate p. An analogous argument demonstrates that h^+ preserves functions and constants. □

We now turn to the proof of Theorem 7.3.1.

Proof of Theorem 7.3.1 Let M be a model of T. First note that if M^+ exists, then it is unique up to isomorphism. For if N is a model of T^+ with $N|_\Sigma = M$, then by letting h be the identity map (which is an isomorphism) Lemma 7.3.2 implies that $M^+ \cong N$. We need only to define the Σ^+-structure M^+. To guarantee that M^+ is an expansion of M we interpret every symbol in Σ the same way that M does. We need to say how the symbols in $\Sigma^+\backslash\Sigma$ are interpreted. There are a number of cases to consider.

Suppose that $p \in \Sigma^+\backslash\Sigma$ is a predicate symbol of arity $\sigma_1 \times \ldots \times \sigma_n$ with $\sigma_1, \ldots, \sigma_n \in \Sigma$. There must be a Σ-formula $\phi(x_1, \ldots, x_n)$ such that $T^+ \vDash \forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n (p(x_1, \ldots, x_n) \leftrightarrow \phi(x_1, \ldots, x_n))$. We define the interpretation of the symbol p in M^+ by letting $M^+(p) = M(\phi)$. Obviously this definition implies that $M^+ \vDash \delta_p$. The cases of function and constant symbols are handled similarly.

Let $\sigma \in \Sigma^+\backslash\Sigma$ be a sort symbol. We describe the cases where T^+ defines σ as a product sort or a subsort. The coproduct and quotient sort cases follow analogously. Suppose first that σ is defined as a product sort with π_1 and π_2 the projections of arity $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$, respectively. We define $M^+_\sigma = M^+_{\sigma_1} \times M^+_{\sigma_2}$ with $\pi_1^{M^+} : M^+_\sigma \to M^+_{\sigma_1}$ and $\pi_2^{M^+} : M^+_\sigma \to M^+_{\sigma_2}$ the canonical projections. One can easily verify that $M^+ \vDash \delta_\sigma$. On the other hand, suppose that σ is defined as a subsort with defining Σ-formula $\phi(x)$ and inclusion i of arity $\sigma \to \sigma_1$. We define $M^+_\sigma = M(\phi) \subseteq M_{\sigma_1}$ with $i^{M^+} : M^+_\sigma \to M^+_{\sigma_1}$ the inclusion map. One can again verify that $M^+ \vDash \delta_\sigma$. □

The previous result immediately yields an important corollary:

THEOREM 7.3.3 (Barrett) *If T^+ is a Morita extension of T, then T^+ is a conservative extension of T.*

Proof Suppose that T^+ is not a conservative extension of T. One can easily see that $T \vdash \phi$ implies that $T^+ \vdash \phi$ for every Σ-sentence ϕ. So there must be some Σ-sentence ϕ such that $T^+ \vdash \phi$, but $T \nvdash \phi$. This implies that there is a model M of T such that $M \vDash \neg\phi$. This model M has no expansion that is a model of T^+ since $T^+ \vdash \phi$, contradicting Theorem 7.3.1. □

Theorems 7.3.1 and 7.3.3 are natural generalizations from definition extensions to Morita extensions. In the case that T^+ is a definitional extension of T, there are natural

maps $I : T \to T^+$ and $R : T^+ \to T$ that form a homotopy equivalence. We now define a reduction map $R : T^+ \to T$ for the case where T^+ is a Morita extension of T.

Lemma 4.6.11 shows that if T^+ is a definitional extension of T to Σ^+, then for every Σ^+-formula ϕ there is a corresponding Σ-formula $R\phi$ such that $T^+ \vdash \phi \leftrightarrow R\phi$. The following example demonstrates that this result does not generalize to the case of Morita extensions in a perfectly straightforward manner.

Example 7.3.4 Recall the theories T and T^+ from Example 5.2.3, and consider the Σ^+-formula $\phi(x, z)$ defined by $i(z) = x$. One can easily see that there is no Σ-formula $\phi^*(x, z)$ that is equivalent to $\phi(x, z)$ according to the theory T^+. Indeed, the variable z cannot appear in any Σ-formula since it is of sort $\sigma^+ \in \Sigma^+ \backslash \Sigma$. A Σ-formula simply cannot say how variables with sorts in Σ relate to variables with sorts in Σ^+. ⌟

In order to define $R : T^+ \to T$, therefore, we need a way of specifying how variables with sorts in $\Sigma^+ \backslash \Sigma$ relate to variables with sorts in Σ. We do this by defining the concept of a "code" (see Szczerba, 1977).

DEFINITION 7.3.5 Let $\Sigma \subseteq \Sigma^+$ be signatures with T a Σ-theory and T^+ a Morita extension of T to Σ^+. We define a **code** formula $\xi(x, y_1, y_2)$ for each variable x of sort $\sigma \in \Sigma^+ \backslash \Sigma$ as follows:

- Suppose that T^+ defines σ as a product sort with π_1 and π_2 the corresponding projections. Then $\xi(x, y_1, y_2)$ is the Σ^+-formula

$$(y_1 = \pi_1(x)) \wedge (y_2 = \pi_2(x)).$$

- Suppose that T^+ defines σ as a coproduct sort with corresponding function symbols $\rho_1 : \sigma_1 \to \sigma$ and $\rho_2 : \sigma_2 \to \sigma$. Then $\xi(x, y_1, y_2)$ is either the Σ^+-formula $\rho_1(y_1) = x$ or the Σ^+-formula $\rho_2(y_2) = x$, where y_i is a variable of sort σ_i. (Note: $\xi(x, y_1, y_2)$ is *not* the disjunction of these two formulas.)
- Suppose that T^+ defines σ as a subsort with $i : \sigma \to \sigma'$ the corresponding function symbol. Then $\xi(x, y)$ is the formula $i(x) = y$, where y is a variable of sort $\sigma' \in \Sigma$.
- Suppose that T^+ defines σ as a quotient sort with $\epsilon : \sigma' \to \sigma$ the corresponding function symbol. Then $\xi(x, y)$ is the Σ^+-formula $\epsilon(y) = x$, where y is again a variable of sort $\sigma' \in \Sigma$.
- Given the empty sequence of variables, we let the **empty code** be the tautology $\exists x(x =_\sigma x)$, where $\sigma \in \Sigma$ is a sort symbol.

Given the conjuncts $\xi_1, \ldots, \xi_n$, we will use the notation $\xi(x_1, \ldots, y_{n2})$ to denote the code $\xi_1(x_1, y_{11}, y_{12}) \wedge \ldots \wedge \xi_n(x_n, y_{n1}, y_{n2})$ for the variables $x_1, \ldots, x_n$. Note that the variables y_{i1} and y_{i2} have sorts in Σ for each $i = 1, \ldots, n$. One should think of a code $\xi(x_1, \ldots, y_{n2})$ for $x_1, \ldots, x_n$ as encoding one way that the variables $x_1, \ldots, x_n$ with sorts in $\Sigma^+ \backslash \Sigma$ might be related to variables $y_{11}, \ldots, y_{n2}$ that have sorts in Σ. One additional piece of notation will be useful in what follows. Given a Σ^+-formula

ϕ, we will write $\phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ to indicate that the variables $x_1, \ldots, x_n$ have sorts $\sigma_1, \ldots, \sigma_n \in \Sigma^+ \backslash \Sigma$ and that the variables $\overline{x}_1, \ldots, \overline{x}_m$ have sorts $\overline{\sigma}_1, \ldots, \overline{\sigma}_m \in \Sigma$.

LEMMA 7.3.6 (Functionality of codes) *Let T be a Σ-theory and T^+ a Morita extension of T to the signature Σ^+. Let $\vec{x}, \vec{z}$ be n-tuples of variables of the same sorts in Σ^+ and let $\xi(\vec{x}, \vec{y})$ be a code for $\vec{x}$. Then we have*

$$T^+ \vdash (\xi(\vec{x}, \vec{y}) \wedge \xi(\vec{z}, \vec{y})) \rightarrow \vec{x} = \vec{z},$$

where $\vec{x} = \vec{z}$ is shorthand for $(x_1 =_{\sigma_1} z_1) \wedge \cdots \wedge (x_n =_{\sigma_n} z_n)$.

Proof This fact follows immediately from the definition of codes. □

We can now state our generalization of Lemma 4.6.11.

THEOREM 7.3.7 (Barrett) *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. Suppose that T^+ is a Morita extension of T to Σ^+ and that $\phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ is a Σ^+-formula. Then for every code $\xi(x_1, \ldots, y_{n2})$ for the variables $x_1, \ldots, x_n$ there is a Σ-formula $\phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})$ such that T^+ entails*

$$\xi(x_1, \ldots, y_{n2}) \rightarrow (\phi(x_1, \ldots x_n, \overline{x}_1, \ldots, \overline{x}_m) \leftrightarrow \phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})).$$

The idea behind Theorem 7.3.7 is simple. Although one might not initially be able to translate a Σ^+-formula ϕ into an equivalent Σ-formula ϕ^*, such a translation is possible after one specifies how the variables in ϕ with sorts in $\Sigma^+ \backslash \Sigma$ are related to variables with sorts in Σ.

We first prove the following lemma. Given a Σ^+-term t, we will again write $t(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ to indicate that the variables $x_1, \ldots, x_n$ have sorts $\sigma_1, \ldots, \sigma_n \in \Sigma^+ \backslash \Sigma$ and that the variables $\overline{x}_1, \ldots, \overline{x}_m$ have sorts $\overline{\sigma}_1, \ldots, \overline{\sigma}_m \in \Sigma$.

LEMMA 7.3.8 *Let $t(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ be a Σ^+-term of sort σ and x a variable of sort σ. Let $\xi(x, x_1, \ldots, x_n, y_1, y_2, y_{11}, \ldots, y_{n2})$ be a code for the variables $x, x_1, \ldots, x_n$, where the variables y_1 and y_2 are used for coding the variable x. Then there is a Σ-formula $\phi_t(x, \overline{x}_1, \ldots, \overline{x}_m, y_{01}, \ldots, y_{n2})$ such that T^+ implies*

$$\xi(x, \ldots, y_{n2}) \rightarrow \big(t(x_1, \ldots, \overline{x}_m) = x \leftrightarrow \phi_t(x, \overline{x}_1, \ldots, \overline{x}_m, y_1, \ldots, y_{n2})\big).$$

If $\sigma \in \Sigma$, then x will not appear in the code ξ. If $\sigma \in \Sigma^+ \backslash \Sigma$, then x will not appear in the Σ-formula ϕ_t.

Proof We induct on the complexity of t. First, suppose that t is a variable x_i of sort σ. If $\sigma \in \Sigma$, then there are no variables in t with sorts in $\Sigma^+ \backslash \Sigma$. So ξ must be the empty code. Let $\phi_t(x, x_i)$ be the Σ-formula $x = x_i$. This choice of ϕ_t trivially satisfies the desired property. If $\sigma \in \Sigma^+ \backslash \Sigma$, then there are four cases to consider. We consider the cases where σ is a product sort and a subsort. The coproduct and quotient cases follow

analogously. Suppose that T^+ defines σ as a product sort with projections π_1 and π_2 of arity $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$. A code ξ for the variables x and x_i must therefore be the formula

$$\pi_1(x) = y_1 \wedge \pi_2(x) = y_2 \wedge \pi_1(x_i) = y_{i1} \wedge \pi_2(x_i) = y_{i2}.$$

One defines the Σ-formula ϕ_t to be $y_1 = y_{i1} \wedge y_2 = y_{i2}$ and verifies that it satisfies the desired property. On the other hand, suppose that T^+ defines σ as a subsort with injection i of arity $\sigma \to \sigma_1$. A code ξ for the variables x and x_i is therefore the formula

$$i(x) = y \wedge i(x_i) = y_{i1}.$$

Let ϕ_t be the Σ-formula $y = y_{i1}$. The desired property again holds.

Second, suppose that t is the constant symbol c. Note that it must be the case that c is of sort $\sigma \in \Sigma$. If $c \in \Sigma$, then letting ϕ_t be the Σ-formula $x = c$ trivially yields the result. If $c \in \Sigma^+ \backslash \Sigma$, then there is some Σ-formula $\psi(x)$ that T^+ uses to explicitly define c. Letting $\phi_t = \psi$ yields the desired result.

For the third (and final) step of the induction, we suppose that t is a term of the form

$$f\big(t_1(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m), \ldots, t_k(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)\big),$$

where $f \in \Sigma^+$ is a function symbol. We show that the result holds for t if it holds for all of the terms $t_1, \ldots, t_k$. There are three cases to consider. First, if $f \in \Sigma$, then it must be that f has arity $\sigma_1 \times \ldots \times \sigma_k \to \sigma$, where $\sigma, \sigma_1, \ldots, \sigma_k \in \Sigma$. Let ξ be a code for $x_1, \ldots, x_n$. We define ϕ_t to be the Σ-formula

$$\exists_{\sigma_1} z_1 \ldots \exists_{\sigma_k} z_k \big(\phi_{t_1}(z_1, \overline{x}_1, \ldots, y_{n2}) \wedge \ldots \wedge \phi_{t_k}(z_k, \overline{x}_1, \ldots y_{n2}) \wedge f(z_1, \ldots, z_k) = x\big),$$

where each of the ϕ_{t_i} exists by our inductive hypothesis. One can verify that ϕ_t satisfies the desired property. Second, if $f \in \Sigma^+ \backslash \Sigma$ is defined by a Σ-formula $\psi(z_1, \ldots, z_k, x)$, then one defines ϕ_t in an analogous manner to above. (Note that, in this case, the arity of f is again $\sigma_1 \times \ldots \times \sigma_k \to \sigma$ with $\sigma_1, \ldots, \sigma_k, \sigma \in \Sigma$.)

Third, we need to verify that the result holds if f is a function symbol that is used in the definition of a new sort. We discuss the cases where f is π_1 and where f is ϵ. Suppose that f is π_1 with arity $\sigma \to \sigma_1$. Then it must be that the term t_1 is a variable x_i of sort σ since there are no other Σ^+-terms of sort σ. So the term t is $\pi_1(x_i)$. Let $\xi(x_i, y_{i1}, y_{i2})$ be a code for x_i. It must be that ξ is the formula

$$\pi_1(x_i) = y_{i1} \wedge \pi_2(x_i) = y_{i2}.$$

Letting ϕ_t be the formula $y_{i1} = x$ yields the desired result. On the other hand, suppose that f is the function symbol ϵ of arity $\sigma_1 \to \sigma$, where σ is a quotient sort defined by the Σ-formula $\psi(z_1, z_2)$. The term t in this case is $\epsilon(t_1(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m))$, and we assume that the result holds for the Σ^+-term t_1 of sort $\sigma_1 \in \Sigma$. Let ξ be a code for the variables $x, x_1, \ldots, x_n$. This code determines a code $\overline{\xi}$ for the variables $x_1, \ldots, x_n$ by "forgetting" the conjunct $\epsilon(y) = x$ that involves the variable x. We use the code $\overline{\xi}$ and the inductive hypothesis to obtain the formula ϕ_{t_1}. Then we define ϕ_t to be the Σ-formula

$$\exists_{\sigma_1} z\big(\phi_{t_1}(z, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}) \wedge \psi(y, z)\big).$$

Considering the original code ξ, one verifies that the result holds for ϕ_{t_1}. □

We now turn to the proof of the main result.

Proof We induct on the complexity of ϕ. Suppose that ϕ is the formula $t(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m) = s(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$, where t and s are Σ^+-terms of sort σ. Let $\xi(x_1, \ldots, y_{n2})$ be a code for $x_1, \ldots, x_n$, and let x be a variable of sort σ. By Lemma 7.3.8, there are corresponding Σ-formulas $\phi_t(x, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})$ and $\phi_s(x, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})$. The Σ-formula ϕ^* is then defined to be

$$\exists_\sigma x\big(\phi_t(x, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}) \wedge \phi_s(x, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})\big).$$

One can verify that this definition of ϕ^* satisfies the desired result.

If t and s are of sort $\sigma \in \Sigma^+ \backslash \Sigma$, then there are four cases to consider. We show that the result holds when T^+ defines σ as a product sort or a quotient sort. The coproduct and subsort cases follow analogously. If T^+ defines σ as a product sort with projections π_1 and π_2 of arity $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$, then we define a code $\overline{\xi}(x, x_1, \ldots, y_{n2}, v_1, v_2)$ for the variables $x, x_1, \ldots, x_n$ by

$$\xi(x_1, \ldots, y_{n2}) \wedge \pi_1(x) = v_1 \wedge \pi_2(x) = v_2.$$

Lemma 7.3.8 and the code $\overline{\xi}$ for the variables $x, x_1, \ldots, x_n$ generate the Σ-formulas $\phi_t(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2)$ and $\phi_s(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2)$. We then define the Σ-formula ϕ^* to be

$$\begin{aligned}\exists_{\sigma_1} v_1 \exists_{\sigma_2} v_2\big(&\phi_t(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2) \\ &\wedge \phi_s(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2)\big).\end{aligned}$$

One can verify that ϕ^* again satisfies the desired result.

If T^+ defines σ as a quotient sort with projection ϵ of arity $\sigma_1 \to \sigma$, then we again define a new code $\overline{\xi}(x, x_1, \ldots, y_{n2}, v)$ for the variables $x, x_1, \ldots, x_n$ by

$$\xi(x_1, \ldots, y_{n2}) \wedge \epsilon(v) = x,$$

Lemma 7.3.8 and the code $\overline{\xi}$ for the variables $x, x_1, \ldots, x_n$ again generate the Σ-formulas $\phi_t(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v)$ and $\phi_s(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v)$. We define the Σ-formula ϕ^* to be

$$\exists_{\sigma_1} v\big(\phi_t(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v) \wedge \phi_s(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v)\big).$$

One again verifies that this ϕ^* satisfies the desired property. So the result holds when ϕ is of the form $t = s$ for Σ^+-terms t and s.

Now suppose that $\phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ is a Σ^+-formula of the form

$$p(t_1(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m), \ldots, t_k(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)),$$

where p has arity $\sigma_1 \times \ldots \times \sigma_k$. Note that it must be that $\sigma_1, \ldots, \sigma_k \in \Sigma$. Either $p \in \Sigma$ or $p \in \Sigma^+ \backslash \Sigma$. We consider the second case. (The first is analogous.) Let $\psi(z_1, \ldots, z_k)$

be the Σ-formula that T^+ uses to explicitly define p and let $\xi(x_1, \ldots, y_{n2})$ be a code for $x_1, \ldots, x_n$. Lemma 7.3.8 and ξ generate the Σ-formulas $\phi_{t_i}(z_i, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})$ for each $i = 1, \ldots, k$. We define ϕ^* to be the Σ-formula

$$\exists_{\sigma_1} z_1 \ldots \exists_{\sigma_k} z_k \big(\phi_{t_1}(z_1, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}) \wedge \ldots \\ \wedge \phi_{t_k}(z_k, \overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}) \wedge \psi(z_1, \ldots, z_k)\big).$$

One can again verify that the result holds for this choice of ϕ^*.

We have covered the "base cases" for our induction. We now turn to the inductive step. We consider the cases of $\neg$, $\wedge$, and $\forall$. Suppose that the result holds for Σ^+-formulas ϕ_1 and ϕ_2. Then it trivially holds for $\neg\phi_1$ by letting $(\neg\phi)^*$ be $\neg(\phi^*)$. It also trivially holds for $\phi_1 \wedge \phi_2$ by letting $(\phi_1 \wedge \phi_2)^*$ be $\phi_1^* \wedge \phi_2^*$.

The $\forall_{\sigma_i}$ case requires more work. If x_i is a variable of sort $\sigma_i \in \Sigma$, we let $(\forall_{\sigma_i} x_i \phi_1)^*$ be $\forall_{\sigma_i} x_i(\phi_1^*)$. The only nontrivial part of the inductive step is when one quantifies over variables with sorts in $\Sigma^+ \backslash \Sigma$. Suppose that $\phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ is a Σ^+-formula and that the result holds for it. We let x_i be a variable of sort $\sigma_i \in \Sigma^+ \backslash \Sigma$, and we show that the result also holds for the Σ-formula $\forall_{\sigma_i} x_i \phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$. There are again four cases. We show that the result holds when σ_i is a product sort and a coproduct sort. The cases of subsorts and quotient sorts follow analogously.

Suppose that T^+ defines σ_i as a product sort with projections π_1 and π_2 of arity $\sigma_i \to \sigma_{i1}$ and $\sigma_i \to \sigma_{i2}$. Quantifying over a variable x_i of product sort σ_i can be thought of as "quantifying over pairs of elements of sorts σ_{i1} and σ_{i2}." Indeed, let $\xi(x_1, \ldots, y_{n2})$ be a code for the variables $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ (these are all of the free variables in $\forall_{\sigma_i} x_i \phi$ with sorts in $\Sigma^+ \backslash \Sigma$). We define a code $\overline{\xi}$ for the variables $x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n$ by

$$\xi(x_1, \ldots, y_{n2}) \wedge \pi_1(x_i) = v_1 \wedge \pi_2(x_i) = v_2.$$

One uses the code $\overline{\xi}$ and the inductive hypothesis to generate the Σ-formula $\phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2)$. We then define the Σ-formula $(\forall_{\sigma_i} x_i \phi)^*$ to be

$$\forall_{\sigma_{i1}} v_1 \forall_{\sigma_{i2}} v_2 \phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_1, v_2).$$

And one verifies that the desired result holds for this choice of $(\forall_{\sigma_i} x_i \phi)^*$.

Suppose that T^+ defines σ_i as a coproduct sort with injections ρ_1 and ρ_2 of arity $\sigma_{i1} \to \sigma_i$ and $\sigma_{i2} \to \sigma_i$. Quantifying over a variable x_i of coproduct sort σ_i can be thought of as "quantifying over *both* elements of sort σ_{i1} and elements of sort σ_{i2}." Indeed, let $\xi(x_1, \ldots, y_{n2})$ be a code for the variables $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ (these are again all of the free variables in $\forall_{\sigma_i} x_i \phi$ with sorts in $\Sigma^+ \backslash \Sigma$). We define two different codes $\overline{\xi}$ for the variables $x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n$ by

$$\xi(x_1, \ldots, y_{n2}) \wedge \rho_1(v_1) = x_i$$
$$\xi(x_1, \ldots, y_{n2}) \wedge \rho_2(v_2) = x_i.$$

We will call the first code $\xi'(x_1, \ldots, y_{n2}, v_1)$ and the second $\xi''(x_1, \ldots, y_{n2}, v_2)$. We use these two codes and the inductive hypothesis to generate Σ-formulas $\phi^{*'}$ and $\phi^{*''}$. We then define the Σ-formula $(\forall_{\sigma_i} x_i \phi)^*$ to be

$$\forall_{\sigma_{i1}} v_1 \forall_{\sigma_{i2}} v_2 \big(\phi^{*'}(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_2) \\ \wedge \phi^{*''}(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2}, v_2)\big).$$

One can verify that the desired result holds again for this definition of $(\forall_{\sigma_i} x_i \phi)^*$. □

Theorem 7.3.7 has the following immediate corollary.

COROLLARY 7.3.9 *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. If T^+ is a Morita extension of T to Σ^+, then for every Σ^+-sentence ϕ there is a Σ-sentence ϕ^* such that $T^+ \vdash \phi \leftrightarrow \phi^*$.*

Proof Let ϕ be a Σ^+-sentence, and consider the empty code ξ. Theorem 7.3.7 implies that there is a Σ-sentence ϕ^* such that $T^+ \vdash \xi \rightarrow (\phi \leftrightarrow \phi^*)$. Since ξ is a tautology, we trivially have that $T^+ \vdash \phi \leftrightarrow \phi^*$. □

The theorems in this section capture different senses in which a Morita extension of a theory "says no more" than the original theory. In this way, Morita equivalence is analogous to definitional equivalence.

At first glance, Morita equivalence might strike one as different from definitional equivalence in an important way. To show that theories are Morita equivalent, one is allowed to take any finite number of Morita extensions of the theories. On the other hand, to show that two theories are definitionally equivalent, it appears that one is only allowed to take *one* definitional extension of each theory. One might worry that Morita equivalence is therefore not perfectly analogous to definitional equivalence.

Fortunately, this is not the case. Theorem 3.3 implies that if theories $T_1, \ldots, T_n$ are such that each T_{i+1} is a definitional extension of T_i, then T_n is, in fact, a definitional extension of T_1. (One can easily verify that this is not true of Morita extensions.) To show that two theories are definitionally equivalent, therefore, one actually *is* allowed to take any finite number of definitional extensions of each theory.

If two theories are definitionally equivalent, then they are trivially Morita equivalent. Unlike definitional equivalence, however, Morita equivalence is capable of capturing a sense in which theories with different sort symbols are equivalent. The following example demonstrates that Morita equivalence is a more liberal criterion for theoretical equivalence.

Example 7.3.10 Let $\Sigma_1 = \{\sigma_1, p, q\}$ and $\Sigma_2 = \{\sigma_2, \sigma_3\}$ be signatures with σ_i sort symbols, and p and q predicate symbols of arity σ_1. Let T_1 be the Σ_1-theory that says p and q are nonempty, mutually exclusive, and exhaustive. Let T_2 be the empty theory in Σ_2. Since the signatures Σ_1 and Σ_2 have different sort symbols, T_1 and T_2 can't possibly be definitionally equivalent. Nonetheless, it's easy to see that T_1 and T_2 are Morita equivalent. Let $\Sigma = \Sigma_1 \cup \Sigma_2 \cup \{i_2, i_3\}$ be a signature with i_2 and i_3 function symbols of arity $\sigma_2 \rightarrow \sigma_1$ and $\sigma_3 \rightarrow \sigma_1$. Consider the following Σ-sentences.

$$\forall_{\sigma_1} x\big(p(x) \leftrightarrow \exists_{\sigma_2} y(i_2(y) = x)\big) \\ \quad \wedge \forall_{\sigma_2} y_1 \forall_{\sigma_2} y_2\big(i_2(y_1) = i_2(y_2) \rightarrow y_1 = y_2\big) \tag{δ_{σ_2}}$$

$$\forall_{\sigma_1} x\big(q(x) \leftrightarrow \exists_{\sigma_3} z(i_3(z) = x)\big) \\ \quad \wedge \forall_{\sigma_3} z_1 \forall_{\sigma_3} z_2\big(i_3(z_1) = i_3(z_2) \rightarrow z_1 = z_2\big) \tag{δ_{σ_3}}$$

$$\forall_{\sigma_1} x\big(\exists_{\sigma_2=1} y(i_2(y) = x) \vee \exists_{\sigma_3=1} z(i_3(z) = x)\big) \\ \quad \wedge \forall_{\sigma_2} y \forall_{\sigma_3} z \neg\big(i_2(y) = i_3(z)\big) \tag{δ_{σ_1}}$$

$$\forall_{\sigma_1} x\big(p(x) \leftrightarrow \exists_{\sigma_2} y(i_2(y) = x)\big) \tag{δ_p}$$

$$\forall_{\sigma_1} x\big(q(x) \leftrightarrow \exists_{\sigma_3} z(i_3(z) = x)\big) \tag{δ_q}$$

The Σ-theory $T_1^1 = T_1 \cup \{\delta_{\sigma_2}, \delta_{\sigma_3}\}$ is a Morita extension of T_1 to the signature Σ. It defines σ_2 to be the subsort of "elements that are p" and σ_3 to be the subsort of "elements that are q."

The theory $T_2^1 = T_2 \cup \{\delta_{\sigma_1}\}$ is a Morita extension of T_2 to the signature $\Sigma_2 \cup \{\sigma_1, i_2, i_3\}$. It defines σ_1 to be the coproduct sort of σ_2 and σ_3. Lastly, the Σ-theory $T_2^2 = T_2^1 \cup \{\delta_p, \delta_q\}$ is a Morita extension of T_2^1 to the signature Σ. It defines the predicates p and q to apply to elements in the "images" of i_2 and i_3, respectively. One can verify that T_1^1 and T_2^2 are logically equivalent, so T_1 and T_2 are Morita equivalent. ⌟

Morita equivalence captures a clear and robust sense in which theories might be equivalent, but it is a difficult criterion to apply outside of the framework of first-order logic. Indeed, without a formal language one does not have the resources to say what an explicit definition is. Questions of equivalence and inequivalence of theories, however, still come up outside of this framework. It is well known, for example, that there are different ways of formulating the theory of smooth manifolds (Nestruev, 2002). There are also different formulations of the theory of topological spaces (Kuratowski, 1966). None of these formulations are first-order theories. Physical theories, too, are rarely formulated in first-order logic, and there are many pairs of physical theories that have been considered to be equivalent. We list just a few examples.

- According to the standard view in physics, Heisenberg's matrix mechanics is equivalent to Schrödinger's wave mechanics – despite the fact that these theories use completely different formalisms, and neither is axiomatizable in first-order logic. Or, if you prefer to be more mathematically rigorous, quantum mechanics can be formulated either in terms of Hilbert spaces or in terms of C^*-algebras. There are good reasons, however, to think that these two formulations are equivalent.
- A model of Einstein's general theory of relativity (GTR) is typically taken to be a smooth manifold with a Lorentzian metric. However, we have a free choice: either we can use a metric of signature $(3, 1)$ or a metric of signature $(1, 3)$. These two formulations of GTR seem to be equivalent – but it's doubtful that we could explicate that equivalence in terms of some regimentation of these theories in first-order logic.

In fact, GTR can also be formulated with a completely different mathematical apparatus, viz. "Einstein algebras," and there is a precise sense in which this formulation is equivalent to the formulation in terms of manifolds (see Rosenstock et al., 2015; Weatherall, 2018).

- GTR seems to differ radically from classical Newtonian gravitation, since the latter posits a static spacetime structure. Some have claimed, in fact, that GTR has a special property, called "general covariance," that disguishes it from all previous spacetime theories. However, in the mid-twentieth century, Henri Cartan formulated a coordinate-free version of Newtonian gravitation on a curved spacetime. If this Newton–Cartan gravitational theory is equivalent to Newtonian gravity, then the latter is also generally covariant. For discussion of this example, see Glymour (1977); Knox (2014); Weatherall (2016a).
- In typical presentations of rigorous methods in classical physics, it is usually assumed (or even partially demonstrated) that the Lagrangian formalism is equivalent to the Hamiltonian formalism. However, North (2009) argues that these two theories have different structure, and hence are inequivalent. For further discussion, see Halvorson (2011); Swanson and Halvorson (2012); Curiel (2014); Barrett (2015).
- Most cutting-edge theories in physics make use of the so-called gauge formalism, and this raises many challenging interpretive issues (see Healey, 2007). Philosophers of science have recently entered into a dispute about whether gauge theories are better thought of in terms of the fiber bundle formalism, or in terms of the holonomy formalism. However, Rosenstock and Weatherall (2016) argue that the two formalisms are equivalent.

Since none of the theories admits a first-order formulation (at least not in any obvious sense), Morita equivalence is incapable of validating these claims of equivalence. Philosophers of science are left with two options: either claim or deny equivalence without a precise account of the standards or propose a more broadly applicable explication of equivalence. We pursue the second option here.

Among the many ways we could explicate theoretical equivalence, we find it most promising to look for hints from contemporary mathematics. In other words, we look to which ideas are working well in contemporary mathematics, and we try to put them to work in the service of philosophy of science. One such fruitful ideas is the notion of **categorical equivalence**, which we first mentioned in Chapter 3. Historically speaking, categorical equivalence was first defined by Eilenberg and Mac Lane (1942, 1945), made a brief appearance in some earlier work in philosophy of science (Pearce, 1985), and has recently by reintroduced in philosophical discussion by Halvorson (2012, 2016); Weatherall (2016a). In the remainder of this section, we review this notion, and prove a few results that relate categorical and Morita equivalence. In summary, for theories with a first-order formulation, Morita equivalence implies categorical equivalence, but not vice versa.

Categorical equivalence is motivated by the following simple observation: First-order theories have categories of models. If T is a Σ-theory, we will use the notation $\mathrm{Mod}(T)$

to denote the **category of models** of T. The objects of $\mathrm{Mod}(T)$ are models of T. For the arrows of $\mathrm{Mod}(T)$, we have a couple of salient choices. On the one hand, we could choose arrows to be homomorphisms – i.e., $f : M \to N$ is a function (or family of functions) that preserves the extensions of the terms in the signature Σ. On the other hand, we could choose arrows to be elementary embeddings – i.e., $f : M \to N$ is an injective function (or family of functions) that preserves the extensions of all Σ formulas.

Let $\mathrm{Mod}(T)$ denote the category with elementary embeddings as arrows, and let $\mathrm{Mod}_h(T)$ denote the category with homomorphisms as arrows. But which of these two categories, $\mathrm{Mod}(T)$ or $\mathrm{Mod}_h(T)$, should we think of as representing the theory T? We will choose the category $\mathrm{Mod}(T)$, with elementary embeddings as arrows, for the following reasons. First, the image of a model of T under a homomorphism f is not necessarily a model of T. For example, let T be the theory (in a single-sorted signature) that says there are exactly two things. Then a model M of T is a set with two elements. However, the mapping $f : M \to M$ that takes both elements to a single element is a homomorphism, and its image $f(M)$ is not a model of T. Such a situation is not necessarily a disaster, but it shows that homomorphisms do not mesh well with full first-order logic. Second, $\mathrm{Mod}_h(\cdot)$ does not even preserve definitional equivalence – i.e., there are definitionally equivalent theories T_1 and T_2 such that $\mathrm{Mod}_h(T_1)$ is not categorically equivalent to $\mathrm{Mod}_h(T_2)$.

Example 7.3.11 Let $\Sigma_1 = \{\sigma\}$, where σ is a sort symbol, and let T_1 be the theory in Σ_1 that says there are exactly two things. Let $\Sigma_2 = \{\sigma, \theta\}$ where θ is a relation of arity $\sigma \times \sigma$, and let T_2 be the theory in Σ_2 that says there are exactly two things, and $T_2 \vDash \theta(x, y) \leftrightarrow (x \neq y)$. Obviously T_2 is a definitional extension of T_1. Now, every arrow of $\mathrm{Mod}_h(T_2)$ is an injection, since it preserves θ and hence $\neq$. But arrows of $\mathrm{Mod}_h(T_1)$ need not be injections. Therefore, $\mathrm{Mod}_h(T_1)$ and $\mathrm{Mod}_h(T_2)$ are not categorically equivalent. ⌟

Because of these issues with homomorphisms, we will continue to associate a theory T with the category $\mathrm{Mod}(T)$ whose objects are models of T and whose arrows are elementary embeddings between these models. We recall now the definition of an equivalence of categories.

DEFINITION 7.3.12 A functor $F : \mathbf{C} \to \mathbf{D}$ is called an **equivalence of categories** just in case there is a functor $G : \mathbf{D} \to \mathbf{C}$, and natural isomorphisms $\eta : GF \Rightarrow 1_{\mathbf{C}}$ and $\varepsilon : FG \Rightarrow 1_{\mathbf{D}}$.

We will also need the following fact, a standard result of category theory (see Mac Lane, 1971, p. 93).

PROPOSITION 7.3.13 *A functor $F : \mathbf{C} \to \mathbf{D}$ is equivalence of categories iff F is full, faithful, and essentially surjective.*

While each first-order theory T defines a category $\mathrm{Mod}(T)$, this structure is not particular to first-order theories. Indeed, one can easily define categories of models

for the different formulations of the theory of smooth manifolds and for the different formulations of the theory of topological spaces. The arrows in these categories are simply the structure-preserving maps between the objects in the categories. One can also define categories of models for physical theories; see, for example, Barrett (2015); Rosenstock et al. (2015); Weatherall (2016a,c, 2018). This means that the following criterion for theoretical equivalence is applicable in a more general setting than definitional equivalence and Morita equivalence. In particular, it can be applied outside of the framework of first-order logic.

DEFINITION 7.3.14 Theories T_1 and T_2 are **categorically equivalent** if their categories of models $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are equivalent.

Categorical equivalence captures a sense in which theories have "isomorphic semantic structure." If T_1 and T_2 are categorically equivalent, then the relationships that models of T_1 bear to one another are "isomorphic" to the relationships that models of T_2 bear to one another.

In order to show how categorical equivalence relates to Morita equivalence, we focus on first-order theories. We will show that categorical equivalence is a strictly weaker criterion for theoretical equivalence than Morita equivalence is. We first need some preliminaries about the category of models $\mathrm{Mod}(T)$ for a first-order theory T. Suppose that $\Sigma \subseteq \Sigma^+$ are signatures and that the Σ^+-theory T^+ is an extension of the Σ-theory T. There is a natural "projection" functor $\Pi : \mathrm{Mod}(T^+) \to \mathrm{Mod}(T)$ from the category of models of T^+ to the category of models of T. The functor Π is defined as follows.

- $\Pi(M) = M|_\Sigma$ for every object M in $\mathrm{Mod}(T^+)$.
- $\Pi(h) = h|_\Sigma$ for every arrow $h : M \to N$ in $\mathrm{Mod}(T^+)$, where the family of maps $h|_\Sigma$ is defined to be $h|_\Sigma = \{h_\sigma : M_\sigma \to N_\sigma \text{ such that } \sigma \in \Sigma\}$.

Since T^+ is an extension of T, the Σ-structure $\Pi(M)$ is guaranteed to be a model of T. Likewise, the map $\Pi(h) : M|_\Sigma \to N|_\Sigma$ is guaranteed to be an elementary embedding. One can easily verify that $\Pi : \mathrm{Mod}(T^+) \to \mathrm{Mod}(T)$ is a functor.

The following three propositions will together establish the relationship between $\mathrm{Mod}(T^+)$ and $\mathrm{Mod}(T)$ when T^+ is a Morita extension of T. They imply that when T^+ is a Morita extension of T, the functor $\Pi : \mathrm{Mod}(T^+) \to \mathrm{Mod}(T)$ is full, faithful, and essentially surjective. The categories $\mathrm{Mod}(T^+)$ and $\mathrm{Mod}(T)$ are therefore equivalent.

PROPOSITION 7.3.15 *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. If T^+ is a Morita extension of T to Σ^+, then Π is essentially surjective.*

Proof If M is a model of T, then Theorem 7.3.1 implies that there is a model M^+ of T^+ that is an expansion of M. Since $\Pi(M^+) = M^+|_\Sigma = M$ the functor Π is essentially surjective. □

PROPOSITION 7.3.16 *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. If T^+ is a Morita extension of T to Σ^+, then Π is faithful.*

Proof Let $h : M \to N$ and $g : M \to N$ be arrows in $\mathrm{Mod}(T^+)$, and suppose that $\Pi(h) = \Pi(g)$. We show that $h = g$. By assumption, $h_\sigma = g_\sigma$ for every sort symbol

$\sigma \in \Sigma$. We show that $h_\sigma = g_\sigma$ also for $\sigma \in \Sigma^+ \backslash \Sigma$. We consider the cases where T^+ defines σ as a product sort or a subsort. The coproduct and quotient sort cases follow analogously.

Suppose that T^+ defines σ as a product sort with projections π_1 and π_2 of arity $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$. Then the following equalities hold.

$$\pi_1^N \circ h_\sigma = h_{\sigma_1} \circ \pi_1^M = g_{\sigma_1} \circ \pi_1^M = \pi_1^N \circ g_\sigma$$

The first and third equalities hold since h and g are elementary embeddings, and the second since $h_{\sigma_1} = g_{\sigma_1}$. One can verify in the same manner that $\pi_2^N \circ h_\sigma = \pi_2^N \circ g_\sigma$. Since N is a model of T^+ and T^+ defines σ as a product sort, we know that $N \models \forall_{\sigma_1} x \forall_{\sigma_2} y \exists_{\sigma=1} z(\pi_1(z) = x \wedge \pi_2(z) = y)$. This implies that $h_\sigma = g_\sigma$.

On the other hand, if T^+ defines σ as a subsort with injection i of arity $\sigma \to \sigma_1$, then the following equalities hold:

$$i^N \circ h_\sigma = h_{\sigma_1} \circ i^M = g_{\sigma_1} \circ i^M = i^N \circ g_\sigma.$$

These equalities follow in the same manner as previously. Since i^N is an injection it must be that $h_\sigma = g_\sigma$. □

Before proving that Π is full, we need the following simple lemma.

LEMMA 7.3.17 *Let M be a model of T^+ with $a_1, \ldots, a_n$ elements of M of sorts $\sigma_1, \ldots, \sigma_n \in \Sigma^+ \backslash \Sigma$. If $x_1, \ldots, x_n$ are variables sorts $\sigma_1, \ldots, \sigma_n$, then there is a code $\xi(x_1, \ldots, x_n, y_{11}, \ldots, y_{n2})$ and elements $b_{11}, \ldots, b_{n2}$ of M such that $M \models \xi[a_1, \ldots, a_n, b_{11}, \ldots, b_{n2}]$.*

Proof We define the code $\xi(x_1, \ldots, y_{n2})$. If T^+ defines σ_i as a product sort, quotient sort, or subsort, then we have no choice about what the conjunct $\xi_i(x_i, y_{i1}, y_{i2})$ is. If T^+ defines σ_i as a coproduct sort, then we know that either there is an element b_{i1} of M such that $\rho_1(b_{i1}) = a_i$ or there is an element b_{i2} of M such that $\rho_2(b_{i2}) = a_i$. If the former, we let ξ_i be $\rho_1(y_{i1}) = x_i$, and if the latter, we let ξ_i be $\rho_2(y_{i2}) = x_i$. One defines the elements $b_{11}, \ldots, b_{n2}$ in the obvious way. For example, if σ_i is a product sort, then we let $b_{i1} = \pi_1^M(a_i)$ and $b_{i2} = \pi_2^M(a_i)$. By construction, we have that $M \models \xi[a_1, \ldots, a_n, b_{11}, \ldots, b_{n2}]$. □

We now use this lemma to show that Π is full.

PROPOSITION 7.3.18 *Let $\Sigma \subseteq \Sigma^+$ be signatures and T a Σ-theory. If T^+ is a Morita extension of T to Σ^+, then Π is full.*

Proof Let M and N be models of T^+ with $h : \Pi(M) \to \Pi(N)$ an arrow in $\mathrm{Mod}(T)$. This means that $h : M|_\Sigma \to N|_\Sigma$ is an elementary embedding. We show that the map $h^+ : M \to N$ is an elementary embedding and therefore an arrow in $\mathrm{Mod}(T^+)$. Since $\Pi(h^+) = h$, this will imply that Π is full.

Let $\phi(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m)$ be a Σ^+-formula, and let $a_1, \ldots, a_n, \overline{a}_1, \ldots, \overline{a}_m$ be elements of M of the same sorts as the variables $x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_m$. Lemma 7.3.17 implies that there is a code $\xi(x_1, \ldots, x_n, y_{11}, \ldots, y_{n2})$ and elements $b_{11}, \ldots, b_{n2}$ of M such that $M \models \xi[a_1, \ldots, a_n, b_{11}, \ldots, b_{n2}]$. The definition of the map h^+ implies that

$N \vDash \xi[h^+(a_1, \ldots, a_n, b_{11}, \ldots, b_{n2})]$. We now show that $M \vDash \phi[a_1, \ldots, a_n, \overline{a}_1, \ldots, \overline{a}_m]$ if and only if $N \vDash \phi[h^+(a_1, \ldots, a_n, \overline{a}_1, \ldots, \overline{a}_m)]$. By Theorem 7.3.7, there is a Σ-formula $\phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})$ such that

$$T^+ \vDash \forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n \forall_{\overline{\sigma}_1} \overline{x}_1 \ldots \forall_{\overline{\sigma}_m} \overline{x}_m \forall_{\sigma_{11}} y_{11} \ldots \forall_{\sigma_{n2}} y_{n2} \big(\xi(x_1, \ldots, y_{n2}) \rightarrow \big(\phi(x_1, \ldots x_n, \overline{x}_1, \ldots, \overline{x}_m) \leftrightarrow \phi^*(\overline{x}_1, \ldots, \overline{x}_m, y_{11}, \ldots, y_{n2})\big)\big) \tag{7.1}$$

We then see that the following string of equivalences holds.

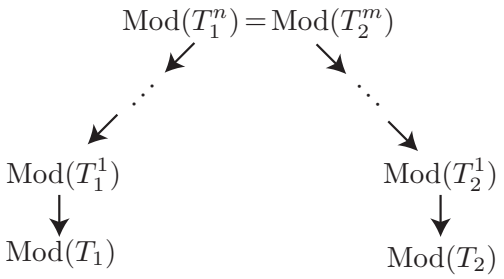
$$\begin{aligned} M \vDash \phi[a_1, \ldots, a_n, \overline{a}_1, \ldots, \overline{a}_m] &\Longleftrightarrow M \vDash \phi^*[\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2}] \\ &\Longleftrightarrow M|_\Sigma \vDash \phi^*[\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2}] \\ &\Longleftrightarrow N|_\Sigma \vDash \phi^*[h(\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2})] \\ &\Longleftrightarrow N \vDash \phi^*[h(\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2})] \\ &\Longleftrightarrow N \vDash \phi^*[h^+(\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2})] \\ &\Longleftrightarrow N \vDash \phi[h^+(a_1, \ldots, a_n, \overline{a}_1, \ldots, \overline{a}_m)] \end{aligned}$$

The first and sixth equivalences hold by (5) and the fact that M and N are models of T^+, the second and fourth hold since ϕ^* is a Σ-formula, the third since $h : M|_\Sigma \rightarrow N|_\Sigma$ is an elementary embedding, and the fifth by the definition of h^+ and the fact that the elements $\overline{a}_1, \ldots, \overline{a}_m, b_{11}, \ldots, b_{n2}$ have sorts in Σ. □

These three propositions provide us with the resources to show how categorical equivalence is related to Morita equivalence. Our first result follows as an immediate corollary.

THEOREM 7.3.19 (Barrett) *Morita equivalence entails categorical equivalence.*

Proof Suppose that T_1 and T_2 are Morita equivalent. Then there are theories $T_1^1, \ldots, T_1^n$ and $T_2^1, \ldots, T_2^m$ that satisfy the three conditions in the definition of Morita equivalence. Propositions 7.3.15, 7.3.16, and 7.3.18 imply that the Π functors between these theories, represented by the arrows in the following figure, are all equivalences.



This implies that $\mathrm{Mod}(T_1)$ is equivalent to $\mathrm{Mod}(T_2)$, and so T_1 and T_2 are categorically equivalent. □

The converse to Theorem 7.3.19, however, does not hold. There are theories that are categorically equivalent but not Morita equivalent. In order to show this, we need one piece of terminology.

DEFINITION 7.3.20 A category $\mathbf{C}$ is **discrete** if it is equivalent to a category whose only arrows are identity arrows.

Note that discrete categories are essentially just sets. In other words, each discrete category is uniquely determined by its underlying set of objects.

THEOREM 7.3.21 *Categorical equivalence does not entail Morita equivalence.*

Proof Let $\Sigma_1 = \{\sigma_1, p_0, p_1, p_2, \ldots\}$ be a signature with a single sort symbol σ_1 and a countable infinity of predicate symbols p_i of arity σ_1. Let $\Sigma_2 = \{\sigma_2, q_0, q_1, q_2, \ldots\}$ be a signature with a single sort symbol σ_2 and a countable infinity of predicate symbols q_i of arity σ_2. Define the Σ_1-theory T_1 and Σ_2-theory T_2 as follows.

$$T_1 = \{\exists_{\sigma_1=1} x(x = x)\}$$
$$T_2 = \{\exists_{\sigma_2=1} y(y = y), \forall_{\sigma_2} y(q_0(y) \to q_1(y)), \forall_{\sigma_2} y(q_0(y) \to q_2(y)), \ldots\}$$

The theory T_2 has the sentence $\forall_{\sigma_2} y(q_0(y) \to q_i(y))$ as an axiom for each $i \in \mathbb{N}$.

We first show that T_1 and T_2 are categorically equivalent. It is easy to see that $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ both have $2^{\aleph_0}$ (non-isomorphic) objects. Furthermore, $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are both discrete categories. We show here that $\mathrm{Mod}(T_1)$ is discrete. Suppose that there is an elementary embedding $f : M \to N$ between models M and N of T_1. It must be that f maps the unique element $m \in M$ to the unique element $n \in N$. Furthermore, since f is an elementary embedding, $M \vDash p_i[m]$ if and only if $N \vDash p_i[n]$ for every predicate $p_i \in \Sigma_1$. This implies that $f : M \to N$ is actually an isomorphism. Every arrow $f : M \to N$ in $\mathrm{Mod}(T_1)$ is therefore an isomorphism, and there is at most one arrow between any two objects of $\mathrm{Mod}(T_1)$. This immediately implies that $\mathrm{Mod}(T_1)$ is discrete. An analogous argument demonstrates that $\mathrm{Mod}(T_2)$ is discrete. Any bijection between the objects of $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ is therefore an equivalence of categories.

But T_1 and T_2 are not Morita equivalent. Suppose, for contradiction, that T is a "common Morita extension" of T_1 and T_2. Corollary 7.3.9 implies that there is a Σ_1-sentence ϕ such that $T \vdash \forall y q_0(y) \leftrightarrow \phi$. One can verify using Theorem 7.3.1 and Corollary 7.3.9 that the sentence ϕ has the following property: If ψ is a Σ_1-sentence and $T_1 \vdash \psi \to \phi$, then either (i) $T_1 \vdash \neg\psi$ or (ii) $T_1 \vdash \phi \to \psi$. But ϕ cannot have this property. Consider the Σ_1-sentence

$$\psi := \phi \wedge \forall x p_i(x),$$

where p_i is a predicate symbol that does not occur in ϕ. We trivially see that $T_1 \vdash \psi \to \phi$, but neither (i) nor (ii) hold of ψ. This implies that T_1 and T_2 are not Morita equivalent. □

7.4 From Geometry to Conceptual Relativity

The twentieth century saw wide swings in prevailing philosophical opinion. In the 1920s, the logical positivists staked out a decidedly antirealist position, particularly in

their rejection of the possibility of metaphysical knowledge. Only a few decades later, prevailing opinion had reached the opposite end of the spectrum. The great analytic philosophers of the 1970s and 1980s – Putnam, Lewis, Kripke, etc. – were unabashed proponents of scientific and metaphysical realism. Or perhaps it would be more accurate to say that these philosophers presupposed realism and built their philosophical programs on the assumption that there is a kind of knowledge that transcends the claims of the empirical sciences.

But the pendulum didn't rest there. By the end of the twentieth century, several analytic philosophers were giving arguments against realism, saying that it didn't mesh well with the way that the sciences actually work. For example, Putnam and Goodman pointed to the existence of different formulations of Euclidean geometry, some of which take points as primitives, and some of which take lines as primitives, saying that realists must render the incorrect verdict that these are inequivalent theories. We will call the invocation of this particular example the *argument from geometry against realism*.

According to the argument from geometry, certain situations can equally well be described using a theory that takes points as fundamental entities or, instead, using a theory that takes lines as fundamental entities. Someone who adopts the first theory is committed to the existence of points and not lines, while someone who adopts the second theory is committed to the existence of lines and not points. But points and lines are different kinds of things, and, in general, the number of points (according to the first theory) will be different from the number of lines (according to the second theory). Since both parties correctly describe the world but use different ontologies to do so, it's supposed to follow that there is no matter of fact about what the ontology of the world is – in direct contradiction with a fundamental tenet of metaphysical realism.

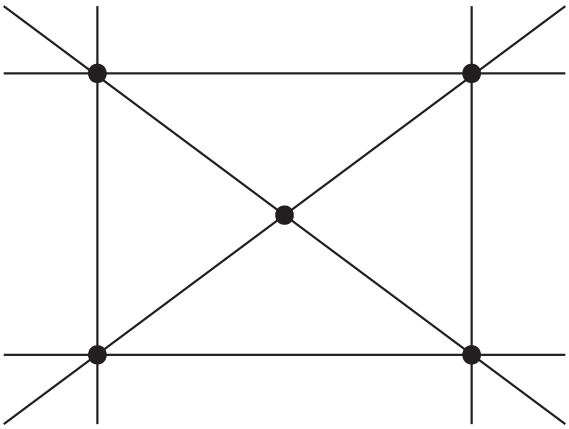
In responding to examples of this sort, metaphysical realists typically agree that the two theories in question involve incompatible ontological commitments (see Sider, 2009; van Inwagen, 2009). These realists then claim, however, that at most one of the two theories can be correct, at least in a fundamental sense. The upshot of this kind of response, of course, is that a realist ontology has been purchased at the price of an epistemic predicament: Only one of the theories is correct, but we will never know which one.

In this section, we propose another reply to arguments of this sort, and specifically to the argument from geometry. We show that geometries with points can naturally be considered equivalent to geometries with lines, and we argue that this equivalence does not in any way threaten the idea that there is an objective world. In other words, since these two theories are equivalent, there is a sense in which they involve exactly the same ontological commitments. The example of geometries with points and geometries with lines does not undermine metaphysical realism in the way that Putnam and Goodman suggested.

There are many ways to formulate a particular geometric theory, and these formulations often differ with respect to the kinds of objects that are taken as primitive. The most famous example of this phenomenon is Euclidean geometry. Tarski first formulated Euclidean geometry using open balls (Tarski, 1929), and later using points (Tarski, 1959). Schwabhäuser and Szczerba (1975) formulated Euclidean geometry

using lines, and Hilbert (1930) used points, lines, planes, and angles. These formulations of Euclidean geometry all take different kinds of objects to be primitive, but despite this ostensible difference, they nonetheless manage to express the same geometric facts. Indeed, it is standard to recognize some sense in which all of these formulations of Euclidean geometry are *equivalent*. This sense of equivalence, however, is rarely made perfectly precise.

In fact, from a certain point of view, it might seem that these theories cannot be equivalent. Consider a simple example: Take six lines in the Euclidean plane, as in the following diagram.



On the one hand, if this diagram were described in terms of the point-based version of Euclidean geometry (T_p), then we would say that there are exactly five things. On the other hand, if this diagram were described in terms of the line-based version of Euclidean geometry (T_ℓ), then we would say that there are exactly six things. The point-based and line-based descriptions therefore seem to disagree about a feature of the diagram – namely, how many things there are in the diagram.

Indeed, according to one natural notion of theoretical equivalence, the first description T_p is not equivalent to the second description T_ℓ. The notion we have in mind is **definitional equivalence**, which we introduced in Section 4.6, and which first entered into philosophy of science through the work of Glymour (1971, 1977, 1980). If two theories are definitionally equivalent, then the cardinalities of their respective domains will be equal. Since the domains of T_p and T_ℓ do not have the same cardinality, these descriptions cannot be definitionally equivalent.

This would be the end of the matter if definitional equivalence were the only legitimate notion of theoretical equivalence. But, as we now know, there is a better notion of theoretical equivalence that does not prejudge issues about the cardinality of domains.

All of the geometries that we will consider are formulated using (some subset of) the following vocabulary. Here we follow Schwabhäuser et al. (1983).

- The sort symbols σ_p and σ_ℓ will indicate the sort of points and the sort of lines, respectively. We will use letters from the beginning of the alphabet like a, b, c to denote variables of sort σ_p, and letters from the end of the alphabet like x, y, z to denote variables of sort σ_ℓ.

- The predicate symbol $r(a,x)$ of arity $\sigma_p \times \sigma_\ell$ indicates that the point a lies on the line x.
- The predicate symbol $s(a,b,c)$ of arity $\sigma_p \times \sigma_p \times \sigma_p$ indicates that the points a, b, and c are colinear.
- The predicate symbol $p(x,y)$ of arity $\sigma_\ell \times \sigma_\ell$ indicates that the lines x and y intersect.
- Lastly, the predicate symbol $o(x,y,z)$ of arity $\sigma_\ell \times \sigma_\ell \times \sigma_\ell$ indicates that the lines x, y, and z are compunctual – i.e., that they all intersect at a single point.

We now prove two theorems that capture the equivalence between geometries with points and geometries with lines. We then provide three examples that illustrate the generality of these results.

Suppose that we are given a formulation of geometry T that uses both of the sort symbols σ_p and σ_ℓ. The two theorems that we will prove in this section show that, given some natural assumptions, the theory T is Morita equivalent both to a theory T_p that only uses the sort σ_p and to a theory T_ℓ that only uses the sort σ_ℓ. In this sense, therefore, the geometry T can be formulated using only points, only lines, or both points and lines.

Our first theorem captures a sense in which the geometry T can be formulated using only points. In order to prove this theorem, we will need the following important result. The proof of this proposition is given by Schwabhäuser et al. (1983, Proposition 4.59).

PROPOSITION 7.4.1 (Elimination of line variables) *Let T be a theory formulated in the signature $\Sigma = \{\sigma_p, \sigma_\ell, r, s\}$, and suppose that T entails the following sentences:*

1. $(a \neq b) \rightarrow \exists_{=1} x\, (r(a,x) \wedge r(b,x))$
2. $\forall x \exists a \exists b\, (r(a,x) \wedge r(b,x) \wedge (a \neq b))$
3. $s(a,b,c) \leftrightarrow \exists x\, (r(a,x) \wedge r(b,x) \wedge r(c,x))$

Then for every Σ-formula ϕ without free variables of sort σ_l, there is a Σ-formula ϕ^*, whose free variables are included in those of ϕ, that contains no variables of sort σ_ℓ, and such that $T \vDash \forall \vec{a}(\phi(\vec{a}) \leftrightarrow \phi^*(\vec{a}))$.

We should take a moment here to unravel the intuition behind this proposition. The theory T can be thought of as a geometry that is formulated in terms of points and lines, using the basic notions of a point lying on a line and three points being colinear. Since the theory T is a geometry, the sentences 1, 2, and 3 are sentences that one should naturally expect T to satisfy. Given these assumptions on T, Proposition 7.4.1 simply guarantees that Σ-formulas ϕ can be "translated" into corresponding formulas ϕ^* that do not use the apparatus of lines. This translation eliminates the line variables from every Σ-formula in two steps. First, one uses the fact that every line is uniquely characterized by two nonidentical points lying on it to replace equalities between line variables with more complex expressions using the predicate r. Second, one replaces instances of the predicate $r(a,x)$ by using complex expressions involving the colinearity predicate $s(a,b,c)$. The reader is encouraged to consult Schwabhäuser et al. (1983, Proposition 4.59) for details.

With this proposition in hand, we have the following result.

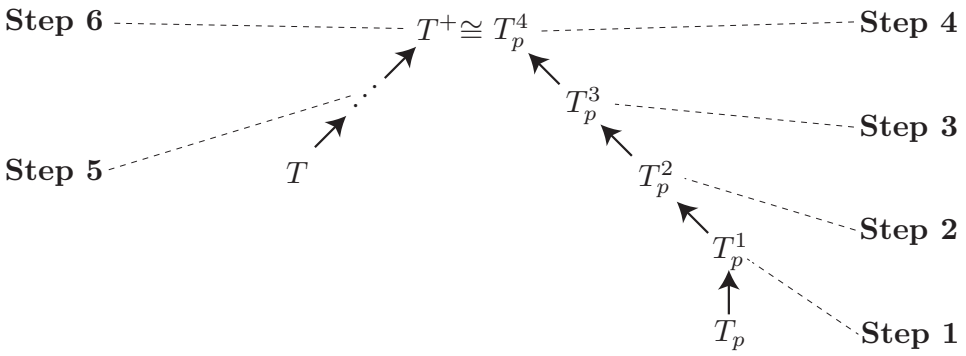
THEOREM 7.4.2 (Barrett) *Let T be a theory that satisfies the hypotheses of Proposition 7.4.1. Then there is a theory T_p in the restricted signature $\Sigma_0 = \Sigma \backslash \{\sigma_\ell, r\}$ that is Morita equivalent to T.*

Theorem 7.4.2 captures a sense in which every geometry that is formulated with points and lines could be formulated equally well using only points. The idea behind the proof of Theorem 7.4.2 should be clear. Consider the Σ_0-theory defined by

$$T_p = \{\phi^* : T \vdash \phi\},$$

where the existence of the sentences ϕ^* is guaranteed by the fact that T satisfies the hypotheses of Proposition 7.4.1. The theory T_p can be thought of as a theory that "says the same thing as T" but uses only the apparatus of points. One proves Theorem 7.4.2 by showing that this theory T_p has the resources to define the sort σ_ℓ of lines. (Note that in the following proof we abuse our convention and occasionally use the variables x, y, z as variables that are not of sort σ_ℓ. But the sort of variables should always be clear from context.)

Proof of Theorem 7.4.2 It suffices to show that the theories T and T_p are Morita equivalent. The following figure illustrates the structure of our argument:



We begin on the right-hand side of the figure by building four theories T_p^1, T_p^2, T_p^3, and T_p^4. The purpose of these theories is to define, using the resources of the theory T_p, the symbols σ_ℓ and r.

Step 1: The theory T_p^1 is the Morita extension of T_p obtained by defining a new sort symbol $\sigma_p \times \sigma_p$ as a product sort (of the sort σ_p with itself). We can think of the elements of the sort $\sigma_p \times \sigma_p$ as pairs of points. The theory T_p^1 is a Morita extension of T_p to the signature $\Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2\}$, where π_1 and π_2 are both function symbols of arity $\sigma_p \times \sigma_p \to \sigma_p$.

Step 2: The theory T_p^2 is the Morita extension of T_p^1 obtained by defining a new sort symbol σ_s as a subsort of $\sigma_p \times \sigma_p$. The elements of sort σ_s are the elements (a, b) of sort $\sigma_p \times \sigma_p$ such that $a \neq b$. One can easily write out the defining formula for the subsort σ_s to guarantee that this is the case. We can think of the elements of sort σ_s as the pairs of distinct points or, more intuitively, as the "line segments formed between

distinct points." The theory T_p^2 is a Morita extension of T_p^1 to the signature $\Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i\}$, where i is a function symbol of arity $\sigma_s \to \sigma_p \times \sigma_p$.

Step 3: The theory T_p^2 employs a sort of "line segments," but we do not yet have a sort of lines. Indeed, we need to take care of the fact that some line segments determine the same line. We do this by considering the theory T_p^3, the Morita extension of T_p^2 obtained by defining the sort symbol σ_ℓ as a quotient sort of σ_s using the formula

$$s(\pi_1 \circ i(x), \pi_1 \circ i(y), \pi_2 \circ i(y)) \wedge s(\pi_2 \circ i(x), \pi_1 \circ i(y), \pi_2 \circ i(y)).$$

Using the fact that T is a conservative extension of T_p, one can easily verify that T_p^2 satisfies the admissibility conditions for this definition – i.e., the preceding formula is an equivalence relation according to T_p^2. The idea here is simple: two line segments (a_1, a_2) and (b_1, b_2) determine the same line just in case the points a_1, b_1, b_2 are colinear, and the points a_2, b_1, b_2 are, too. The theory T_p^3 simply identifies the line segments that determine the same line in this sense. We have now defined the sort σ_ℓ of lines. The theory T_p^3 is a Morita extension of T_p^2 to the signature $\Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \sigma_\ell, \epsilon\}$, where ϵ is a function symbol of sort $\sigma_s \to \sigma_\ell$.

Step 4: All that remains on the right-hand side of the figure is to define the predicate symbol r. The theory T_p^4 is the Morita extension of T_p^3 obtained by defining the predicate $r(a, z)$ using the formula

$$\exists_{\sigma_p \times \sigma_p} x \exists_{\sigma_s} y(\pi_1(x) = a \wedge i(y) = x \wedge \epsilon(y) = z).$$

The idea here is again intuitive. A point a is on a line z just in case there is another point b such that the pair of points (a, b) determines the line l. (In the preceding formula, one can think of the variable x as playing the role of this pair (a, b).) The theory T_p^4 is a Morita extension of T_p^3 to the signature $\Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \sigma_\ell, \epsilon, r\}$.

Step 5: We now turn to the left-hand side of our organizational figure. The theory T is formulated in the signature Σ, so it needs to define all of the new symbols that we added to the theory T_p in the course of defining σ_p and r. The theory T defines the symbols $\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i$ in the obvious manner. For example, it defines $\sigma_p \times \sigma_p$ as the product sort (of σ_p with itself) with the projections π_1 and π_2.

We still need, however, to define the function symbol ϵ. The function ϵ intuitively maps a pair of distinct points to the line that they determine. This suggests that we define $\epsilon(x) = y$ using the formula

$$r(\pi_1 \circ i(x), y) \wedge r(\pi_2 \circ i(x), y).$$

Intuitively, this formula is saying that a pair of points $x = (x_1, x_2)$ determines a line y just in case x_1 is on y and x_2 is on y. We call the theory that results from defining all of these symbols T^+.

Step 6: All that remains now is to show that the theory T_p^4 is logically equivalent to the theory T^+. This argument is mainly a tedious verification. The only nontrivial part of the argument is the following: one needs to show that $T_p^4 \vDash \phi$ for every sentence ϕ such that $T \vDash \phi$. One does this by verifying that T_p^4 itself entails the three sentences 1, 2, and 3 in the statement of Proposition 1. This means that T_p^4 entails the sentences

$\phi \leftrightarrow \phi^*$ for every Σ-sentence ϕ. In conjunction with the fact that $T_p^4 \vDash \phi^*$ for every consequence ϕ of T, this implies that $T_p^4 \vDash \phi$. The theories T_p^4 and T^+ are logically equivalent, so T_p and T must be Morita equivalent. □

Our second theorem is perfectly analogous to Theorem 7.4.2. It captures a sense in which a geometry T can be formulated using only lines. As with Theorem 7.4.2, we will need a preliminary result. The proof of the following proposition is given by Schwabhäuser et al. (1983, Proposition 4.89).

PROPOSITION 7.4.3 (Elimination of point variables) *Let T be a theory formulated in the signature $\Sigma = \{\sigma_p, \sigma_\ell, r, p, o\}$, and suppose that T implies the following sentences.*

1. $(x \neq y) \rightarrow \exists_{\leq 1} a(r(a,x) \wedge r(a,y))$
2. $\forall a \exists x \exists y((x \neq y) \wedge r(a,x) \wedge r(a,y))$
3. $o(x,y,z) \leftrightarrow \exists a(r(a,x) \wedge r(a,y) \wedge r(a,z))$
4. $p(x,y) \leftrightarrow ((x \neq y) \wedge s(x,y,y))$
5. $p(x,y) \leftrightarrow ((x \neq y) \wedge \exists a(r(a,x) \wedge r(a,y)))$

Then for every Σ-formula ϕ without free variables of sort σ_p, there is a Σ-formula ϕ^*, whose free variables are included in those of ϕ, that contains no variables of sort σ_p, and such that $T \vDash \forall \vec{x}(\phi(\vec{x}) \leftrightarrow \phi^*(\vec{x}))$.

Proposition 7.4.3 is perfectly analogous to Proposition 7.4.1. One again thinks of the theory T as a geometry, and so sentences 1–5 are sentences that one naturally expects T to satisfy. Proposition 7.4.3 guarantees that Σ-formulas can be "translated" into formulas ϕ^* that do not use the apparatus of points. Analogous to Proposition 7.4.1, one proves this proposition by showing that variables of sort σ_p can be eliminated in the following manner. One first replaces equalities between these variables and then interprets $r(a,x)$ in terms of $o(y,z,x)$, where y and z have a as their intersection point. The reader is invited to consult Schwabhäuser et al. (1983, Proposition 4.89) for further details.

With Proposition 7.4.3 in hand, we have the following result.

THEOREM 7.4.4 (Barrett) *Let T be a theory that satisfies the hypotheses of Proposition 7.4.3. There is a theory T_ℓ in the restricted signature $\Sigma_0 = \Sigma \backslash \{\sigma_p, r\}$ that is Morita equivalent to T.*

Proof The proof is analogous to the proof of Theorem 7.4.2, so we will not go into as much detail. Consider the Σ_0-theory T_ℓ defined by $T_\ell = \{\phi^* : T \vdash \phi\}$, where the existence of the sentences ϕ^* is guaranteed since T satisfies the hypotheses of Proposition 7.4.3. One shows that the theory T_ℓ is Morita equivalent to T. The theory T_ℓ needs to define the sort symbol σ_p. It does this by first defining a product sort of "pairs of lines," and then a subsort of "pairs of intersecting lines." The sort of points is then the quotient sort that results from identifying two pairs of intersecting lines (w,x) and (y,z) just in case both w,x,y and w,x,z are compunctual. The theory T_ℓ also needs to define the symbol r. It does this simply by requiring that $r(a,x)$ holds of a point a and a line x

just in case there is another line y such that the pair of lines (x, y) intersect at the point a. As in the proof of Theorem 7.4.2, T defines the symbols of T_ℓ in the natural way. □

Theorem 7.4.2 shows that every geometry formulated using points and lines could be formulated equally well using only points; Theorem 7.4.4 shows that it could be formulated equally well using only lines. These two results together capture a robust sense in which geometries with points and geometries with lines are equivalent theories.

Theorems 7.4.2 and 7.4.4 are quite general. Indeed, one can verify that many of the theories that we usually think of as geometries satisfy the hypotheses of the two theorems. We provide three examples here. We begin by revisiting a simple geometric theory that we considered earlier.

Example 7.4.5 Recall the earlier diagram of six lines and five points in the Euclidean plane. By interpreting the symbols $\sigma_p, \sigma_\ell, r, s, p,$ and o in the natural way, one can easily convert this diagram into a $\{\sigma_p, \sigma_\ell, r, s, p, o\}$-structure M. We now consider the geometric theory $\text{Th}(M) = \{\phi : M \vdash \phi\}$. One can verify by inspection that $\text{Th}(M)$ satisfies the hypotheses of both Theorems 7.4.2 and 7.4.4. Theorem 7.4.2 implies that this diagram can be fully described using only the apparatus of points (using the theory $\text{Th}(M)_p$), while Theorem 7.4.4 implies that it can be fully described using only the apparatus of lines (using the theory $\text{Th}(M)_\ell$), and all three of these theories are Morita equivalent. ⌟

In our next two examples, we consider more general geometric theories: projective geometry and affine geometry.

Example 7.4.6 (Projective geometry) Projective geometry is a theory T_{proj} formulated in the signature $\{\sigma_p, \sigma_\ell, r\}$, where all of these symbols are understood exactly as they were earlier. The theory T_{proj} has the following three axioms (Barnes and Mack, 1975).

- $a \neq b \rightarrow \exists_{=1} x(r(a, x) \wedge r(b, x))$
- $x \neq y \rightarrow \exists_{=1} a(r(a, x) \wedge r(a, y)$
- There are at least four points, no three of which lie on the same line.

(One can easily express the third axiom as a sentence of first-order logic, but we here refrain for the sake of clarity.)

Projective geometry satisfies the hypotheses of both Theorems 7.4.2 and 7.4.4. We consider Theorem 7.4.4. In order to apply this result, we need to add the following two axioms that define the symbols p and o:

$$p(x, y) \leftrightarrow (x \neq y \wedge \exists a(r(a, x) \wedge r(a, y))) \qquad (\theta_p)$$

$$o(x, y, z) \leftrightarrow \exists a(r(a, x) \wedge r(a, y) \wedge r(a, z)) \qquad (\theta_o)$$

One can easily verify that the $\{\sigma_\ell, \sigma_p, r, p, o\}$-theory T^+_{proj} obtained by adding the definitions θ_p and θ_o to the axioms of T_{proj} satisfies sentences 1–5 of Proposition 7.4.3. Theorem 7.4.4 then implies that there is a theory in the restricted signature $\{\sigma_\ell, p, c\}$ that is Morita equivalent to T^+_{proj}. Projective geometry can therefore be formulated using only the apparatus of lines. One argues in a perfectly analogous manner to show that Theorem 7.4.2 also applies to projective geometry, so it can also be formulated using only the apparatus of points. ⌟

Example 7.4.7 (Affine geometry) Affine geometry is a theory T_{aff} formulated in the signature $\{\sigma_\ell, \sigma_p, r\}$, where all of these symbols are again understood exactly as earlier. The theory T_{aff} has the following five axioms (Veblen and Young, 1918, 118).

- $a \neq b \rightarrow \exists x(r(a,x) \wedge r(b,x))$
- $\neg r(a,x) \rightarrow \exists_{=1} y(r(a,y) \wedge \forall b(r(b,y) \rightarrow \neg r(b,x)))$
- $\forall x \exists a \exists b(a \neq b \wedge r(a,x) \wedge r(b,x))$
- $\exists a \exists b \exists c(a \neq b \wedge a \neq c \wedge b \neq c \wedge \neg \exists x(r(a,x) \wedge r(b,x) \wedge r(c,x)))$
- Pappus' theorem (Veblen and Young, 1918, p. 103 and Figure 40).

The fifth axiom can easily be written as a first-order sentence in the signature $\{\sigma_\ell, \sigma_p, r\}$, but since this axiom is not used in the following argument, we leave its translation to the reader. (Indeed, one only needs the first, third, and fourth axioms of T_{aff} to complete all of the following verifications.)

Affine geometry satisfies the hypotheses of both Theorems 7.4.2 and 7.4.4. We consider Theorem 7.4.2. In order to apply this result, we need to add one additional axiom to T_{aff} that defines the symbol s as follows:

$$s(a,b,c) \leftrightarrow \exists x(r(a,x) \wedge r(b,x) \wedge r(c,x)). \qquad (\theta_s)$$

It is now trivial to verify that sentences 1–3 of Proposition 7.4.1 are satisfied by the $\{\sigma_\ell, \sigma_p, r, s\}$-theory T^+_{aff} that is obtained by adding the sentence θ_s to the axioms of T_{aff}. Theorem 7.4.2 therefore implies that there is a theory in the restricted signature $\{\sigma_p, s\}$ that is Morita equivalent to T^+_{aff}, capturing a sense in which affine geometry can be formulated using only the apparatus of points. In a perfectly analogous manner, one can apply the Theorem 7.4.4 to the case of affine geometry. This captures a sense in which affine geometry can also be formulated using only lines. ⌟

The previous example is more general than it might initially appear. Indeed, affine geometry serves as the foundation for many of our most familiar geometries. For example, by supplementing the affine geometry with the proper notion of orthogonality, one can obtain two dimensional Euclidean geometry or two dimensional Minkowski geometry. (See Coxeter [1955], Szczerba and Tarski [1979], Szczerba [1986, p. 910], or Goldblatt [1987] for details.) Theorems 7.4.2 and 7.4.4 therefore capture a sense in which both Euclidean geometry and Minkowski geometry can be formulated using either points or lines.

7.5 Morita Equivalence Is Intertranslatability

The results of the previous section might seem to validate Putnam's arguments against metaphysical realism. After all, we proved that line-based geometries are (Morita) equivalent to point-based geometries. However, in order to make a case against realism, one would need say something more about why Morita equivalence is the right notion of equivalence. Perhaps, you might worry that the inventors of Morita equivalence cooked it up precisely to deliver this kind of antirealistic verdict. Indeed, the definition of Morita equivalence seems to include several arbitrary choices. Why, for example, allow the construction of just these kinds of sorts (products, coproducts, subsorts, quotient sorts) and not others (such as exponential sorts)?

In this section, we provide independent motivation for Morita equivalence. In particular, we show that Morita equivalence corresponds to the notion of intertranslatability described in Section 5.4. The coincidence between these two notions is remarkable, as they were developed independently of each other. On the one hand, Morita equivalence was proposed by Barrett and Halvorson (2016b) and was motivated by results in topos theory (see Johnstone, 2003). On the other hand, many-sorted intertranslatability was being used already in model theory in the 1970s. It was given a precise formulation by van Benthem and Pearce (1984), and has been further articulated by Visser (2006). The coincidence of these two notions – Morita equivalence and intertranslatability – suggests that there is something natural about them, at least from a mathematical point of view.

We established previously that definitional equivalence of single-sorted theories corresponds to strong intertranslatability (4.6.17 and 6.6.21). We now generalize this result as follows: for many-sorted theories without trivial sorts (i.e., sorts that are restricted to only one thing), Morita equivalence corresponds to weak intertranslatability. Our argument proceeds as follows: first we show that if T^+ is a Morita extension of T, then there is a reduction $R : T^+ \to T$ that is inverse (up to homotopy) to the inclusion $I : T \to T^+$. Intuitively speaking, R expands each definiendum in Σ^+ into its definiens in Σ. The trick here is figuring out how an equality relation $x =_\sigma y$, with $\sigma \in \Sigma^+$, can be reconstrued in terms of Σ-formulas. For product sorts in Σ^+, the answer is simple: $x =_{\sigma_1 \times \sigma_2} y$ can be reconstrued as $(x_1 =_{\sigma_1} y_1) \wedge (x_2 =_{\sigma_2} y_2)$. For coproduct sorts in Σ^+, the answer is more complicated. The problem here is that an equality $x =_{(\sigma_1 + \sigma_2)} y$ defines an equivalence relation, but $(x_1 =_{\sigma_1} y_1) \vee (x_2 =_{\sigma_2} y_2)$ does not define an equivalence relation; hence the former cannot be reconstrued as the latter. Thus, to reconstrue equality statements over a coproduct sort, we will need a more roundabout construction. To this end, we borrow the following definition from Harnik (2011).

DEFINITION 7.5.1 Let T be a theory in signature Σ. We say that T is **proper** just in case there is a Σ-formula $\phi(z)$ such that $T \vdash \exists z \phi(z)$ and $T \vdash \exists z \neg \phi(z)$. Here we allow z also to be a sequence of variables, possibly of various sorts.

NOTE 7.5.2 Suppose that Σ has a sort symbol σ and that $T \vdash \exists x \exists y (x \neq_\sigma y)$. Then T is proper, as witnessed by the formula $\phi(x, y) \equiv (x =_\sigma y)$.

THEOREM 7.5.3 (Washington) *Let T be a proper theory, and let T^+ be a Morita extension of T. Then there is a translation $R : T^+ \to T$ that is inverse to the inclusion $I : T \to T^+$.*

Before we give a proof of this result, we give an example to show why it's necessary to restrict to proper theories.

Example 7.5.4 Let T be the theory of equality over a single sort σ. Let T^+ be the Morita extension of T to the signature $\{\sigma, \sigma', \rho_1, \rho_2\}$, where T^+ defines σ' as a coproduct with coprojections $\rho_1 : \sigma \to \sigma'$ and $\rho_2 : \sigma \to \sigma'$. In this case, $T^+ \vdash \rho_1(x_1) \neq \rho_2(x_2)$; hence $T^+ \vdash \exists y_1 \exists y_2 (y_1 \neq y_2)$, with y_1, y_2 variables of sort σ'.

If there were a translation $R : T^+ \to T$, then we would have a corresponding model functor $R^* : \mathrm{Mod}(T) \to \mathrm{Mod}(T^+)$. But consider the model M of T with $M(\sigma)$ a singleton set. In that case, $(R^*M)(\sigma')$ would be a quotient of a subset of $M(\sigma) \times \cdots \times M(\sigma)$, which is again a singleton set. This contradicts the fact that $T^+ \vdash \exists y_1 \exists y_2 (y_1 \neq y_2)$. Therefore, there is no translation (in the sense of 5.4.2) from T^+ to T. ⌟

Proof of 7.5.3 Since T is proper, there is a sort σ_* of Σ and a formula $\phi(z)$ with $z : \sigma_*$ such that $T \vdash \exists z \phi(z)$ and $T \vdash \exists z \neg\phi(z)$. We first define $R : S \to (S^+)^*$. All cases are straightforward, except for coproduct sorts, which require a special treatment.

- Suppose that T^+ defines σ as a product with projections $\pi_1 : \sigma \to \sigma_1$ and $\pi_2 : \sigma \to \sigma_2$. Then we define $R(\sigma) = \sigma_1, \sigma_2$.
- Suppose that T^+ defines σ as a coproduct with coprojections $\rho_1 : \sigma_1 \to \sigma$ and $\rho_2 : \sigma_2 \to \sigma$. Then we define $R(\sigma) = \sigma_1, \sigma_2, \sigma_*$. Here the final sort σ_* plays an auxiliary role that permits us to define a coproduct of two sorts as a quotient of a product of sorts.
- Suppose that T^+ defines σ as a subsort with injection $i : \sigma \to \sigma'$. Then we define $R(\sigma) = \sigma'$.
- Suppose that T^+ defines σ as a quotient sort with projection $p : \sigma' \to \sigma$. Then we define $R(\sigma) = \sigma'$.
- Finally, if $\sigma \in \Sigma$, we define $R(\sigma) = \sigma$.

We now define the formulas E_σ for each sort symbol $\sigma \in \Sigma^+$.

- If σ is defined as a product sort $\sigma_1 \times \sigma_2$, then we set

$$E(x_1, x_2, y_1, y_2) \equiv (x_1 = y_1) \wedge (x_2 = y_2).$$

- Suppose that T^+ defines σ as a coproduct sort $\sigma_1 + \sigma_2$, in which case $R(\sigma) = \sigma_1, \sigma_2, \sigma_*$. Intuitively speaking, we will use a triple x, y, z to represent a variable of sort $\sigma_1 + \sigma_2$. We will think of the triples satisfying $\phi(z)$ as ranging over σ_1 (with y and z as dummy variables), and we will think of the triples satisfying $\neg\phi(z)$ as ranging over σ_2 (with x and z as dummy variables). Since $\vdash \phi(z) \vee \neg\phi(z)$, any triple x, y, z satisfies exactly one of these two conditions. We can then explicitly define the relevant formula $E(x_1, x_2, z; x_1', x_2', z')$ as

$$(\phi(z) \wedge \phi(z') \wedge (x_1 = x_1')) \vee (\neg\phi(z) \wedge \neg\phi(z') \wedge (x_2 = x_2')).$$

- If T^+ defines σ as a quotient sort in terms of a Σ-formula ϕ, then define $E(x, y) \equiv \phi(x, y)$.
- If σ is defined as a subsort in terms of a Σ-formula ϕ, then define $E(x, y) \equiv \phi(x) \wedge \phi(y) \wedge (x = y)$.

To complete the definition of the reconstrual, we need to give the mapping from predicate symbols and function symbols of Σ^+ to Σ.

- If $p \in \Sigma^+ \backslash \Sigma$ is a predicate symbol with explicit definition $p \leftrightarrow \psi_p$, then we define $R(p)(x_1, \ldots, x_n)$ as $\psi_p(x_1, \ldots, x_n)$. If $p \in \Sigma$, then we define the image to be $p(x_1, \ldots, x_n)$.
- If $f \in \Sigma^+ \backslash \Sigma$ is a function symbol that is not used in an explicit definition of a sort symbol, and if f has explicit definition $(f(\vec{x}) =_\sigma y) \leftrightarrow \psi_f(\vec{x}, y)$, then we define $R(f)(\vec{x}, y)$ as $\psi_f(\vec{x}, y)$. If $f \in \Sigma$, then we define the image to be $f(\vec{x}) =_\sigma y$.
- For function symbols $\pi_i : \sigma \to \sigma_i$ that define a product sort, we define $R(\pi_1)(x_1, x_2, y_1) \equiv (x_1 =_{\sigma_1} y_1)$ and $R(\pi_2)(x_1, x_2, y_2) \equiv (x_2 =_{\sigma_2} y_2)$.
- For function symbols $\rho_i : \sigma_i \to \sigma$ that define a coproduct sort, we define $R(\rho_1)(v_1, x_1, x_2, z) \equiv (v_1 =_{\sigma_1} x_1)$ and $R(\rho_1)(v_2, x_1, x_2, z) \equiv (v_2 =_{\sigma_2} x_2)$.
- For a function symbol $\epsilon : \sigma' \to \sigma$ that defines a quotient sort, we define $R(\epsilon)(x, y) \equiv \phi(x, y)$.
- For a function symbol $i : \sigma \to \sigma'$ that defines a subsort, we define $R(i)(x, y) \equiv \phi(x) \wedge \phi(y) \wedge (x = y)$.

We now show that $RI \simeq 1_T$ and $IR \simeq 1_{T^+}$. The former case is trivial: since R acts as the identity on elements of Σ, it follows that $RI = 1_T$. For the proof that $IR \simeq 1_{T+}$, we will define a t-map $\chi : IR \Rightarrow 1_{T^+}$, and we will show that χ is a homotopy.

Recall that a homotopy is a family of formulas, one for each sort symbol $\sigma \in \Sigma^+$. We will treat only the case where T^+ defines σ as a coproduct over $\rho_1 : \sigma_1 \to \sigma$ and $\rho_2 : \sigma_2 \to \sigma$. We need to define a Σ^+-formula χ whose free variables are of sorts $R(\sigma)$ and σ. Intuitively speaking, χ should establish a bijection between elements of sort $(\sigma_1, \sigma_2, \sigma_*)/E$ and elements of sort σ. We define

$$\chi(x_1, x_2, z, x) \equiv (\phi(z) \wedge (\rho_1(x_1) = x)) \vee (\neg\phi(z) \wedge (\rho_2(x_2) = x)).$$

We sketch the argument for the various conditions in the definition of a t-map (5.4.11). Throughout, we argue internally to the theory T^+.

- We show that χ is well defined relative to the equivalence relation E on $\sigma_1, \sigma_2, \sigma_*$. That is,

  $$E(x_1, x_2, z; x_1', x_2', z') \wedge \chi(x_1, x_2, z, x) \to \chi(x_1', x_2', z', x).$$

  Indeed, if $E(x_1, x_2, z; x_1', x_2', z')$, then there are two cases: either $\phi(z) \wedge \phi(z')$ or $\neg\phi(z) \wedge \neg\phi(z')$. In the former case, we have both $x_1 = x_1'$ and $\chi(x_1, x_2, z, x) \leftrightarrow (\rho(x_1) = x)$. Hence $\chi(x_1', x_2', z', x)$. The second case is similar.

- The "exists" property – i.e., $\exists x\, \chi(x_1, x_2, z, x)$ – follows immediately from the fact that $\phi(z) \vee \neg\phi(z)$ and the fact that ρ_1, ρ_2 are functions.
- We show now that χ is one-to-one (relative to the equivalence relation E on $\sigma_1, \sigma_2, \sigma_1, \sigma_1$); that is,

$$\chi(x_1, x_2, z, x) \wedge \chi(x_1', x_2', z', x) \rightarrow E(x_1, x_2, z; x_1', x_2', z').$$

  Assume that $\chi(x_1, x_2, z, x) \wedge \chi(x_1', x_2', z', x)$, which expands to

$$\begin{array}{l}\big[(\phi(z) \wedge \rho_1(x_1) = x) \vee (\neg\phi(z) \wedge \rho_2(x_2) = x)\big] \\ \wedge \big[(\phi(z') \wedge \rho_1(x_1') = x) \vee (\neg\phi(z') \wedge \rho_2(x_2') = x)\big].\end{array}$$

  Since $\rho_1(y_1) \neq \rho_2(y_2)$, the first conjunct is inconsistent with the fourth, and the second is inconsistent with the third. Since ρ_1 and ρ_2 are injective, that formula is equivalent to

$$(\phi(z) \wedge \phi(z') \wedge (x_1 = x_1')) \vee (\neg\phi(z) \wedge \neg\phi(z') \wedge (x_2 = x_2')),$$

  which, of course, is $E(x_1, x_2, z; x_1', x_2', z')$.
- Finally, we show that χ is onto, i.e., $\exists z \exists x_1 \exists x_2\ \chi(x_1, x_2, z, x)$. Fix x, in which case, we have $\exists x_1(\rho_1(x_1) = x) \vee \exists x_2(\rho_2(x_2) = x)$. Since T^+ is proper, $\exists z \phi(z)$. Hence, in the case that $\exists x_1(\rho_1(x_1) = x)$, we have

$$\exists z \exists x_1(\phi(z) \wedge (\rho_1(x_1) = x)),$$

  from which it follows that

$$\exists z \exists x_1 \exists x_2\ \big[(\phi(z) \wedge \rho_1(x_1) = x) \vee (\neg\phi(z) \wedge \rho_2(x_2) = x)\big].$$

  Again, since T is proper, $\exists z \neg\phi(z)$, hence the same holds in the case that $\exists x_2(\rho_2(x_2) = x)$. In either case, $\exists z \exists x_1 \exists x_2\ \chi(x_1, x_2, z, x)$, as we needed to prove.

Thus, we have shown how to define the component of $\chi : IR \Rightarrow 1_{T^+}$ where $\sigma \in \Sigma^+$ is defined to be a coproduct sort. The other cases are simpler, and we leave them to the reader. □

This completes the proof that Morita equivalence implies weak intertranslatability. We now turn to the converse implication.

THEOREM 7.5.5 (Washington) *If T_1 and T_2 are weakly intertranslatable, then T_1 and T_2 are Morita equivalent.*

While this result is not surprising, it turns out that the proof is extremely complicated because of needing to keep track of all the newly defined symbols. Thus, before we descend into the details of the proof, we discuss the intuition behind it.

A weak translation $F : T_1 \rightarrow T_2$ doesn't necessarily map a sort symbol σ of T_1 to a sort symbol of T_2. Nor does it exactly map a sort symbol σ of T_1 to a "product" $\sigma_1 \times \cdots \times \sigma_n$ of sort symbols of T_2, because the domain formula D_F restricts to a "subsort" $F_\bullet(\sigma)$ of $\sigma_1 \times \cdots \times \sigma_n$. What's more, the equality relation $=_\sigma$ is translated to the equivalence relation E_σ, which means that σ is really translated into something

like the "quotient sort" of $F_\bullet(\sigma)$ modulo E_σ. In what follows, we will frequently write $F(=_\sigma)$ instead of E_σ in order to explicitly indicate the reconstrual F.

Now, notice that each of the constructions we mentioned earlier is permitted in taking a Morita extension of T_2. Intuitively, then, T_2 has a Morita extension T_2^+ that has enough sorts so that the translation $F : T_1 \to T_2$ can be extended to a one-dimensional translation $\hat{F} : T_1 \to T_2^+$, i.e., such that $\hat{F}(\sigma)$ is a single sort symbol of T_2^+. Intuitively, then, this extended translation $\hat{F}$ should be one-half of a homotopy equivalence in the strict sense.

One can then repeat this process to define a one-dimensional translation $\hat{G} : T_2 \to T_2^+$. Then, using the reductions $R_i : T_i^+ \to T_i$, one hopes to show that T_1^+ and T_2^+ are intertranslatable in the strict (one-dimensional) sense, which entails that they have a common definitional extension.

In practice, there are many complications in working out this idea. Thus, in the following proof, it will be convenient to allow ourselves a liberalized notion of a Morita extension where we can, in one step, add subsorts of product sorts. Suppose that Σ has sort symbols $\sigma_1, \ldots, \sigma_n$, and a formula $\phi(\vec{x})$, with $x_i : \sigma_i$, and such that $T \vdash \exists \vec{x} \phi(\vec{x})$. Then we may take

$$\Sigma^+ = \Sigma \cup \{\sigma\} \cup \{\pi_1, \ldots, \pi_n\},$$

where $\pi_i : \sigma \to \sigma_i$, and we may add explicit definitions that specify σ as the subsort of $\sigma_1 \times \cdots \times \sigma_n$ determined by the formula $\phi(\vec{x})$:

1. The projections π_i are jointly injective, i.e.,

$$\bigwedge_{i=1}^{n} (\pi_i(x) = \pi_i(y)) \to (x = y).$$

2. The projections π_i are jointly surjective, with image in $\phi(\vec{x})$, i.e.,

$$\phi(x_1, \ldots, x_n) \leftrightarrow \exists x : \sigma \bigwedge_{i=1}^{n} (\pi_i(x) = x_i).$$

This liberalized notion of Morita equivalence is clearly equivalent to the original. So, there is no harm in allowing the direct construction of subsorts $\sigma \rightarrowtail \sigma_1 \times \cdots \times \sigma_n$, given that there is an appropriate formula $\phi(\vec{x})$.

Proof Let T_1 be a Σ_1-theory and T_2 a Σ_2 theory that are intertranslatable by the translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$, and homotopies $\chi : GF \cong 1_{T_1}$ and $\chi' : FG \cong 1_{T_2}$. We will create Morita extensions of T_1 and T_2 in several stages, first defining new sort symbols and then defining new relation and function symbols.

Step 1: Suppose that $\sigma \in \Sigma_1$ is a sort symbol and that $F(\sigma) = F(\sigma)_1, \ldots, F(\sigma)_n$. Let $F_\bullet(\sigma)$ be a new sort symbol, and let

$$\Sigma_2^1 = \Sigma_2 \cup \{F_\bullet(\sigma) \mid \sigma \in S_1\} \cup \{\pi_{F(\sigma)_i} \mid \sigma \in S_1\},$$

where $\pi_{F(\sigma)_i}$ is a function symbol of sort $F_\bullet(\sigma) \to F(\sigma)_i$. Let T_2^1 be the Morita extension of T_2 that defines $F_\bullet(\sigma) \rightarrowtail F(\sigma)_1 \times \cdots \times F(\sigma)_n$, with projections $\pi_{F(\sigma)_i}$, using the domain formula $D_F(\vec{x})$.

Similarly, let

$$\Sigma_1^1 \;=\; \Sigma_1 \cup \{G_\bullet(\sigma) \mid \sigma \in S_2\} \cup \{\pi_{G(\sigma)_i} \mid \sigma \in S_2\},$$

and let T_1^1 be the Morita extension of T_1 that defines each such $G_\bullet(\sigma)$ as a product of $G(\sigma)_1, \ldots, G(\sigma)_m$, with projections $\pi_{G(\sigma)_i}$.

Before proceeding to the next step, recall that $G(=_\sigma)$ is a T_1-provable equivalence relation on the domain $D_G(\vec{x}) \rightarrowtail G(\sigma)_1, \ldots, G(\sigma)_n$. Thus, we can use the projections $\pi_i \equiv \pi_{G(\sigma)_i}$ to define an equivalence relation $G_\bullet(=_\sigma)(x, y)$ on $G_\bullet(\sigma)$:

$$G_\bullet(=_\sigma)(x, y) \;\equiv\; G(=_\sigma)\,(\pi_1(x), \ldots, \pi_n(x); \pi_1(y), \ldots, \pi_n(y))\,.$$

Step 2: For $\sigma \in S_2$, we use T_1^1 to define σ as the quotient of $G_\bullet(\sigma)$ modulo $G_\bullet(=_\sigma)$. Let

$$\Sigma_1^2 \;=\; \Sigma_1^1 \cup \{\sigma \mid \sigma \in S_2\} \cup \{\epsilon_\sigma \mid \sigma \in S_2\},$$

where ϵ_σ is a new function symbol of sort $G_\bullet(\sigma) \to \sigma$. Let δ_σ be the explicit definition

$$\delta_\sigma \;\equiv\; ((\epsilon_\sigma(x) = \epsilon_\sigma(y)) \leftrightarrow G_\bullet(=_\sigma)(x, y)) \wedge \forall y \exists x (\epsilon_\sigma(x) = y). \tag{7.2}$$

We then define a Morita extension

$$T_1^2 \;=\; T_1^1 \cup \{\delta_\sigma \mid \sigma \in S_2\}.$$

Similarly, let

$$\Sigma_2^2 \;=\; \Sigma_2^1 \cup \{\sigma \mid \sigma \in S_1\} \cup \{\epsilon_\sigma \mid \sigma \in S_1\},$$

where $\epsilon_\sigma : F_\bullet(\sigma) \to \sigma$, and let T_2^2 be the Morita extension of T_2^1 that defines each $\sigma \in S_1$ as a quotient sort.

Before proceeding to the next step, we show that T_1^2 defines a functional relation ξ from the domain D_{GF} to $G_\bullet(F(\sigma)_1), \ldots, G_\bullet(F(\sigma)_n)$ or, more precisely, to the image of the latter in $GF(\sigma)$. Recall that the domain formulas of the composite GF are given by the general recipe $D_{GF} = G(D_F)$; and that G is defined so that

$$G(\phi)(\vec{x}_1, \ldots, \vec{x}_n) \;\vdash\; D_G(\vec{x}_i),$$

for any Σ_2-formula ϕ. Thus, $D_{GF}(\vec{x}_1, \ldots, \vec{x}_n) \vdash D_G(\vec{x}_i)$. Furthermore, $G_\bullet(F(\sigma)_i)$ is defined as a subsort of $G(F(\sigma)_i)_1, \ldots, G(F(\sigma)_i)_m$ via the formula $D_G(\vec{x}_i)$.

$$\begin{array}{ccc} & & D_{GF}(\vec{x}_1, \ldots, \vec{x}_n) \\ & & \big\downarrow \xi \\ G_\bullet(F(\sigma)_1), \ldots, G_\bullet(F(\sigma)_n) & \xrightarrow{\;\zeta\;} & D_G(\vec{x}_1) \wedge \cdots \wedge D_G(\vec{x}_n) \\ & & \big\downarrow \\ & & GF(\sigma) \end{array}$$

Step 3: In Step 1, we equipped T_2^1 with subsorts $F_\bullet(\sigma) \rightarrowtail F(\sigma)_1, \ldots, F(\sigma)_n$. Now we add these sorts to T_1^2 as well. Given $\sigma \in \Sigma_1$, each $F(\sigma)_i$ is a sort in Σ_2, hence by Step 2, also in Σ_1^2. Now let

$$\Sigma_1^3 \;=\; \Sigma_1^2 \cup \{F_\bullet(\sigma) \mid \sigma \in S_1\} \cup \{\pi_{F(\sigma)_i} \mid \sigma \in S_1\},$$

where $\pi_{F(\sigma)_i}$ is the Σ_2^1 function symbol of sort $F_\bullet(\sigma) \to F(\sigma)_i$. In order to define $F_\bullet(\sigma)$, we need an appropriate formula $U(x_1, \ldots, x_n) \rightarrowtail F(\sigma)_1, \ldots, F(\sigma)_n$. We choose the image of D_{GF} under the function $\rho \equiv \epsilon_{F(\sigma)_1} \wedge \cdots \wedge \epsilon_{F(\sigma)_n}$.

$$\begin{array}{ccc}
D_{GF}(\vec{x}_1, \ldots, \vec{x}_n) & \overset{\rho}{\dashrightarrow} & U(x_1, \ldots, x_n) \\
\downarrow & & \vdots \\
D_G(\vec{x}_1) \wedge \cdots \wedge D_G(\vec{x}_n) & \xrightarrow{\rho} & F(\sigma)_1, \ldots, F(\sigma)_n \\
\downarrow & & \\
GF(\sigma) & &
\end{array}$$

That is,

$$U(x_1, \ldots, x_n) \;\equiv\; \exists \vec{x}_1 \ldots \exists \vec{x}_n \left(D_{GF}(\vec{x}_1, \ldots, \vec{x}_n) \wedge \bigwedge_{i=1}^{n} (\epsilon_{F(\sigma)_i}(\vec{x}_i) = x_i) \right).$$

Since $T_1 \vdash \exists X\, D_{GF}(X)$, it follows that $T_1^2 \vdash \exists \vec{x}\, U(\vec{x})$. Thus, we can use $U(\vec{x})$ to define $F_\bullet(\sigma)$ as a subsort of $F(\sigma)_1, \ldots, F(\sigma)_n$, and we let T_1^3 denote the resulting Morita extension of T_1^2.

Similarly, let

$$\Sigma_2^3 \;=\; \Sigma_2^2 \cup \{G_\bullet(\sigma) \mid \sigma \in S_2\} \cup \{\pi_{G(\sigma)_i} \mid \sigma \in S_2\},$$

and let T_2^3 be the Morita extension of T_2^2 that defines each $G_\bullet(\sigma)$ as a subsort of $G(\sigma)_1, \ldots, G(\sigma)_n$.

Step 4: Let Σ_1^4 be the union of Σ_1^3 with all relation and function symbols from Σ_2. We extend T_1^3 to T_1^4 by adding explicit definitions for all the new symbols. For notational simplicity, we treat only the case of a predicate symbol $p \in \Sigma_2$ of sort $\sigma \in \Sigma_2$. We leave the other cases to the reader. Recall that T_1^3 defines $\epsilon_\sigma : G_\bullet(\sigma) \to \sigma$ as a quotient, and also the projections $\pi_{G(\sigma)_i} : G_\bullet(\sigma) \to G(\sigma)_i$ can be conjoined to give a bijection θ between $G_\bullet(\sigma)$ and $D_G(\vec{x})$.

$$\begin{array}{ccccc}
G(p)(\vec{x}) & \dashleftarrow & G_\bullet(p) & \dashrightarrow & \phi_p(x) \\
\downarrow & & \vdots & & \vdots \\
D_G(\vec{x}) & \xleftarrow{\theta} & G_\bullet(\sigma) & \xrightarrow{\epsilon_\sigma} & \sigma
\end{array}$$

To define ϕ_p, first pull $G(p)(\vec{x})$ back along π to obtain $G_\bullet(p)$; then take the image of $G_\bullet(p)$ under ϵ_σ. That is,

$$\phi_p(x) \;\equiv\; \exists y\; (G(p)(\pi_1(y), \ldots, \pi_n(y)) \wedge (\epsilon_\sigma(y) = x))\,.$$

Recall that

$$G(=_\sigma)(\pi_1(y),\ldots,\pi_n(y);\pi_1(z),\ldots,\pi_n(z)) \vdash \epsilon_\sigma(y) = \epsilon_\sigma(z),$$

and also that

$$G(p)(\vec{x}), G(=_\sigma)(\vec{x},\vec{y}) \vdash G(p)(\vec{y}).$$

Hence the preceding diagram defines a functional relation from $G(p)(\vec{x})$ to $\phi_p(x)$, relative to the notion of equality given by $G(=_\sigma)$.

We now add explicit definitions $\delta_p \equiv p(x) \leftrightarrow \phi_p(x)$ for each relation symbol $p \in \Sigma_2$, creating a Morita extension T_1^4 of T_1^3. We perform the analogous construction to obtain extensions $\Sigma_2^4 \supseteq \Sigma_2^3$ and $T_2^4 \supseteq T_2^3$.

Before proceeding, we note that at this stage, the expanded signature Σ_2^4 has copies of the Σ_1-formulas D_{GF} and χ that define the homotopy $\chi : GF \Rightarrow 1_T$ for T.

$$\begin{array}{ccc} & & F_\bullet(\sigma) \\ & & \downarrow{\scriptstyle \epsilon_\sigma} \\ D_{GF}(\vec{x}_1,\ldots,\vec{x}_n) & \xrightarrow{\chi} & \sigma \end{array}$$

Step 5: In Step 3, we equipped T_2^3 with function symbols $\epsilon_\sigma : F_\bullet(\sigma) \to \sigma$, for $\sigma \in S_1$. We now add these function symbols to T_1^4 as well. Let

$$\Sigma_1^5 = \Sigma_1^4 \cup \{\epsilon_\sigma \mid \sigma \in S_1\}.$$

We need to find a Σ_1^4-formula that can serve as a suitable definiens for ϵ_σ. We construct a span of relations.

$$\begin{array}{ccccc} D_G(\vec{x}_1) \wedge \cdots \wedge D_G(\vec{x}_n) & \longleftarrow & D_{GF}(\vec{x}_1,\ldots,\vec{x}_n) & \xrightarrow{\chi} & \sigma \\ \downarrow{\scriptstyle \rho} & & \downarrow{\scriptstyle \rho} & & \\ F(\sigma)_1,\ldots,F(\sigma)_n & \longleftarrow & U(x_1,\ldots,x_n) & & \\ & & \downarrow{\scriptstyle \theta} & & \\ & & F_\bullet(\sigma) & & \end{array}$$

Here $D_G(\vec{x}_i)$ is the domain formula corresponding to the assignment $F(\sigma)_i \mapsto G(F(\sigma)_i)$; and $\rho \equiv \epsilon_{F(\sigma)_1} \wedge \cdots \wedge \epsilon_{F(\sigma)_n}$, where $\epsilon_{F(\sigma)_i} : G(F(\sigma)_i) \to F(\sigma)_i$ defines $F(\sigma)_i$ as a quotient sort via the equivalence relation $G(=_{F(\sigma)_i})$; and θ is given by

$$\theta(x_1,\ldots,x_n;y) \equiv U(x_1,\ldots,x_n) \wedge \bigwedge_{i=1}^{n}(x_i = \pi_i(y)),$$

for $U(x_1,\ldots,x_n)$, as defined in Step 3. Here θ is a bijection, so we ignore it. We show that the span of $\rho : D_{GF} \to U$ and $\chi : D_{GF} \to \sigma$ defines a functional relation from U to σ.

Since the homotopy formula χ is well defined relative to the equivalence relation $GF(=_\sigma)$, and surjective onto σ, we have

$$GF(=_\sigma)(Y,Z) \vdash \exists_{\sigma=1} x\,(\chi(Y,x) \wedge \chi(Z,x)).$$

Here we have used $Y = \vec{y}_1, \ldots, \vec{y}_n$ and $Z = \vec{z}_1, \ldots, \vec{z}_n$ for sequences of variables of sort $GF(\sigma)$. It will suffice then to show that

$$\rho(Y; x_1, \ldots, x_n), \rho(Z; x_1, \ldots, x_n) \vdash GF(=_\sigma)(Y, Z). \tag{7.3}$$

Now, the definition of ρ yields

$$\rho(Y; x_1, \ldots, x_n), \rho(Z; x_1, \ldots, x_n) \vdash \bigwedge_{i=1}^{n} G(=_{F(\sigma)_i})(\vec{y}_i, \vec{z}_i). \tag{7.4}$$

Moreover, since F is a translation, T_2 entails that $F(=_\sigma)$ is an equivalence relation on $F(\sigma)_1, \ldots F(\sigma)_n$. Hence, by reflexivity,

$$\bigwedge_{i=1}^{n}(y =_{F(\sigma)_i} z) \vdash F(=_\sigma)(y_1, \ldots, y_n; z_1, \ldots, z_n).$$

Since $G : T_2 \to T_1$ is a translation, the substitution theorem gives

$$\bigwedge_{i=1}^{n} G(=_{F(\sigma)_i})(\vec{y}_i, \vec{z}_i) \vdash GF(=_\sigma)(Y, Z). \tag{7.5}$$

The implications (7.4) and (7.5) together show that $\chi \circ \rho^{-1}$ is a functional relation from U to σ, where equality on the former is given by $GF(=_\sigma)$. Thus, $\chi \circ (\theta \circ \rho)^{-1}$ is a functional relation from $F_\bullet(\sigma)$ to σ. Using ψ to denote this relation, we introduce the explicit definition

$$\delta_{\epsilon_\sigma} \equiv (\epsilon_\sigma(x) = y) \leftrightarrow \psi(x, y), \tag{7.6}$$

and we define a Morita extension

$$T_1^5 = T_1^4 \cup \{\delta_{\epsilon_\sigma} \mid \sigma \in S_1\}.$$

We define a Morita extension T_2^5 of T_2^4 in an analogous fashion. Therefore, $\Sigma_1^5 = \Sigma_2^5$. This completes our construction of the Morita extensions T_1^5 of T_1, and T_2^5 of T_2.

We will now show that T_1^5 and T_2^5 are logically equivalent, thereby establishing the Morita equivalence of T_1 and T_2. To this end, note first that since T_1^5 is a Morita extension of T_1, the two theories are intertranslatable, and similarly for T_2^5 and T_2. (Note that the construction does not use coproduct sorts. Hence, the result holds even when T_1 and T_2 are not proper theories.) Composing these translations gives translations $F : T_1^5 \to T_2^5$ and $G : T_2^5 \to T_1^5$ that extend the original translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$. We will use these translations to show that T_1^5 and T_2^5 have the same models in their shared signature $\Sigma_1^5 = \Sigma_2^5$. The intuition behind the result is clear: since T_1^5 is a Morita extension of T_1, each model of T_1 uniquely expands to a model of T_1^5, and similarly for T_2 and T_2^5. Since the original model functor $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ is an equivalence of categories, the lifted model functor $F^* : \mathrm{Mod}(T_2^5) \to \mathrm{Mod}(T_1^5)$ is also an equivalence of categories. We proceed now to the details of the argument.

Recall that we defined a reconstrual $\tilde{F} : \Sigma_1 \to \Sigma_2^3$ that is constant on sorts. (Hence, we may treat $\tilde{F}$ as a reconstrual in the more narrow sense.) We extend $\tilde{F}$ as usual to a

map from Σ_1- formulas to Σ_2^3 formulas. Since F is a translation, $\tilde{F}$ is also a translation. Thus, the corresponding model map $\tilde{F}^*$ has the feature that

$$(\tilde{F}^* M)(\phi) = M(\tilde{F}(\phi)),$$

for each Σ_1-formula ϕ. In particular, $\tilde{F}^* M \vDash \phi$ iff $M \vDash \tilde{F}(\phi)$. The translation $\tilde{F} : T_1 \to T_2^3$ also has the feature that $T_2^5 \vdash F(\phi) \leftrightarrow \tilde{F}(\phi)$, for any sentence ϕ of T_1.

Now let M be a model of T_2^5. First we show that $M \vDash \phi$ for any Σ_1-sentence ϕ such that $T_1 \vdash \phi$. Since M satisfies the explicit definitions we gave for all the symbols in Σ_1, it follows (by induction) that $M(\tilde{F}(\phi)) = M(\phi)$ for any Σ_1-formula ϕ. Since M is a model of T_2^3, $F^* M$ is a model of T_1, and $\tilde{F}^* M \vDash \phi$. By the previous paragraph, $M \vDash \tilde{F}(\phi)$, hence $M \vDash F(\phi)$, and, therefore, $M \vDash \phi$.

We now show that M satisfies the explicit definitions we added to T_1^5 in Steps 1–5. In Step 1, we added the definition of $G(\sigma)$ as a product sort. However, we added the same definition to T_2^3 in Step 3. Thus, since M is a model of T_2^3, these definitions are satisfied by M.

In Step 2, we expand Σ_2^1 to Σ_2^2 by adding sort symbols $\sigma \in S_1$ and function symbols $\epsilon_\sigma : F_\bullet(\sigma) \to \sigma$, and we let T_2^2 define $\epsilon_\sigma : F_\bullet(\sigma) \to \sigma$ as a quotient map corresponding to the equivalence relation $F_\bullet(=_\sigma)$. Hence, in any model M of T_2^5, we have

$$\epsilon_\sigma(a) = \epsilon_\sigma(b) \qquad \text{iff} \qquad F_\bullet(=_\sigma)(a, b),$$

for $a, b \in M_{F_\bullet(\sigma)}$. Recall also that T_2^5 explicitly defines $F_\bullet(\sigma)$ as a subsort of $F(\sigma)_1, \ldots, F(\sigma)_n$, and that

$$F_\bullet(=_\sigma)(a, b) \qquad \text{iff} \qquad F(=_\sigma)(\vec{a}, \vec{b}).$$

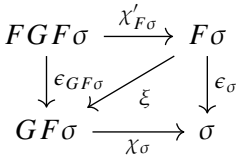
In Step 5, we stipulate that $T_1^5 \vdash \delta_{\epsilon_\sigma}$, where δ_{ϵ_σ} is the explicit definition:

$$\delta_{\epsilon_\sigma} \equiv (\epsilon_\sigma(x) = z) \leftrightarrow (\chi \circ \rho^{-1})(x, z).$$

We need to show that $T_2^5 \vdash \delta_{\epsilon_\sigma}$, and for this, we need to see how T_2^5 defines the symbols χ and ρ. First, $\chi : D_{GF} \to \sigma$ is the homotopy map, which is originally a Σ_1-formula. Thus, the symbols in χ are explicitly defined by T_2^4 in Step 4.

Next, $\rho \equiv \epsilon_{F(\sigma)_1} \wedge \cdots \wedge \epsilon_{F(\sigma)_n}$, where $F(\sigma)_i$ is a Σ_2 sort symbol, and $\epsilon_{F(\sigma)_i} : G_\bullet(F(\sigma)_i) \to F(\sigma)_i$ is a function symbol. In Step 1, we have T_1^1 define $G_\bullet(F(\sigma)_i)$ as a sub-product sort of $G(F(\sigma)_i)_1, \ldots, G(F(\sigma)_i)_m$. In Step 5, we have T_2^5 define the function symbol $\epsilon_{F(\sigma)_i}$ in terms of the Σ_2 homotopy map χ'.

We need to show now that $M \vDash \delta_{\epsilon_\sigma}$ or, in other words, that $\epsilon_\sigma(x) = z$ and $\psi(x, z)$ define the same relation in M. We can show that the following diagram commutes (where the objects are meant to be domains of the sort symbols in the model M).



We can thus characterize χ_σ as the map that makes the preceding diagram commute. A key observation is that $F(\chi_\sigma) = \chi'_{F(\sigma)}$ for each sort $\sigma \in S_1$. □

TECHNICAL ASIDE 7.5.6 The sheer complexity of the previous proof shows one reason why it can be convenient to move to the context of categorical logic, where theories are treated as certain kinds of categories. We conjecture that a more intuitive (but conceptually laden) proof of this result could be obtained as follows.

Each first-order theory T has a unique classifying (Boolean) pretopos in the sense of Makkai (1987). Intuitively speaking, T and T' should have the same classifying pretopos iff T and T' are weakly intertranslatable in the sense we have described here. Furthermore, Tsementzis (2017b) shows that T and T' have the same classifying pretopos iff T and T' are Morita equivalent.

Having completed this result, we now have a much clearer picture of the various options for a precise notion of theoretical equivalence. We have placed the most salient options in the following chart.

$$\begin{array}{ccccc} \text{SI} & \Longrightarrow & \text{WI} & \Longrightarrow & \text{CE} \\ \Updownarrow & & \Updownarrow & & \\ \text{DE} & & \text{ME} & & \end{array}$$

Here "I" represents the intertranslatability notions (strong and weak), and "E" represents the equivalence notions (definitional, Morita, and categorical). In this chart, the further to the right, the more liberal the notion of theoretical equivalence, and the fewer the invariants of equivalence. For example, if $F : T \to T'$ is a strong (one-dimensional) translation, then the dual functor $F^* : \text{Mod}(T') \to \text{Mod}(T)$ preserves the size of the underlying domains of models, which isn't necessarily the case for Morita equivalent theories. Similarly, if $F : T \to T'$ is a weak translation, then $F^* : \text{Mod}(T') \to \text{Mod}(T)$ preserves ultraproducts, which isn't necessarily the case for an arbitrary categorical equivalence between $\text{Mod}(T')$ and $\text{Mod}(T)$.

7.6 Open Questions

We do not mean to give the impression that we have answered all of the interesting questions that could be raised about theories and the relations between them. Quite to the contrary, we hope that our investigations serve to reinvigorate the sort of "exact philosophy" that Rudolf Carnap envisioned. We conclude this section, then, with a list of some open questions and lines of investigation that might be pursued.

1. We encourage philosophers of science to return to previous discussions of specific scientific theories, where claims of equivalence (or inequivalence) play a central role, but where the relevant notion of equivalence was not explicated. Can the tools we have developed here help clarify the commitments that led to certain judgments of equivalence or inequivalence?

2. It would be interesting to look again at the possibilities for providing perspicuous first-order formalizations of interesting scientific theories. Some work in this direction continues, e.g., with the Budapest group working on axiomatizations of relativity theory (see Andréka and Németi, 2014).
3. Some theories are so strong that new sorts (e.g., product sorts) seem to be encoded already into the original sorts. For example, in Peano arithmetic, n-tuples of natural numbers can be encoded as individual natural numbers. This encoding could perhaps be represented as an isomorphism $f : (\sigma \times \sigma) \to \sigma$ in a Morita extension T^+ of T. One might conjecture that for theories like Peano arithmetic, strict (one-dimensional) intertranslatability is equivalent to weak (many-dimensional) intertranslatability.
4. It's tempting to think that one could resort to "ontological maximalism" in the following sense: for a model M of theory T, the ontology for M consists of *all* the objects in every set that can be *constructed from* the domain M or, if the theory is many-sorted, from the domains $M(\sigma_1), \ldots, M(\sigma_n)$. (This idea is in the spirit of the suggestion of Hawthorne [2006].)

   There are three immediate difficulties with this proposal. First, this proposal would make the ontology of every nontrivial theory infinite. In particular, infinitely many distinct elements occur in the tower of Cartesian products: $M, M \times M, M \times M \times M, \ldots$ And that's even before we construct equivalence classes and coproducts from these sets. Second, it's not clear which constructions should be permitted. Should we allow the constructions from a Morita extension, or should we also allow, say, the construction of powersets? Third and finally, ontological maximalism runs contrary to the spirit of Ockham's razor.
5. One might worry that the definition of Morita equivalence is *arbitrary*. Why do we allow the particular definitions we do, and not others? Is there any intrinsic motivation for this choice? There is an intuition that the definitions permitted in a Morita extension are precisely those definitions that can be expressed in first-order language. How can we make that intuition precise?

7.7 Notes

- The notion of a dual functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ makes an appearance already in Makkai and Reyes (1977), who explore the correspondence between properties of F and properties of F^*. This exploration is part of their proof of the "conceptual completeness" of coherent logic.

  However, for Makkai and Reyes, first-order theories are replaced by coherent theories, and the latter are replaced by their corresponding pretoposes – all of which make their discussion a bit inaccessible for most philosophers. For an even more sophisticated investigation in this direction, see Breiner (2014). The dual functor makes an appearance in classical first-order logic in Gajda et al. (1987). The dual functor also seems to be quite closely related to the notion of a "model mapping" due to Gaifman (see Myers, 1997).

In later work, Makkai (1991) explores the question of which functors G : $\text{Mod}(T') \rightarrow \text{Mod}(T)$ are duals of translations. He makes some progress by assuming that $\text{Mod}(T)$ and $\text{Mod}(T')$ are not just categories, but *ultracategories*, i.e., categories with ultraproduct structure. In this case, the dual functors are those that preserve the ultraproduct structure.

- The proof that Morita equivalence implies categorical equivalence is from Barrett and Halvorson (2016b). In one sense, the result was no surprise all: the notion of Morita equivalence for first-order theories was modelled after the notion of Morita equivalence in categorical logic, i.e., when two theories T and T' have equivalent classifying toposes $\mathscr{E}_T$ and $\mathscr{E}_{T'}$. And when $\mathscr{E}_T \simeq \mathscr{E}_{T'}$, standard topos-theoretic methods show that $\text{Mod}(T) \simeq \text{Mod}(T')$ (see Johnstone, 2003, D1.4.13). Tsementzis (2017b) calls the notion we use here "T-Morita equivalence," and he gives a precise description of the relation between it at the topos-theoretic notion.
- The Morita equivalence of point and line geometries was demonstrated by Barrett and Halvorson (2017a). The arguments about geometry are novel, but not without precedent. Beth and Tarski (1956), Scott (1956), Tarski (1956), Robinson (1959), and Royden (1959) focus on the relationships between formulations of geometry that use different primitive *predicate* symbols, but not different primitive *sort* symbols. Szczerba (1977) and Schwabhäuser et al. (1983) take crucial steps toward capturing the relationships between geometries with different sorts but do not explicitly prove their equivalence. Andréka et al. (2008) and Andréka and Németi (2014), however, introduce a collection of tools from definability theory that allows one to demonstrate a precise equivalence.
- The proof that Morita equivalence coincides with weak intertranslatability is due to Washington (2018).

8 From Metatheory to Philosophy

Much of twentieth-century analytic philosophy was concerned – when not explicitly, then implicitly – with theories and with the relationships between them. For example, is every spacetime theory equivalent to one with Euclidean background geometry? Or is folk psychology reducible to neuroscience? Or can there be a good reason to choose a theory over an empirically equivalent rival theory?

But what is a theory? And what does it mean to say that two theories are equivalent or that one theory is reducible to another? Carnap had the audacious idea that philosophy can follow mathematics' method of explication: to take an intuitive notion and to find a nearby neighbor in the realm of precisely defined mathematical concepts. In this book, we've tried to follow Carnap's lead; and indeed, we hope that we've done a bit better than Carnap, because mathematics has come a long way in the past hundred years. We now have mathematical concepts – such as categories, functors, and natural transformations – the likes of which Carnap never dreamed about.

In this book, we've attempted to explicate the concept of a theory, as well as some of the relations between theories that scientists and philosophers find it useful to discuss. With these explications in the background, we can now return to some of the big questions of philosophy of science, such as, "what is the proper attitude to take toward a successful scientific theory?"

8.1 Ramsey Sentences

No analytic philosopher's education is complete until she learns the magic of the Ramsey sentence. The idea was proposed by Frank Ramsey (1929) and was reinvented by Carnap in the 1950s – or, more accurately, Carnap forgot that he learned about it from Herbert Bohnert (see Psillos, 2000). Most contemporary philosophers know of the idea because David Lewis (1970) argued that it solves the problem of theoretical terms. In the years since Lewis' seminal paper, Ramsey sentences have become a sort of deus ex machina of analytic philosophy.

Let's start with a simple example. Suppose that P is a theoretical predicate and that O is an observational predicate. (Or, in Lewis' preferred terminology, O is antecedently understood vocabulary, and P is new vocabulary.) Now suppose that our theory T consists of a single sentence $P(c) \rightarrow O(c)$, which might be paraphrased as saying that $O(c)$ is an empirical sign that $P(c)$. (Here c is a constant symbol. We omit

first-order quantifiers to keep things simple.) To form the Ramsey sentence of T, we simply perform an instance of second-order existential generalization:

$$\frac{P(c) \to O(c)}{\exists X(X(c) \to O(c))}.$$

The sentence below the line is called the Ramsey sentence T^R of the theory T. Thus, while the original theoretical statement T mentions some particular property P, the Ramsey sentence T^R simply says that there is some or other property that plays the appropriate role. It may feel – and has felt to many philosophers – that the truth of T^R somehow magically endows the term P with meaning. In particular, philosophers are wont to say things like, "P is whatever it is that plays the role described by T^R."

Since Ramsey sentences draw upon the resources of second-order logic, the neophyte is left to wonder: does the philosophical magic here depend on something special that happens in second-order logic, something that only the most technically sophisticated philosophers can understand? We think that the answer to this question is no. In fact, Ramsifying a theory simply weakens that theory in the same way that existentially quantifying a first-order sentence weakens that sentence. Consider the following pedestrian example.

Example 8.1.1 Let $\Sigma = \{m\}$, where the name m is a theoretical term. Let T be the theory $\exists x(x = m)$ in Σ. Then the Ramsey sentence T^R of T is the sentence $\exists x(x = x)$, which is just a tautology. That is, T^R is the empty theory in the empty signature. It is easy to see that the inclusion $I : T^R \to T$ is conservative but not essentially surjective. In particular, there is no formula ϕ of Σ such that $(I\phi)(x) \equiv (x = m)$. The fact that I is not essentially surjective corresponds to the fact that $I^* : \mathrm{Mod}(T) \to \mathrm{Mod}(T^R)$ is not full. Here I^* is the functor that takes a model of T and forgets the extension of m. In general, then, I^*M has more symmetries than M.

We can be yet more precise about the differences between $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T^R)$. In short, a model of T^R is simply a nonempty set X (and two such models are isomorphic if they have the same cardinality). For each $p \in X$, there is a corresponding model X_p of T where $X_p(m) = p$. For a fixed X, and $p, q \in X$, there is an isomorphism $h : X_p \to X_q$ that maps p to q. However, the automorphism group of X_p is smaller than the automorphism group of X. Indeed, $\mathrm{Aut}(X_p)$ consists of all permutations of X that fix p, hence is isomorphic to $\mathrm{Aut}(X \backslash \{p\})$.

We can see then that T and T^R are not intertranslatable (or definitionally equivalent). Nonetheless, there is a sense in which mathematicians would have no qualms about passing from T^R to the more structured theory T. Indeed, once we've established that the domain X is nonempty (which, of course, is a presupposition of first-order logic), we could say, "let m be one of the elements of X." This latter statement does not involve any further theoretical commitment over what T^R asserts. ⌟

Our advice then to the neophyte is not to allow herself to be intimidated by second-order quantification. In fact, we will argue that passage from a theory T to its

Ramsified version T^R either forgets too much of what the original theory said or says *more* than what the original theory said – depending on which notion of second-order logical equivalence one adopts. Before we do this, let's pause to recall just how much philosophical work Ramsey sentences have been asked to do. We will look at three applications. First, Carnap claims that Ramsey sentences solve the problem of dividing the analytic and synthetic parts of a scientific theory. Second, Lewis claims that Ramsey sentences solve the problem of theoretical terms and, in particular, the problem of giving meaning to "mentalese" in a physical world. Third, contemporary structural realists claim that Ramsey sentences give a way of isolating the structural claims of a scientific theory.

Carnap's Irenic Realism

One theme running throughout Carnap's work is a rejection of what he sees as false dilemmas. In one sense, Carnap is one of the most pragmatic philosophers ever in the Western tradition, as he places extreme emphasis on questions such as: what questions are worth asking, and what problems are worth working on? Now, one can imagine a philosophy graduate student asking herself: what question should I try to answer in my dissertation? If she's a particularly ambitious (or perhaps overconfident) student, she might decide to determine whether materialism or dualism is true. Or she might decide to determine whether scientific realism or instrumentalism is true. Carnap's advice to her would be to work on such questions is not a good use of your time.

In the early twentieth century, the debate between scientific realism and instrumentalism centered around the question: do theoretical entities – i.e., the things named by scientific theories, but which are not evident in our everyday experience – exist? Or, shifting to a more explicitly normative manner of speech: are we entitled to believe in the existence of these entities, and perhaps even obliged to do so? The realist says yes to these questions, and the instrumentalist says no. Carnap attempts to steer a middle way. He says that the questions are ill-posed.

Toward the end of his career, Carnap hoped that Ramsey sentences could help show why there is no real argument between realism and instrumentalism. In particular, if T is a scientific theory containing some theoretical terms $r_1, \ldots, r_n$, then Carnap parses T into two parts: the Ramsey sentence T^R and the sentence $T^R \rightarrow T$ that has since been dubbed the "Carnap sentence." Carnap claims that the Ramsey sentence T^R gives the empirical (synthetic) content of T, whereas $T^R \rightarrow T$ gives the definitional (analytic) part of T. The latter claim can be made plausible by realizing that $T^R \rightarrow T$ is trivially satisfiable, simply by stipulating appropriate extensions for $r_1, \ldots, r_n$.

Psillos (2000) argues that Carnap's equation of synthetic content with the Ramsey sentence makes him a structural realist – in which case he is subject to Newman's objection, which impales him on the horns of the realism–instrumentalism dilemma. Friedman (2011) disagrees, arguing that Carnap's invocation of the Ramsey sentence successfully implements his neutralist stance. Debate on this issue continues in the literature – see, e.g., Uebel (2011); Beni (2015).

Ramsey Sentence Functionalism

In the philosophy of mind, Ramsey sentences came to play a central role through the work of Lewis (1966, 1972, 1994) and, more generally, in a point of view known as **functionalism**. To be sure, Lewis claims not to know whether or not he is a functionalist, and most functionalists don't talk explicitly about Ramsey sentences. However, by the 1980s, the connection between functionalism and Ramsey had been firmly established (see Shoemaker, 1981).

Around 1970, materialist reductionism had gone out of style. Philosophers concluded that folk psychology cannot, and should not, be reduced – neither to descriptions of behavior nor to physiological descriptions. However, philosophers weren't ready to give up the physicalist project, and, in particular, they didn't want to entertain the possibility that there is an autonomous realm of mental objects or properties. The goal then is to explain how mental properties are anchored in physical properties, even if the former cannot be explicitly defined in terms of the latter.

Functionalism, and functional definitions, are supposed to provide a solution to this problem. According to functionalism, mental properties are defined by the role that they play in our total theory T, which involves both mental concepts (such as "belief" and "desire") and physical concepts (such as "smiling" or "synapse firing"). How then are we supposed to cash out this notion of being "defined by role"? It's here that Ramsey sentences are invoked as providing the best formal explication of functional definitions.

Contemporary analytic philosophers routinely mention Ramsey sentences in this connection. Nonetheless, long ago, Bealer (1978) argued that this attempt to define mental properties – call it "Ramsey sentence functionalism" – is inconsistent. According to Bealer, functionalism has both a negative and a positive theses. On the negative side, functionalism is committed to the non-reductionist thesis: mental properties (m-properties) cannot be explicitly defined in terms of physical properties (p-properties). On the positive side, m-properties *are* defined in terms of the role they play vis-à-vis each other and the p-properties.

Let T be a theory in signature $\Sigma \cup \{r_1, \ldots, r_n\}$, where we think of Σ as p-vocabulary, and of $r_1, \ldots, r_n$ as m-vocabulary. We then adopt the following proposal (which defenders of functionalism are welcome to reject or modify):

> T provides functional definitions of $r_1, \ldots, r_n$ in terms of Σ just in case, in each model M of the Ramsey sentence T^R, there are unique realizing properties $M(r_1), \ldots, M(r_n)$.

It's easy to see then that T provides functional definitions of $r_1, \ldots, r_n$ in terms of Σ only if T implicitly defines $r_1, \ldots, r_n$ in terms of Σ. Indeed, if M and N are models of T, then $M|_\Sigma$ and $N|_\Sigma$ are models of T^R, and it follows from the uniqueness clause that $M(r_i) = N(r_i)$. It then follows from Beth's theorem that T explicitly defines $r_1, \ldots, r_n$ in terms of Σ.

Bealer's argument, if successful, shows that functionalism is inconsistent: the positive thesis of functionalism entails the negation of the negative thesis. Surprisingly, however, functionalism lives on, apparently oblivious of this little problem of inconsistency. In fact, functionalism hasn't just survived; it is flourishing and spreading its

tendrils – indeed, it has become an overarching philosophical ideology: **the Canberra plan**. The goal of the Canberra plan is to find a place in the causal nexus of physical properties for all the stuff that makes up our daily lives – things like moral and aesthetic values, laws, society, love, etc. (For further discussion, see Menzies and Price [2009].)

Structural Realism

In more recent times, Ramsey sentences have been invoked in support of a trendy view in philosophy of science: structural realism. In the early 1990s, structural realism was the new kid on the block in discussions of scientific realism and antirealism. As forcefully recounted by Worrall (1989), there are good arguments against both scientific realism and scientific antirealism. Against scientific realism, there is the *pessimistic metainduction*, which points to the long history of failed scientific theories as evidence that our current favorite scientific theories will probably also fail. Against scientific antirealism, there is the *no miracles argument*, which points to the success of scientific theories as something crying out for an explanation. In good Hegelian fashion, Worrall seeks a synthesis of the extremes of realism and antirealism – a position that offers the best of both worlds. His proposal is structural realism, according to which the part of a theory to take seriously is its pronouncements on issues of *structure*.

Worrall illustrates the idea of "preserved structure" with a specific example. In particular, before Einstein's special theory of relativity, it was thought that there was a substance, the "aether," in which electric and magentic waves propagated. After the Michelson–Morley experiment and the success of special relativity, there was no longer any use for the aether. Thus, the transition to special relativity might be taken to be a particularly clear example of failed reference – showing, in particular, that pre-Einsteinian physicists ought not to have taken their theory so seriously.

Nonetheless, says Worrall, it would have been a mistake for pre-Einsteinian physicists to treat their theory instrumentally, i.e., merely as a tool for making predictions. For the form of the equations of motion was preserved through the transition to special relativity – hence, they would have done well to trust their equations. The general lesson, says Worrall, is to trust your theory's structure but not the underlying stuff it purports to be talking about.

Worrall's example is highly suggestive, and we might like to apply it in a forward-looking direction. In particular, take one of our current-day successful scientific theories T, such as quantum mechanics. The pessimistic metainduction suggests that T will be wrong about something. But can we already make an educated guess about which parts of T will be preserved and which part will go on the scrap heap with other rejected theories?

Worrall and Zahar (2001), Cruse and Papineau (2002), and Zahar (2004) provide a specific proposal for picking out the structural commitments of a theory T: they are given by its Ramsey sentence T^R. This idea certainly has some intuitive appeal – trading on an analogy to coordinate-free descriptions of space. For a naive or straightforward description of physical space, we might use triples of real numbers, i.e., the mathematical space $\mathbb{R}^3$. But now our description of space has superfluous structure. In particular,

we assigned the origin $0 \in \mathbb{R}^3$ to some particular point in space – but we didn't mean to indicate that the denoted point is any different than any other point in space. Thus, our description breaks the natural symmetry of space, and it would be natural to look for another description that respects these symmetries. Indeed, that's precisely the idea behind the move from using vector spaces to using affine spaces to describe space.

Now, just as a vector-space description of space breaks its symmetry, so our theoretical descriptions in general might fail to respect the symmetry between properties. For example, we didn't need to use the word "electron" to describe those things that are found in the energy shells around an atom's nucleus – we could simply say that something or other plays the relevant role. And that's exactly what the Ramsified theory says. Thus, it might seem that T^R provides a more intrinsic description than the original theory T.

Nonetheless, the intuitive appeal of Ramsey sentences fades quickly in the light of critical scrutiny. Most famously, already in 1928, Newman argued that Bertrand Russell's structuralism trivializes, for these structural claims are true whenever their observational consequences are true (see Newman, 1928). The so-called Newman objection to structural realism has been the centerpiece of recent debates about Ramsey-sentence structuralism. But even before we get to that level of scrutiny, there is something quite strange in the idea of passing to the Ramsified theory T^R to get rid of redundancy. Let's recall that a formal theory T doesn't actually refer to things like electrons or protons – it's formulated in an uninterpreted calculus. Hence, T doesn't actually have any referring terms.

It seems that the impulse to Ramsify is no other than the original impulse to use uninterpreted mathematical symbols to represent physical reality. You'll recall that one of the key maneuvers in the development of non-Euclidean geometries was de-interpreting words like "line," thereby liberating mathematicians to focus attention solely on the relation that "line" plays relative to other (uninterpreted) terms in their formal calculus.

In any case, what's really at stake here is the question of what attitude we should take toward the best scientific theories of our day and age. At one extreme, radical scientific realists assert that we should give nothing less than *full* assent to these theories, interpreted literally. To draw an analogy (that scientific realists will surely eschew), the extreme scientific realist is akin to the radical religious fundamentalist, and in particular to those fundamentalists who say that one must interpret scriptures literally. The point of that injunction, we all know, is to enable religious leaders to foist their opinions on others. At the opposite extreme, an extreme scientific antirealist sees science as having no epistemic authority whatsoever – i.e., a successful scientific theory doesn't call for any more epistemic attention on our part than, say, Zoroastrianism.

In the light of this somewhat hyperbolic characterization of the anti/realism debate, we can see various alternative positions as granting a selective epistemic authority to successful scientific theories. Consider an analogy: suppose that you know a highly skilled car mechanic, Jacob. You completely trust Jacob when it comes to his opinions on automobile-related issues. For example, if he says that you need a new alternator, then you won't doubt him, even if it costs you a lot of money. Nonetheless, if Jacob tells you that you need a new kidney, or that you should vote for a certain candidate,

you might well ignore his opinion – since he's speaking on a topic that lies outside his proper expertise.

Now, selective scientific realists consider successful scientific theories to be epistemically authoritative, but only when they speak on topics within their expertise. The different brands of selective realism are distinguished by how they understand the expertise of science. For example, a constructive empiricist (such as van Fraassen) trusts a successful scientific theory T when it makes predictions about empirical phenomena (presupposing, as he does, that it makes sense to speak of predictions and empirical phenomena – precisely the point to which Boyd and Putnam object). Similarly, a structural realist (such as Worrall) trusts a successful scientific theory T on its structural pronouncements. But if T says something about things in themselves (or whatever is *not* structure), then the structural realist treats it as no more of an authority than your auto mechanic is on politics.

The previous considerations suggest that varieties of selective scientific realism can be classified by means of different notions of theoretical equivalence. For example, the strict empiricist thinks that the important part of a theory is its empirical content; and hence, if two theories T_1 and T_2 agree on empirical content, then there is no epistemically relevant difference between them. Similarly, a structural realist thinks that the important part of a theory is its pronouncements about structure; and hence, if two theories T_1 and T_2 agree on structure, then there is no epistemically relevant difference between them. In the particular case of Ramsey-sentence structuralism, the structural pronouncements of a theory T_i are captured by its Ramsey sentence T_i^R. Hence, if $T_1^R \equiv T_2^R$, then there is no epistemically relevant difference between T_1 and T_2.

Unfortunately, the statement "$T_1^R \equiv T_2^R$" doesn't have an obvious meaning, since there is no single, obviously correct notion of second-order logical consequence. What this means is that we get different notions of "same structure" depending on which notion of second-order consequence we adopt. Let's review, then, some salient notions of second-order logical consequence.

Second-order logic is a complicated subject in its own right, and has been the source of much dispute among analytic philosophers. We refer the reader to studies such as Shapiro (1991) and Bueno (2010) for more details. For present purposes, it will suffice to make some minor modifications of first-order logical grammar: first, we add a list of second-order variables $X, Y, \ldots$ Each second-order variable has a specific arity $n \in \mathbb{N}$, which means that it can stand in the place of an n-ary relation symbol. We then permit formulas such as $X(x_1, \ldots, x_n)$, with a second-order variable of arity n applied to n first-order variables. We also add an existential quantifier $\exists X$ that can be applied to quantify over second-order variables.

Now there are two important facts to keep in mind about second-order logic. The first fact to keep in mind is that second-order logic has is intrinsically incomplete – hence there is no tractable syntactic relation "$\vdash$" of second-order provability. The second fact to keep in mind is that there are several candidates for the semantic relation "$\models$" of entailment. Depending on which choice we make for this relation, we will get a different notion of logical equivalence.

DEFINITION 8.1.2 A second-order Σ-frame $\mathscr{F} = (M, (\mathscr{E})_{n\in\mathbb{N}})$ consists of a first order Σ-structure M and, for each $n \in \mathbb{N}$, a subset $\mathscr{E}_n$ of $\mathscr{P}(M^n)$. We let $\mathscr{E}^{\mathscr{F}} = \bigcup_{n\in\mathbb{N}} \mathscr{E}_n$. Here the sets in $\mathscr{E}^{\mathscr{F}}$ will give the domain of the second-order quantifiers in frame $\mathscr{F}$.

In order to define the relation $\vDash$, we will also make use of the notion of a variable assignment. Given a Σ-frame $\mathscr{F}$, a first-order variable assignment g assigns each variable x to an element $g(x) \in M$. A second-order variable assignment G assigns each variable X of arity n an element $G(X) \in \mathscr{E}_n$. We then define

$M[G, g] \vDash \exists X\phi$ iff for some $E \in \mathscr{E}_n$, $M[G^E_X, g] \vDash \phi$, where G^E_X is the second-order variable assignment that agrees with G on everything besides X, which it assigns to E.

Now to define the relation $\vDash$ between sentences, we have to decide which second-order Σ-frames to quantify over. We get three different notions, depending on the family we choose:

1. For **full semantics**, we permit only those Σ-frames in which $\mathscr{E}_n = \mathscr{P}(M^n)$.
2. For **Henkin semantics**, we permit all Σ-frames in which $\mathscr{E}_n$ is closed under first-order definability.
3. For **frame semantics**, we permit all Σ-frames.

Recall that the more structures there are, the more counterexamples and, hence, the fewer implications. Accordingly, full semantics has more entailments than Henkin semantics, and Henkin semantics has more entailments than frame semantics. Hence, full semantics yields a more liberal notion of equivalence than Henkin semantics, which yields a more liberal notion of equivalence than frame semantics.

In the following discussion, we will take T_i, for $i = 1, 2$, as a theory in signature $\Sigma \cup \Sigma_i$, where Σ_i is disjoint from Σ. We let T_i^* be the result of replacing terms in Σ_i with (possibly second-order) variables, and we let T_i^R be the corresponding Ramsey sentence of T_i. We now give a general schema for Ramsey equivalence of theories.

DEFINITION 8.1.3 Two theories T_1 and T_2 are **Ramsey equivalent** if T_1^R is logically equivalent to T_2^R.

The three choices of frames discussed earlier give rise to three notions of Ramsey equivalence.

- RE_1 = loose Ramsey equivalence = Ramsey sentences are equivalent relative to full semantics.
- RE_2 = moderate Ramsey equivalence = Ramsey sentences are equivalent relative to Henkin semantics.
- RE_3 = strict Ramsey equivalence = Ramsey sentences are equivalent relative to frame semantics.

Obviously, then, we have $\mathrm{RE}_3 \Rightarrow \mathrm{RE}_2 \Rightarrow \mathrm{RE}_1$.

We can now give a sharpened formulation of the Newman problem – in the spirit of Ketland (2004) and Dewar (2019). Recall that on the old-fashioned syntactic view of theories, two theories T_1 and T_2 are considered to be empirically equivalent if they have the same consequences in the observation language. If we now think of Σ as the

observation vocabulary, then we could formulate this criterion as saying that $\mathrm{Cn}(T_1)|_\Sigma = \mathrm{Cn}(T_2)|_\Sigma$, where $\mathrm{Cn}(T_i)|_\Sigma$ indicates the restriction of the set of consequences to those that contain only observation terms.

One might also wish to formulate a more semantically oriented notion of empirical equivalence. For example, we might say that two theories T_1 and T_2 are empirically equivalent if their models agree on Σ-structure.

DEFINITION 8.1.4 We say that T_1 and T_2 are Σ-equivalent just in case, for each model M of T_1, there is a model N of T_2 and an isomorphism $h : M|_\Sigma \to N|_\Sigma$, and vice versa.

The following result shows that this semantic notion of empirical equivalence implies the syntactic notion.

PROPOSITION 8.1.5 *If T_1 and T_2 are Σ-equivalent, then* $\mathrm{Cn}(T_1)|_\Sigma = \mathrm{Cn}(T_2)|_\Sigma$.

Proof Suppose that T_1 and T_2 are Σ-equivalent. Let ϕ be a Σ-sentence such that $\phi \notin \mathrm{Cn}(T_2)$. By completeness, there is a model M of T_2 such that $M \nVdash \phi$. Since T_1 and T_2 are Σ-equivalent, there is a model N of T_1 and an isomorphism $h : M|_\Sigma \to N|_\Sigma$. But then $N \nVdash \phi$, hence $\phi \notin \mathrm{Cn}(T_1)$. It follows that $\mathrm{Cn}(T_1)|_\Sigma \subseteq \mathrm{Cn}(T_2)|_\Sigma$. The result follows by symmetry. □

However, this implication cannot be reversed – i.e., the syntactic notion of empirical equivalence doesn't imply the semantic notion.

Example 8.1.6 Let Σ be the empty signature (with equality). Let $\Sigma_1 = \{c_r \mid r \in \mathbb{R}\}$, and let T_1 be the theory in $\Sigma \cup \Sigma_1$ with axioms $c_r \neq c_s$, for all $r \neq s$. Let T_2 be the theory in Σ that says there are infinitely many things. Then $\mathrm{Cn}(T_1)|_\Sigma = \mathrm{Cn}(T_2)|_\Sigma$. However, T_2 has a countable model M, and T_1 has no countable model. Therefore, T_1 and T_2 are not Σ-equivalent. ⌟

The Newman problem for structural realism is usually phrased as saying that it's too easy for a theory's Ramsey sentence to be true – that the Ramsey sentence is "trivially realizable." We can make precise what is meant here by "too easy" in terms of the notion of theoretical equivalence. In short, Ramsey equivalence – i.e., having logically equivalent Ramsey sentences – is too liberal a notion of equivalence. In particular, empirically equivalent theories are Ramsey equivalent.

PROPOSITION 8.1.7 (Dewar) *If T_1 and T_2 are Σ-equivalent, then T_1^R and T_2^R are logically equivalent relative to full semantics.*

Proof Suppose that T_1 and T_2 are Σ-equivalent. Now let $\mathscr{F}$ be a full Σ-frame such that $\mathscr{F} \vDash T_1^R$. Thus, there is a second-order variable assignment G such that $\mathscr{F}[G] \vDash T_1^*$. Let M be the $\Sigma \cup \Sigma_1$ structure obtained by assigning $M(R) = G(X)$, where X is the variable in T_1^* that replaces R in T_1. Clearly M is a model of T_1. Since T_1 and T_2 are Σ-equivalent, M is Σ-isomorphic to a model N of T_2. This model N of T_2 defines a second-order variable assignment G' such that $\mathscr{F}[G'] \vDash T_2^*$, and hence $\mathscr{F} \vDash T_2^R$. □

The notion of empirical equivalence imposes no constraints whatsoever on what the theories T_1 and T_2 say in their theoretical vocabulary – and for this reason, nobody but the most extreme empiricist should adopt weak Ramsey equivalence as their standard.

Moving back toward the right-wing side of the spectrum of theoretical equivalence, one might hope that moderate Ramsey equivalence would provide a more reasonable standard. But the following result shows that any two mutually interpretable theories satisfy RE_2.

PROPOSITION 8.1.8 (Dewar) *If T_1 and T_2 are Σ-equivalent and mutually interpretable, then T_1^R and T_2^R are logically equivalent relative to Henkin semantics.*

Proof Suppose that T_i is a theory in $\Sigma \cup \Sigma_i$. We will show that if $F : T_1 \to T_2$ is a translation (which is the identity on Σ), then $T_2^R \vDash T_1^R$, where the $\vDash$ symbol is entailment relative to Henkin semantics, and T_i^R is the result of Ramsefying out Σ_i. Suppose then that $F : T_1 \to T_2$ is a translation and that $\mathscr{H}$ is a Henkin structure (of signature Σ) such that $\mathscr{H} \vDash T_2^R$. Thus, $\mathscr{H}[G] \vDash T_2^*$ relative to some second-order variable assignment G. Consider then the first-order structure M for signature $\Sigma \cup \Sigma_1$ that agrees with $\mathscr{H}$ on Σ, and such that $M(P) = G(X_P)$, for each $P \in \Sigma_2$, where X_P is the second-order variable that replaces P in T_1^*. It is clear then that $M \vDash T_2$. Now we will use the fact that the translation $F : T_1 \to T_2$ gives rise to a functor $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ (6.6.5). In particular, $(F^*M)(Q) = M(F(Q))$ for each relation symbol $Q \in \Sigma \cup \Sigma_1$. Now define a second-order variable assignment G' by setting

$$G'(X_Q) = (F^*M)(Q) = M(F(Q)),$$

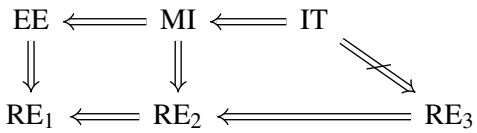
for each variable X_Q that occurs in T_1^*. (Again, we use X_Q to denote the variable that replaces a relation symbol Q that occurs in T_1.) To see that G' is a Henkin-admissible assignment, note that $F(Q)$ is a Σ_2-formula, and so $M(F(Q))$ is a first-order definable subset of M. By construction, each first-order definable subset of M is an element of $\mathscr{E}^{\mathscr{H}}$. Now, it's clear that $\mathscr{H}[G'] \vDash T_1^*$, and hence that $\mathscr{H} \vDash T_1^R$. Since $\mathscr{H}$ was an arbitrary Henkin frame, it follows that $T_2^R \vDash T_1^R$. By symmetry, if there is a translation $G : T_2 \to T_1$, then $T_1^R \vDash T_2^R$. Therefore, if T_1 and T_2 are mutually interpretable, then T_1^R and T_2^R are Henkin equivalent. □

There is one last hope for the Ramsefier: that strict Ramsey equivalence (RE_3) will provide the right notion of structural equivalence. Unfortunately, RE_3 proves to be the worst candidate for structuralism, since intertranslatable theories need not satisfy RE_3.

Example 8.1.9 Let $\Sigma_1 = \{r\}$, and let $\Sigma_2 = \{r'\}$, where both r and r' are unary predicates. Let $T_1 = \{\exists x r(x)\}$, and let $T_2 = \{\exists x \neg r'(x)\}$. The reconstrual $F(r) = \neg r'(x)$ induces a homotopy equivalence between T_1 and T_2 – i.e., T_1 and T_2 are intertranslatable. However, the Ramsey sentences of T_1 and T_2 are not frame equivalent. In particular, consider any frame $\mathscr{F}$ with first-order domain M, and $\mathscr{E}_1^{\mathscr{F}} = \{M\}$ – i.e., M is the only admissible subset of M. Then $\mathscr{F} \vDash T_1^R$ but $\mathscr{F} \nvDash T_2^R$. ⌟

Since strict Ramsey equivalence (RE_3) is more conservative ("right wing") than definitional equivalence, we don't expect structural realists to find it congenial. But what about those hard-core realists – like David Lewis or Ted Sider – who pin their theoretical hopes on natural properties and reference magnetism? Might they actually want a criterion of equivalence that is even more conservative than definitional equivalence? In fact, it seems that frame semantics might be a good way to capture the idea that to describe a possible world, you need to say not only what things exist, but also what the natural properties are. We should note, however, that adopting a first-order signature Σ already goes some way to picking out natural properties. When we specify a Σ-structure M, we get a natural property $M(\phi)$ for each formula ϕ of Σ. It's not clear then why a theorist who has adopted a first-order signature Σ would need to additionally specify a notion of natural properties.

The previous results can be summarized in the following diagram:



Here "EE" is empirical equivalence (explicated semantically), "MI" is mutual interpretability over Σ, and "IT" is intertranslatability over Σ, which is equivalent to definitional equivalence. It appears then that none of the notions of Ramsey equivalence gets us near the promising area in the neighborhood of intertranslatability. Most philosophers, we think, would agree that intertranslatability is a reasonable – if somewhat strict – explication of the idea that two theories have the same logical structure.

8.2 Counting Possibilities

If you page through an analytic philosophy journal, it won't be long before you see the phrase "possible world." Many philosophical discussions focus on this concept, and it is frequently used as a basis from which to explicate other concepts – Humean supervenience, counterfactuals, laws of nature, determinism, physicalism, content, knowledge, etc. When the logically cautious philosopher encounters this concept, she will want to know what rules govern its use. Where things get really tricky is when philosophers start invoking facts about the structure of the space of possible worlds – e.g., how many worlds there are, which worlds are similar, and which worlds are identical. These sorts of assumptions play a significant role in discussions of fundamental ontology. To take a paradigm example, Baker (2010) argues that if two models are isomorphic, then they represent the same possible world.

Analytic philosophers might be the primary users of the phrase "possible world," but they aren't the only ones using the concept. Scientists talk about possible worlds all the time. However, at least in the exact sciences, there are explicit rules governing the use of possible-worlds talk. Indeed, these rules are built into the structure of their theories and, more particularly, in the structure of those theories' spaces of models. Following Belot

(2017), we think that philosophers ought to try to understand the way that scientists' theories guide their use of modal concepts.

Nonetheless, it's not hard to find philosophers scratching their heads and asking themselves questions like the following:

($\star$) Consider two general relativistic spacetimes, M and N, and suppose that $h : M \to N$ is an isomorphism (e.g., a metric preserving diffeomorphism). Do M and N represent the same possible world?

($*$) Consider two Newtonian spacetimes, M and N, and suppose that $h : M \to N$ is an isomorphism (e.g., a shift). Do M and N represent the same possible world?

Belot (2017) helpfully classifies philosophers into two groups according to how they answer these questions: the *shiftless* claim that isomorphisms do not generate new possibilities, and the *shifty* claim that isomorphisms do generate new possibilities. In particular, the shiftless philosopher says that if $h : M \to N$ is an isomorphism, then M and N represent the same possibility. In contrast, the shifty philosopher allows that M and N might represent different possibilities, even though they are isomorphic. While the majority of philosophers of physics and metaphysicians have become shiftless, Belot champions the heterodox, shifty point of view. As we will now argue, all parties to the dispute have adopted a questionable presupposition, viz. that it makes sense to count possibilia.

But first, what hangs on this dispute between the shifty and the shiftless? In the first place, shiftless philosophers believe that they are on the right side of history, ontologically speaking. In particular, they believe that it would be wrong to countenance the existence of two possibilities, represented by M and N, when a single one will do the job. This way of thinking trades on vague associations with Leibniz's principle of the identity of indiscernibles: since M and N are indiscernible, there is no reason to regard them as different. Belot points out, however, that shiftless philosophers have trouble making sense of how theories can guide the use of modal concepts. In particular, he argues that the shiftless view is in danger of collapsing the distinction between deterministic and indeterministic theories.

One is tempted immediately to dismiss the shiftless position, because it patently conflicts with the standard reading of physical theories. Take, for example, a Galilean spacetime M, and let $\gamma : \mathbb{R} \to M$ be an inertial world line in M. Now, a boost $x \mapsto x + vt$ for some fixed $v > 0$ is represented by an isomorphism $h : M \to M$. Does this boost generate a new possibility? The question might seem confusing because the model on the right side of $h : M \to M$ is the same as the model on the left side. It might seem to be trivially true, then, that $h : M \to M$ does *not* generate a new possibility. But let's see what happens if we adopt the shiftless view. If h does *not* generate a new possibility, then we ought to say, of a particle in inertial motion that it could *not* be in some other state of inertial motion (because there is no other such state of inertial motion). But that claim is contrary to the way that physicists use this theory to guide their modal reasoning. When a physicist adopts Galilean relativity, she commits to the claim that there are many distinct possible states of inertial motion, and that a thing that is in one state of inertial motion *could be* in some other state of inertial motion. In other

words, it matters to physicists that the isomorphism $h : M \to M$ is not the identity isomorphism and, in particular, that the world line $h \circ \gamma$ is not the same as the world line γ. Nonetheless, shiftless philosophers can't make sense of these modal claims, because they insist that isomorphisms don't generate new possibilities.

Despite the implausibility of the shiftless view, there are some very serious and smart philosophers who defend it. What is it, then, that really drives their insistence on saying that isomorphism (at the level of representations) implies identity (at the level of the represented)? We suspect that the shiftless are fumbling their way toward an insight – but an insight that is difficult to articulate when one is operating with mistaken views about mathematical objects and, in particular, about the relation between abstract and concrete objects. We blame a lot of this confusion on Quine, who decided that we have no need for the abstract–concrete distinction – in particular, that belief in the existence of abstracta is no different in principle from belief in the existence of concreta.

At risk of oversimplifying, we will first give a simple formulation of the basic insight toward which we think the shiftless philosophers are fumbling:

(†) A theory T is indifferent to the question of the identity of its models. In other words, if M and N are models of T, then T neither says that $M = N$ nor that $M \neq N$. The only question T understands is: are these models isomorphic or not?

Now, please don't get us wrong: (†) does not say that isomorphic models are identical, nor does it say that the theory T treats isomorphic models as if they were identical. No, from the point of view of T, the question, "are they identical?" simply does not make sense. According to this thesis, claims of identity, or nonidentity of models, play *no* explanatory role in the theory.

We realize that this thesis is controversial and that it might take some time for philosophers to become comfortable with it. The problem is that we learned a little bit of set theory in our young years, and we seem to assume that everything lives in a world of sets – where questions of the form "is M equal to N" always have a definite answer. Indeed, the rigid grip of set theory makes philosophers profoundly uncomfortable with contemporary mathematics, which likes to play a fast and loose game with identity conditions. Consider a simple example (due to John Burgess): suppose that we ask two different mathematicians two different questions:

(Q1) How many groups are there with two elements?

(Q2) Inside the group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, how many subgroups are there with two elements?

What we are likely to find is that mathematicians will give apparently conflicting answers. On the one hand, they will tell us that there is only one group with two elements. On the other hand, they will tell us that $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ has two distinct subgroups with two elements. Obviously, if taken literally, these two answers contradict each other. But there is no genuine conflict, and mathematicians are not in crisis about the number of groups with two elements. No, the fact is mathematicians use words and symbols in a different way than we use them in everyday life – e.g., when we count the number of apples in a basket.

To reinforce this point, recall that categorical equivalence doesn't respect the number of objects in a category. Consider, for example, the following two categories: let **C** be the category with one object and one identity morphism. Let **D** be the category with two objects a, b, one identity morphism from each object to itself, and a pair of morphisms $f : a \to b$ and $g : b \to a$ that are inverse to each other. Then **C** and **D** are equivalent categories – which entails that "this category" doesn't really have a definite number of objects. It is not correct to say that it has one object, and it's not correct to say that it has two. Or, perhaps better: it is just as correct to say that it has one object as it is to say that it has two.

Here, then, is our positive proposal:

For the purposes of interpreting a theory T, the collection $\mathrm{Mod}(T)$ of its set-theoretic models should be treated as nothing more nor less than a *category*. In particular, the philosopher of science shouldn't say things about $\mathrm{Mod}(T)$ that are not invariant under categorical equivalence, nor should they argue over questions – such as "how many models does T have?" – whose answer is not invariant under categorical equivalence.

If this proposal is adopted, then there is no debate to be had between the shifty and the shiftless. The question they are asking – do isomorphisms generate new possibilities? – depends on a notion (the number of isomorphic possibilities) that is not invariant under categorical equivalence.

The rationale for this proposal is our belief that models of a theory T in **Sets** are *representations* of that theory; the set-theoretic description of these models is *not* itself a further theory that attempts to describe the world at an even finer-grained level of detail than was done by T. We can further clarify these points by means of a simple example.

Example 8.2.1 Suppose that Berit is a scientist with a very simple theory. Her language Σ has a single predicate symbol P, and her theory T says that there are exactly two things, one of which is a P:

$$\exists x \exists y (P(x) \wedge \neg P(y) \wedge \forall z ((z = x) \vee (z = y))).$$

Now we metatheorists know that a set-theoretic model M of T consists of a two-element set, say $X = \{a, b\}$, with a singleton set $M(P)$. Let M be the model such that $M(P) = \{a\}$, and let N be the model such that $N(P) = \{b\}$. Then the permutation $h(a) = b, h(b) = a$ gives a Σ-isomorphism $h : M \to N$. (But the permutation h is not an automorphism of M.)

Let's consider the shifty–shiftless dilemma with regard to the models M and N, with the isomorphism $h : M \to N$. The shifty philosopher (e.g., Belot) says that M and N represent distinct possibilities. The shiftless philosopher (e.g., Baker) says that M and N represent the same possibility. Who is on the side of truth?

In our opinion, both the shifty and the shiftless say misleading things about this example. On the one hand, the shifty claim is misleading, because the user of T doesn't have the language to say what would be different between M and N. She cannot say, "in

M, a is P, and in N, a is not P," because she herself doesn't have the name "a." The shiftless wants us to start counting how many models there are, but the theory T doesn't answer that question.

On the other hand, the shiftless would insist that there is only one possibility, represented redundantly by M and N. But that claim is misleading for the following reason. Berit's theory T is an extension of the theory T_0, in empty signature, that says there are exactly two things. Let $I : T_0 \to T$ be the translation of T_0 into T, and let $I^* : \mathrm{Mod}(T) \to \mathrm{Mod}(T_0)$ be the functor that forgets the assignment of P. Here I^*M and I^*M are both the bare two-point set X, and the isomorphism $I^*h = h : X \to X$ is the nontrivial permutation. Recall, though, that functors map identity morphisms to identity morphisms. Hence, if the isomorphism $h : M \to N$ is considered to be an identity (as the shiftless seem to do), then it would follow that I^*h is the identity morphism. Thus, contra the shiftless, we cannot identify M and N and forget that there was a nonidentity isomorphism $h : M \to N$. If we do that, then we won't be able to see how the theory T is related to the theory T_0.

The confusion here is somewhat similar to Skølem's paradox (about the existence of uncountable sets in models of ZF set theory), where we run into trouble if we don't distinguish between claims made in the object language and claims made in the metalanguage. In the present case, one might be tempted to think of the theory T as saying things such as

In model M, a is a P.

Of course, T says no such thing, since it doesn't have names for models or for elements in models.

The other problem here is in the way that we've set up the problem – by speaking as if the representation relation holds between M (or N) and the world. To the contrary, the representation relation holds between Berit's language and the world, and we (the metatheorists) are representing Berit's theorizing using our own little toy theory (which presumably includes some fragment of set theory, because that's a convenient way to talk about collections of formulas, etc.). Berit herself doesn't claim that M (or N) represents the world – rather, the metatheorist claims that M and N represent ways that Berit's language could represent the world. Accordingly, Berit doesn't claim that $M = N$, or that $M \neq N$; those are metatheoretical assertions – and do not add to the stock of knowledge about the world. ⌟

Before proceeding, we should deal with an obvious objection to the view we've put forward. Some philosophers will point out that it is simply false to say that physicists don't count the number of possibilities. Indeed, it's precisely by counting the number of possibilities that physicists derive notions such as entropy.

We do not disagree with this point, but it doesn't conflict at all with our positive proposal (to talk about models of a theory as a category). Category theory is a framework that is almost infinitely flexible: what we can talk about in a categorically invariant way depends on how we – or physicists – define the relevant category. For the case at hand, if X is a classical phase space, then it is assumed that X is a discrete category – i.e., that

there are no nontrivial isomorphisms between elements of X. Thus, in this case, there is no question about whether to count two isomorphic possibilities as the same, because we (or better, the physicists) have chosen not to admit isomorphic possibilities.

To be clear, we explicitly reject the idea that there is a single relation "being isomorphic" that either holds or does not hold between concrete objects. On the contrary, the notion of isomorphism applies to abstractions, and different notions of isomorphism are valid for different levels of abstraction. It's up to us to decide which level of abstraction serves our purposes in reasoning about concrete, physical reality. (In particular, models of a theory are not concrete realities, and that's why they cannot either be identical or nonidentical.)

For all of its other virtues, one of the defects of the semantic view of theories is that it obscures the object language-metalanguage distinction, a distinction that is absolutely necessary to make sense of the notion of symmetry of representations. To be more accurate, the targets of this criticism are advocates of the "language-free" or "semantic-L" view (see Halvorson, 2013). The picture we get from the language-free semantic view is that mathematical structures are out there in the world, and that they are either isomorphic to each other or they are not. Of course, that picture completely ignores the fact that isomorphisms are defined in terms of language or, to put it more accurately, that isomorphisms relate mappings $M : \Sigma \to \mathbf{Sets}$ and $N : \Sigma \to \mathbf{Sets}$, which have a common domain Σ. Thus, in particular, arbitrary mathematical structures are neither isomorphic nor non-isomorphic.

The object language Σ serves as the reference point in defining a notion of symmetry. The object language tells us what must be held fixed, and the metalanguage tells us what can be varied. In particular, a model M of a theory T can have a nontrivial automorphism group because of two features of the formal setup:

1. The metalanguage describes the world in finer-grained language than the object language.
2. Distinctions that are not made by the object language are not significant for the kinds of explanations that the theory T gives.

If we drop either one of those components, then we will most likely make a hash of the notion of symmetry. Without the metalanguage, there is no way to see any difference between a and b, and so no way to express the change the occurs in the permutation $a \mapsto b$. But if we think of the metalanguage as a better object language, then we shouldn't count $a \mapsto b$ as a symmetry, since these two things are distinguishable in the metalanguage. Thus, it's precisely the mismatch between object language and metalanguage that provides us with a rich notion of symmetry; and, conversely, the importance of the notion of symmetry gives us reason to maintain a distinction between object language and metalanguage.

The distinction between object language and metalanguage is one of the most interesting ideas in twentieth century logic and philosophy – and it remains one of the least well understood. Obviously, Carnap made a lot of this distinction, and, in fact, he seems to use it as his primary analogy in formulating the distinction between internal and external questions and, more generally, in understanding the relationship between theories in

the exact sciences and our other, nonscientific beliefs and attitudes. In contrast, Quine seems to reject the idea that there is an important difference of status between object and metalanguage. He seems to propose, instead, that the ascent to metalanguage should be seen as an extension of one's object language – and so assertions in the metalanguage have exactly the same force as assertions in the object language.

8.3 Putnam's Paradox

Perhaps the most notorious argument from logical metatheory to philosophy is Hilary Putnam's model-theoretic argument against realism (Putnam, 1977, 1980). Here is how the argument goes.

Suppose that theory T is consistent, i.e., T does not imply $\bot$, or equivalently, T has a model. Now let W represent the collection of all actually existing objects, i.e., W represents "the world." Besides consistency, we will make two other minimal mathematical assumptions about T: First, the cardinality of the language is not so large as to force belief in the existence of too many objects. In short, we require that $|\Sigma| \leq |W|$. Second, the theory T doesn't entail that there are at most n things, for $n \in \mathbb{N}$.

We then proceed as follows: by the Löwenheim–Skølem theorem, there is a model M of T such that $|M| = |W|$. This means, of course, that there is a bijection $f : M \to W$. Now we define another model of T, still called W, by setting $W(p) = f(M(p))$ for each relation symbol p in the theory T. But then the the world is a model of T. That is, T is true.

This argument is intended to show that if T is consistent, then T is true – actually true, in the real world. There is one obvious way to try to block this argument, and that's to say that the model W may not be the "intended" assignment of relation, function, and constant symbols to things in the real world. However, Putnam tries to block that response essentially by calling upon your charity. Imagine that T is the theory held by some other person, and that you're going to try your best to believe that what that person says is true. In other words, you are going to give her the benefit of the doubt whenever possible. Then what Putnam has shown is essentially that there is a way of giving her the benefit of the doubt.

This simple-looking argument is so subtle, and there are many ways we might respond to it. But let me be completely clear about my view of this argument: it is absurd. This version of Putnam's argument is not merely an argument for antirealism, or internal realism, or something like that. This version of the argument would prove that all consistent theories should be treated as equivalent: there is no reason to choose one over the other. Thus, Putnam's paradox is essentially an argument for one of the most radically liberal views of theoretical equivalence imaginable. The only more radical view is the Zenonian view, according to which all theories are equivalent.

To keep things concrete, let's suppose that T is Mette's theory. The goal of Putnam's argument is to show that Mette's theory is true. In my view, the problematic assumption in the argument is the following:

(S) The world can be described as an object W in the universe of **Sets**.

The question to be raised here is: *who* is using the theory of sets to describe the world? Putnam's presentation makes it seem that either: (1) it's unproblematic and theory-neutral to describe the world as a set, or (2) a realist must describe the world as a set. We don't agree with either claim.

Let's remember that nobody here – including Putnam – is free from language and theory. When Putnam describes the world as a set, it might seem that he is making minimal assumptions about it. But the opposite is true. When you have a set, you have all of its subsets; and when you have two sets, you have all of the functions between them. To even say these things, we need the rich and expressive language of set theory.

Thus, Putnam has set things up in a misleading way by (1) describing the world as a set but (2) failing to note who is responsible for this description of the world as a set. Suddenly it becomes clear why Putnam's argument goes through, and why it's trivial. Putnam assumes that Mette's theory T is set-theoretically consistent, which simply means that Mette's theory can be translated into the background theory T_0 that was used to describe the world. That is, there is a translation $F : T \to T_0$. Putnam rightly concludes that the T_0-theorist could take Mette's theory T to be true. What Putnam does not show is that *anybody*, regardless of their background theory, could take Mette's theory to be true.

Putnam's argument should actually not make any assumption about W – i.e., it should be like a black box. However, Putnam begins by assuming that there is already a fixed interpretation of ZF into the world – i.e., we know what objects are, and collections of objects, and functions between objects, etc. He then asks whether T has one (or perhaps even many) interpretations into this already understood domain. And of course, the answer is yes.

Thus, Putnam assumes that he is permitted a trans-theoretical language to speak of the domains M and W. By "trans-theoretical" here, I mean simply that the language of T_0 (in this case, ZF) is not the same as the language of the theory T. In particular, for Putnam's argument to go through, he needs to be able to make distinctions in W that simply cannot be made by users of the theory T.

To make these ideas more concrete, let's consider an example: Let $\Sigma = \{c, d\}$, where c and d are constant symbols. Let T be the theory in Σ that says $c \neq d$, and $\forall x((x = c) \vee (x = d))$. (This example violates the strictures of Putnam's Löwenheim–Skølem based argument, but the point will not depend on those details.) Of course, there is only one model of T up to isomorphism. And yet, a skeptical worry arises! Imagine two people, Mette and Niels, both of whom accept T, and both of whom think that the world is the set $\{a, b\}$. And yet, Mette says that c denotes a, whereas Niels says that c denotes b. Do Mette and Niels disagree? The answer is yes and no.

We have already misdescribed the situation. Mette cannot say that "c denotes a," because a is not a name in her language. Similarly, Niels cannot say that "c denotes b." It is the metatheorist who can say: "Mette uses c to denote a," and "Niels uses d to denote b." But how does the metatheorist's language get a grip on the world? How can he tell what Mette and Niels are denoting, and that they are different things? Now, Putnam might claim that it is not he, but the realist, who thinks that the world is made of things, and that when our language use is successful, our names denote these things.

So far I agree. The realist does think that. But the realist can freely admit that even he has just another theory, and that his theory cannot be used to detect differences in how other people's theories connect up with the world. All of us – Mette, Niels, Hilary, you, and I – are on the same level when it comes to language use. None of us has the metalinguistic point of view that would permit us to see a mismatch between language and world.

Now, I suspect that some people might think that I've simply affirmed Putnam's conclusion – i.e., that I have embraced internal realism. I can neither affirm nor deny that claim (largely because of unclarity in the meaning of "internal realism"). But I insist that *if* Putnam's argument works, then we have no reason to discriminate between (ideal) consistent theories, and we should adopt an absolutely radical left-wing account of theoretical equivalence. I, for one, am loath to think that good theories are so easy to find.

Consider another scenario, where now I, rather than Putnam, get to choose the rules of the game. In other words, I have my own theory T_0 of which I believe the world W is a model. Then along comes Putnam and says that any consistent theory can be interpreted into the world W. But if my background theory T_0 is not ZF, then I don't see W as a set, and Putnam's argument cannot even get started. In particular, I don't necessarily grant that there is an isomorphism $f : M \to W$ between a model M of T and this model W of my theory T_0. For one, what would I even mean by the word "isomorphism"? I, the user of the theory T_0, know about isomorphisms between models of my theory. However, M is a model of a different theory T, written in a different signature Σ, and so there may be no standard of comparison between models of T and models of T_0.

There is still another, more severe problem for Putnam's argument. For a scientific theory to be "ideal," it's really not enough for it to correctly report every actual fact. In must do more! There are a few ways to get a handle on what more a good scientific theory must do. David Lewis recognized that the "best theory" is not simply one that gets every fact correct. Instead, the "best" achieves an ideal balance of strength and simplicity. Here "strength" means reporting the facts, and simplicity means ... well, we all know it when we see it, right? Whether or not we philosophers have a good account of simplicity, the fact is that Lewis was right that there is (at least) a second component to theory evaluation, and it has something to do with systematicity, or choosing the right language, or cutting nature at the joints.

Thus, when I'm looking at a scientific theory, I'm not just interested in whether it's true. You could write down every truth in a massive encyclopedia, and I wouldn't consider it to be the best scientific theory. There are better and worse ways to say the truth. And what this means for our considerations here is that not all true theories are created equal; thus, certainly not all ideal consistent theories are equal.

We might want to go to the trouble of explaining when I, user of theory T_0, would grant that T can be interpreted into a model M of my theory. In the simplest sort of case, I would require that for each relation symbol p of T, there is a formula Fp of the appropriate arity of my language Σ_0 such that p can be interpreted as $M(Fp)$. As a user of theory T_0, I only recognize those subsets of M that can be described via the predicates of my theory. In particular, I don't necessarily have the resources to name the

elements of the domain M, and I don't necessarily have the resources to collect arbitrary elements of M and form subsets out of them. I can only talk about "things that satisfy ϕ", where ϕ is one of the predicates of my language.

So, suppose then that T is consistent relative to my theory T_0: for each model M of my theory, there is a model M^* of T with the same domain as M, and such that for each relation symbol p of Σ, $M^*p = M(Fp)$ for some formula Fp of my language Σ_0. However, even in this scenario, I wouldn't necessarily consider the theory T to be adequate, for it may fail to pick up the relationships between various models of my theory. I'd want to know that the user of T recognizes the same connections between models that I do. In particular, where I see an elementary embedding $h : M \to N$, I would require the user of T to see a corresponding elementary embedding $h^* : M^* \to N^*$ between models of his theory T. And that just means that $h \mapsto h^*$ completes the definition of a functor from $\mathrm{Mod}(T_0)$ to $\mathrm{Mod}(T)$, where the object part is given by $M \mapsto M^*$. We then have the following result.

PROPOSITION 8.3.1 *Let F be a map of Σ-formulas to Σ_0-formulas such that the map $M(\phi) \mapsto M(F\phi)$ defines a functor from $\mathrm{Mod}(T_0)$ to $\mathrm{Mod}(T)$. Then $F : T \to T_0$ is a translation.*

Proof Define a reconstrual $G : \Sigma \to \Sigma_0$ by setting $Gp = Fp$. We claim that $T_0 \vdash G\phi \leftrightarrow F\phi$ for all formulas ϕ of Σ. For this, it suffices to run through the clauses in the definition of F. For example, we need to check that $F(\phi_1 \wedge \phi_2) \equiv F(\phi_1) \wedge F(\phi_2)$, where $\equiv$ means provable equivalence modulo T_0. But this is easy to check: let M be an arbitrary model of T_0. Then

$$\begin{aligned} M(F(\phi_1 \wedge \phi_2)) &\equiv M^*(\phi_1 \wedge \phi_2) \\ &\equiv M^*(\phi_1) \cap M^*(\phi_2) \\ &\equiv M(F(\phi_1)) \cap M(F(\phi_2) \\ &\equiv M(F(\phi_1) \wedge F(\phi_2)). \end{aligned}$$

(Here I've ignored for simplicity the fact that ϕ_1 and ϕ_2 might have different free variables.) The clauses for the other connectives and for the quantifiers are similar. □

The upshot of this result for Putnam's argument is as follows: a user of a theory T_0 should only grant that T can be true if there is a translation of T into T_0. This result is not surprising at all. In real life, this is the sort of criterion we do actually employ. If I hear someone else speaking, I judge that what they are saying "could be true" if I can reconstrue what they are saying in my language. If there is no way that I can interpret their utterances into *my* language, then I am forced to regard those utterances as false or meaningless.

As Otto Neurath pointed out, and as Quine liked to repeat, we cannot start the search for knowledge from scratch. Each of us already has a theory, or theories. And we have a notion of permissible translations between theories that regulates (or describes) our attitude about which other theories could potentially be correct. If a theory T can be conservatively translated into my theory T_0, then I will think that T might possibly say something true (perhaps if its terms are charitably interpreted). But even then, I would

not necessarily judge T to be true. Indeed, if my standard of theoretical equivalence is weak intertranslatability, then I will judge T to be potentially true (even under the most charitable interpretation) only if T and T_0 are weakly intertranslatable. (And do recall that weak intertranslatability is a fairly conservative criterion of equivalence.)

What Putnam has shown, at best, is that *relative* to a background theory T_0 of bare sets, a theory T that has a model in **Sets** could be charitably interpreted as true by a user of T_0. The result is really not very interesting – except insofar as it reminds us of the dangers of uncritically accepting set theory as our background metatheory. Indeed, set theory makes nontrivial existence claims – e.g., the claim that any two points in a model are related by a permutation.

The things I've just said might sound quite similar to Lewis' (1984) response to Putnam's argument. Lewis attempts to block the argument precisely by denying the permissibility of the relevant permutation – or, what's the same, of denying that each subset of WORLD picks out a genuine property. But Lewis' response is not, by itself, sufficient to block Putnam's argument. Suppose indeed that we've identified a privileged subclass $\mathcal{N}$ of natural properties among the subsets of WORLD. We can also require, as Lewis does, that a predicate symbol p of the signature Σ must be assigned to a set $M(p) \in \mathcal{N}$. In other words, M cannot assign p to any old subset of WORLD.

What Lewis has done here, in effect, is to propose an extension of Putnam's background theory T_0, by means of adding predicate symbols to the signature Σ_0 in order to designate the subsets in $\mathcal{N}$. Let T_1 be Lewis' strengthened background theory – the theory that describes the world as a set WORLD, with a privileged family $\mathcal{N}$ of subsets of WORLD to represent the natural properties. Then Lewis' requirement that the predicates of T be interpreted as elements of $\mathcal{N}$ is tantamount to the requirement that there is an interpretation of T into Lewis' background theory T_1. Since T_1 is expressively weaker than Putnam's background theory T_0, it is more demanding to ask for an interpretation of T into T_1 than it is to ask for an interpretation of T into T_0.

Lewis' requirement can block Putnam's trivializing maneuver: for some choices of $\mathcal{N}$, there are theories T that are set-theoretically consistent but that cannot be translated into T_1. To take one trivial example, suppose that T_1 has three natural properties: the empty set, the entire world, and some proper subset of the world. Suppose also that T includes the axiom

$$\exists x\, Px \wedge \exists y\, Qy \wedge \neg\exists z(Pz \wedge Qz).$$

Then T is set-theoretically consistent, but T cannot be translated into T_1.

Nonetheless, Lewis' demands here are not strong enough. In general, for any sufficiently rich family of natural properties $\mathcal{N}$, too many theories T will be interpretable into Lewis' background theory T_1. And hence, if Lewis grants Putnam's call for charitable interpretation, then Lewis must grant that those theories are true. That concedes too much. It is easy to think of examples that would make a realist choke. For example, suppose that Gargamel has a theory that says there are many gods, and there are no electrons. If Lewis countenances just a single natural property with instances, then Gargamel's theory can be translated into Lewis' background theory – and, by the principle of charity, should be counted as true.

We will not engage now in further formal investigation of these matters, e.g., to ask how many natural properties there need to be in order for a given theory T to be interpretable into Lewis' background theory. We don't think that question is very interesting – because we've already gone off on a bad track. There are two interrelated problems here. The first problem is that Lewis' background theory T_1 has little to recommend it, even if we are inclined to accept that there are "natural properties." (And anyone who uses first-order logic implicitly does accept the existence of natural properties – they are precisely the properties that are definable in her language.) The second, and deeper, problem is that Lewis, like Putnam, seems to be supposing that all parties – or at least all metaphysical realists – can agree on some particular fixed background theory T_*. We reject that assumption, and as a result, Putnam's paradox simply dissolves.

8.4 Realism and Equivalence

According to the standard stereotype, the logical positivists were *antirealists* or *instrumentalists* about scientific theories. Moreover, this antirealist stance was facilitated by means of the syntactic analysis of scientific theories, according to which a theory T's language has some purely observational terms O, and its empirical content can be identified with $T|_O$. With this formal analysis, the positivists could then articulate their particular versions of epistemic and semantic empiricism:

- Epistemic empiricism: the reasons we have to believe T derive from reasons we have for believing $T|_O$.
- Semantic empiricism: the meaning of terms in $\Sigma \setminus O$ derives exclusively from the meaning of terms in O.

The extreme instrumentalist would say that the terms in $\Sigma \setminus O$ have no meaning: there are merely instruments to facilitate making predictions. The attenuated instrumentalist tries to find a way for terms in Σ to inherit meaning from terms in O.

In the 1960s and 1970s, the syntactic view of theories was discredited, and the tide seemed to have turned decisively against antirealism – or at least against this stereotyped antirealism. Without a clear delineation of the empirical part of a theory, it was no longer possible to think that warrant or meaning could flow upward from the observationally relevant parts of a theory.

Van Fraassen characterized the state of play in 1976: "After the demise of logical positivism, scientific realism has once more returned as a major philosophical position" (van Fraassen, 1976, 623). He goes on to characterize scientific realism as commitment to the following thesis:

> The aim of science is to give us *a literally true story of what the world is like*; and the proper form of acceptance of a theory is to believe that it is true. (van Fraassen, 1976, 623)

As is well known, van Fraassen then gave several strong arguments against scientific realism, before going on to develop his positive alternative: *constructive empiricism*.

In the years that followed, there was much back-and-forth debate: van Fraassen on the side of constructive empiricism – and dozens of other philosophers on the side of scientific realism. The terms of this debate had been set by van Fraassen, and these terms were rarely (if ever) questioned. In particular, the scientific realists seem to have been happy enough with van Fraassen's characterization of their position; their job was merely to bring out its merits.

However, if we look more closely, it becomes apparent that the debate wasn't so clear-cut. During the 1960s and 1970s, scientific realists were fond of saying that the philosophical position of scientific realism is itself a scientific hypothesis, and that the reasons for believing it are of the same nature as the reasons for believing any other scientific theory. In particular, they claimed that the hypothesis of scientific realism is the *best explanation* for the success of the scientific enterprise.

Now, van Fraassen certainly questioned the latter claim. But more interestingly, he chose not to play by the same game as the scientific realists. For van Fraassen, the reasons for being a constructive empiricist are different in kind from the reasons for accepting a scientific theory. For those who were following the debate closely, it became clear that the choice between realism and antirealism about science was not a simple disagreement about which hypothesis better explains a common domain of phenomena. There was a deeper and more elemental disagreement about the goals of philosophical reflection.

For many philosophers of the next generation, the question of scientific realism versus scientific antirealism had receded too far into the upper reaches of metaphilosophy. The simple "pro and con" arguments of the 1970s and 1980s were not going to get us anywhere, seeing that the opposed parties were using different standards to evaluate these arguments. Thus, the next generation of philosophers of science moved downward – back to the analysis of specific scientific theories. Although they may not openly use these words, I suspect that many philosophers of science now feel that "realism or antirealism?" is a pseudoquestion, or at least not a particularly interesting question.

Speaking of pseudoquestions, what makes a question pseudo? Here is one criterion: a question is pseudo if getting an answer to it wouldn't change anything you do. By that standard, it's easy to see why the realism–antirealism debate might seem like a pseudodebate. Would a scientist do anything differently tomorrow if he converted to constructive empiricism? Wouldn't he go on looking for the elegant and powerful theories, and using them to make predictions and give explanations?

This last thought suggests a better way to understand what's really at stake in the realism–antirealism debate. I suggest that the debate can be fruitfully reconceived as a battle over standards of theoretical equivalence. In particular, a realist is somebody who adopts – or recommends that people adopt – stricter standards of theoretical equivalence. Conversely, an antirealist is somebody who adopts – or recommends that people adopt – looser standards of theoretical equivalence. In short, realists are conservatives about theoretical equivalence, and antirealists are liberals about theoretical equivalence.

This construal of the realism–antirealism debate matches well with various well-known cases. Consider, for example, the case of the logical positivists. We tend to think that they were antirealists because they said, "the content of the theory T resides in its

observable part $T|_O$." But there are a lot of unclear words here, such as "content" and "residing" and "observable," and so this doesn't make for a very sharp statement of a philosophical thesis. However, one concrete implication of these positivist words is that if $T|_O = T'|_O$, then we should treat T and T' as equivalent. For example, suppose that two scientists, say Werner and Erwin, have apparently conflicting theories T and T' with the same empirical content, i.e., $T|_O = T'|_O$. Then the positivist would recommend that Werner and Erwin reconcile, for there can be no reason to prefer T over T' or vice versa. The difference between their theories is no more important than the difference between theories written in German and French. In contrast, if T says *anything* that conflicts with T', then the scientific realist thinks that one of the two must be better than the other, and that we should actively pursue inquiries to determine which it is.

This picture of the realism debate also makes sense of what structural realists were trying to achieve. In short, structural realists urge that not every single detail of a successful scientific theory should be taken with equal seriousness. In particular, they argue that if two theories T and T' differ only with respect to content, and not with respect to structure, then one can have no reason to prefer T over T', or vice versa. The normative core of structural realism, then, is to propose a notion of theoretical equivalence that lies somewhere to the left of the extreme right realist view and somewhere to the right of the extreme left views of the logical positivists, Nelson Goodman, and Putnam in the later stages of his career.

Scientists and philosophers – and, in fact, everyone – have implicit standards of equivalence that they employ to judge between truth claims, especially when those claims seem prima facie to conflict. If you believe "God doesn't exist," and your French colleague believes "*Dieu n'existe pas*," then you know that there is no dispute to be settled. Not only are those two sentences compatible with each other; they are equivalent. Even within a single language, we can say the same thing in different ways. Imagine that your friends Anne and Bent disagree about the number of roses in the vase on the counter. Anne says, "there are six roses," and Bent says, "there are a half dozen roses." In such a case, you would surely advise Anne and Bent to kiss and make up, since their dispute is merely verbal.

Those cases are easy. But there are more difficult cases in life – especially as we move into the more abstruse regions of the sciences. (And that's not even to speak of cases such as differences in matters of politics or religion.) For example, there is a debate among evolutionary biologists about the units of selection: is it the individual or the species? Many a friendship between scientists has been broken because of disagreement on issues like this one. But what if there really was no dispute between them? What if they were saying the same thing in different terms?

You might think such a scenario is unimaginable. But if the history of science can be trusted, then there have been numerous cases where prima facie disagreement has later been judged to be spurious. For example, in the mid-1920s, Werner Heisenberg developed a theory that made use of non-commutative algebra in order to predict the outcomes of measurements. This theory, called *matrix mechanics*, was hailed by many as a breakthrough, for it unified the ad hoc recipes that plagued the old quantum theory of Bohr and Sommerfeld. However, others abjured matrix mechanics, on the grounds that

it was incomprehensible and unvisualizable and entailed bizarre claims, most notably the existence of "quantum jumps." Thus, a competing theory was developed by Erwin Schrödinger, a theory based on completely different ideas and mathematical techniques. According to Schrödinger's theory, there are waves moving through physical space (or a higher-dimensional configuration space), and particles such as electrons are simply harmonic resonances in these waves.

Thus, Heisenberg presented one theory, T_1, to account for the quantum phenomena, and Schrödinger presented another theory, T_2, to account for the same phenomena. While both these were empirically adequate, the battle between Heisenberg and Schrödinger was fierce, including name-calling, a fight for prominence at professional meetings, and competition for funding and university positions. The behavior of Heisenberg and Schrödinger clearly indicated that they saw this debate as *genuine* and in need of resolution.

The conclusion of this story is typically told as follows. Based on some suggestions that Schrödinger himself made, a young mathematician, John von Neumann, formulated a conjecture: Heisenberg's matrix mechanics T_1 is equivalent to Schrödinger's wave mechanics T_2. Von Neumann then went on to prove this theorem, to the great satisfaction of most participants involved – especially those like Niels Bohr, who didn't want to choose between Heisenberg and Schrödinger. As a result, *the debate came to an end*. Since T_1 and T_2 are equivalent theories, there is no question about which one is better, at least not in any epistemically or ontologically relevant sense. There is no decision to be made about whether to accept T_1 or T_2.

Such is the nature of judgments of theoretical equivalence. When one judges that theories T_1 and T_2 are equivalent, one judges that accepting T_1 is tantamount to accepting T_2. Conversely, if one feels that T_1 might be favored over T_2, or vice versa, then one judges that these theories are *not* equivalent.

Are there equivalent theories? Setting aside the Heisenberg-Schrödinger theory as controversial, still every sane person will admit that at least some theories are equivalent. For example, say that T_1 is the theory written down in the textbook *General Relativity* by Robert Wald that is sitting on the shelf in my office, and that T_2 is the theory written down in the textbook *General Relativity* by Robert Wald that is sitting on the shelf in Carlo Rovelli's office. Of course, we all know that T_1 and T_2 are equivalent theories. In fact, most of us just say that these are the *same* theory, and that's why we use a definite description for it: "*the* general theory of relativity." But if we boil everything down to fundamental physics, then we can only say that there are two distinct collections of ink splotches, one in an office in Princeton and another in an office in Marseilles.

In my experience, philosophers tend to react to this silly sort of example by flying to the realm of abstract entities. They say something like this: the two books contain sentences that pick out the same *propositions*, and that's why we say that the sentences represent the same theory. Now, I don't disagree with this claim; I only doubt its utility. If you give me two languages I don't understand, and theories in the respective languages, then I have no way of knowing whether those theories pick out the same propositions. And that's precisely the sort of case we face with something like matrix and wave mechanics. Employing new formalism that is not yet very well understood,

it is unclear whether these theories say the same thing. Thus, we need some criterion for equivalence that is *checkable*, at least in principle. In other words, we need to know when two sentences pick out the same proposition.

There are essentially two ways to proceed from here. On the one hand, we can ask: what features must two theories have in common in order to be equivalent? In philosophical jargon: what are the necessary conditions for theoretical equivalence? This question can also be given a mathematical gloss: what are the *invariants* of theoretical equivalence? For example, some people would say that for two (single-sorted) first-order theories T_1 and T_2 to be equivalent, they must agree on the number of existing objects. There are other conditions we might try to impose, but which are a bit more difficult to cash out in terms of formal logic. For example, many contemporary philosophers would say that two equivalent theories must have the same *primitive notions* – i.e., those objects, properties, etc., that ground the other things that the theory mentions.

The second question we could ask has a more top-down flavor: could we simply define an equivalence relation on the collection **Th** of all theories? Disregarding the fact that **Th** is a proper class and not a set, there are many such equivalence relations, all of which yield some notion of theoretical equivalence. Among these untold number of equivalence relations, some have relatively simple or elegant definitions. Indeed, each one of the notions of equivalence we have canvassed in this book – e.g., definitional equivalence, Morita equivalence, and categorical equivalence – defines an equivalence relation on the class of all first-order theories.

An ideal method, I think, is to take both procedures into account. On the one hand, we need not accept a definition of equivalence if it violates necessary conditions to which we are committed. On the other hand, some of us might feel compelled to abandon an intuitive necessary condition of equivalence – i.e., some intuitive invariant of theoretical equivalence – if it conflicts with what otherwise seems the most reasonable formulation of an equivalence relation on **Th**.

We can see these sorts of choices and trade-offs being made all the time in philosophy. On the more conservative side, philosophers such as David Lewis and Ted Sider lay heavy stress on choosing the right primitives. At times it seems as if they would go so far as to say that there is a *privileged language* for metaphysics so that no theory in this language could be equivalent to a theory that is not in this language. (One wonders, however, how they individuate languages.)

One could imagine an even more conservative stance on theoretical equivalence. For example, suppose that Σ is a fixed signature (say, the preferred signature for metaphysics), and T_1 and T_2 are theories in Σ that have the same consequences (equivalently, have the same models). Should we then consider T_1 and T_2 to be equivalent? I suspect that Sider would say yes. But I also suspect that some philosophers would have said no, for they might have thought that there are preferred ways of axiomatizing a theory. Indeed, if you really believe that some facts are more basic than all the others, then shouldn't those facts be the ones enunciated in the axioms, so that all other facts are seen as flowing from them? Thus, we get an even finer-grained equivalence relation on **Th** if we demand that equivalent theories are in the same signature *and* have the same axioms.

Even that requirement – having the same axioms – is not the most conservative imaginable. We might even require that the theories literally have the same notation. For example, in formulating group theory, we could use the symbol $\circ$ for the binary relation, or we could use the symbol $\bullet$. Who knows, perhaps one of these two symbols more perspicuously represents the structure of the binary function in the world that we are trying to represent. At the farthest end of this spectrum, one could adopt a pure "Heraclitean" account of theoretical equivalence, according to which no two theories are the same. In other words, the criterion of theory identity could be made out to be *literal identity* – of symbols, axioms, etc.

Conservative views of theoretical equivalence tend to align with "realist" views about science or metaphysics. Roughly speaking, if you think that the world has real structure, then you'll think that a good theory has to represent the structure that is out there. If two theories disagree about that structure, then they cannot be equivalent. Going in the opposite direction, liberal views of theoretical equivalent tend to align with "antirealist" views about science and metaphysics. We see this tendency with Nelson Goodman in the 1960s and with Hilary Putnam in the 1970s. Putnam's move toward antirealism was augured by his giving many examples of theories that he says are equivalent, but which realists regard as being inequivalent. For example, Putnam claims that Euclidean geometry based on points is equivalent to Euclidean geometry based on lines – even though the models of these two theories can have different cardinalities.

Long before Putnam turned in this direction, the connection between antirealism and liberal views of theoretical equivalence had already been established. I'm thinking here of the logical positivists and their infamous notion of *empirical equivalence*. The idea here is that two theories T_1 and T_2 are empirically equivalent just in case they share the same observable consequences – and regardless of what else these might say. So, to take an extreme example, if T_2 is T_1 plus the sentence, "there is a new unobservable particle," then T_1 and T_2 are empirically equivalent.

Now, for the logical positivists – or at least, for some of them – empirical equivalence is equivalence enough. For they identified the content of a proposition with that proposition's empirical consequences; and it follows from this that if two propositions ϕ and ψ have the same empirical consequences, then they have the same content – i.e., they are the same proposition. Stepping back up to theories, as collections of propositions, the positivist view of content entails that two theories are equivalent *tout court* if they are empirically equivalent.

The positivist view of theoretical equivalence is quite liberal, and certainly unacceptable to scientific and metaphysical realists. Most of us have the intuition that theories can say different things about unobservable things, even if those theories agree in all their observational consequences. In this case, we have to reject empirical equivalence as a sufficient condition for theoretical equivalence.

A case can be made that Putnam's view of theoretical equivalence eventually became – at least tacitly and in practice – even more liberal than empirical equivalence. In putting forward the model-theoretic argument, Putnam essentially makes an argument for the following claim:

If T is consistent (and has other virtues such as completeness), then T ought to be taken as true.

Now, in application to *two* consistent theories T and T', we have the following result:

If T and T' are consistent (and have other virtues such as completeness), then T and T' ought both to be taken as true.

In other words, consistent, ideal theories are true in all conditions, hence in all the same conditions, and so they are equivalent. That is a radically liberal view, almost Zeno-like in its implications. For in this case, there is only one equivalence class of consistent theories.

What I've left out from this story so far are all the intermediate (and more plausible) views of theoretical equivalence – views that we have been discussing throughout this book, such as definitional equivalence or Morita equivalence. To put everything together, consider the diagram that follows, which places the different views of theoretical equivalence on a one-dimensional spectrum from maximally liberal (Zenonian) to maximally conservative (Heraclitean).

Zeno ← categorical ← w-intertranslatable ↔ Morita ← s-intertranslatable ↔ CDE ← logical ← Heraclitus

So, given this wide range of different notions of equivalence, how are we to choose among them? And do we need to choose among them? I would say that we don't have to explicitly choose among them – but that our attitudes toward them mirror our attitudes toward real life cases, or at least to cases that come up in other philosophical discussions. Consider, for example, North's (2009) argument for the inequivalence of Hamiltonian and Lagrangian mechanics. She says, "Hamiltonian and Lagrangian mechanics are not equivalent in terms of statespace structure. This means that they are not equivalent, period." In other words, she's putting a model of Hamiltonian mechanics next to a model of Lagrangian mechanics and comparing structure. Seeing that these structures are not "equivalent," she declares that the theories are not equivalent. We see then that, at the very least, North adopts a criterion that is more conservative than categorical equivalence, which is blind to the internal structure of individual models. (In fact, Barrett [2018a] shows that Hamiltonian and Lagrangian mechanics are categorically equivalent.) Most likely, North's criterion is further to the right than even Hudetz's definable categorical equivalence (see Hudetz, 2018a), for she doesn't consider questions as to whether Lagrangian structure can be defined in terms of Hamiltonian structure, and vice versa.

We can see a similar thought process going on with critics of quantifier variance. Indeed, we can think of debates about quantifier variance as debates about which notion of theoretical equivalence to adopt. The opponents of quantifier variance insist that equinumerosity of models is a necessary condition for theoretical equivalence. Thus, they draw the line short of Morita equivalence, which allows that equivalent theories can have models of different cardinalities. In contrast, defenders of quantifier variance claim that theories can be intertranslatable even if they violate that cardinality constraint. The question boils down to which criterion of theoretical equivalence is the better one to adopt.

I believe that this is one of the most interesting questions that philosophers can ask, precisely because it's a non-factual question. Or, to put it more accurately, the answer that one gives to such a question determines what one thinks is a factual question – and so it's not the kind of question that two parties can easily resolve by appeal to a shared stock of facts. Nonetheless, we've made a lot of progress on the technical side, so we now have a much more clear sense of what's at stake and the price we must pay for adopting some particular formal notion of equivalence as an intuitive guide to our practice of judging between theories.

Consider, for example, the distinction between definitional equivalence and Morita equivalence – or what is the same, between strong and weak intertranslatability. The line between these two notions of equivalence seems to correspond pretty well to the distinction between metaphysical realists and, well, those who aren't quite metaphysical realists. (The metaphysical realist might insist that if theories are equivalent, then their models have the same number of objects.) However, we shouldn't forget that Morita equivalence isn't all that liberal. It's certainly far more conservative than what Putnam was suggesting in the model-theoretic argument.

We can also see that the ontology of Morita equivalent (i.e., weakly intertranslatable) theories can never be radically different from each other. If $F : T \to T'$ is a homotopy equivalence (between single-sorted theories), then for each model M of T', there is a model F^*M of T, whose domain is explicitly constructed by the recipe:

$$(F^*M)(\sigma) \;=\; M(\sigma') \times \cdots \times M(\sigma')/\sim,$$

where $\sim$ is an equivalence relation defined by the theory T'. There are a couple of important points here. First, the ontology of F^*M results from simple logical constructions of the ontology of M. Borrowing terminology from Bertrand Russell, we could say that the elements of F^*M are *logical constructs* of elements of M. Second, the recipe for constructing F^*M from M is *uniform* – i.e., it doesn't depend on M. In other words, it's not just that each model of T consists of logical constructs of elements of a model of T'; it's that the type of construction is uniform. It's in this extended and, nonetheless, quite strong sense that T has the same ontology as T'.

Moreover, since F is assumed to be a homotopy equivalence, we can say the same thing in reverse order: each model of T' consists of logical construct of elements of a model of T, and this construction is uniform on models. One bonus insight here is seeing how the relation "being a logical construct of" differs from the mereological parthood relation. Consider the specific example of the point and line formulations of affine geometry (see Section 7.4). Here the points are logical constructs of lines, and the lines are logical constructs of points. It's tempting to think then that points are logical constructs of points – but that would be incorrect. The reason that inference doesn't go through is that "being a logical construct of" is not like the mereological notion of parthood. To get a line, we don't simply take two points; we take an equivalence class of two points. Thus, there is no sense here in which a line results from taking a composite of points. The opposite direction is even more clear. We can construct points from lines, but certainly a point is not made out of lines.

The upshot of these considerations is that moving from definitional equivalence to Morita equivalence is not as radical a generalization (or liberalization) as it might seem at first. Even for the ontological purists, a case could be made that Morita equivalence involves only the slightest relaxation of the constraint that equivalent theories should have equinumerous domains.

In contrast, categorical equivalence is extremely liberal from an ontological point of view. It's possible, indeed, to have categorically equivalent theories where there is no reasonable sense in which the ontology of the first's models can be constructed from the ontology of the second's models.

There are, however, some intermediate cases that are worth considering. Some of these are discussed by Hudetz (2018a). Here we just look at one example that will be familiar from Chapter 3. Consider the categories **Bool**, of Boolean algebras, and **Stone**, of Stone spaces. As we proved, **Bool** is equivalent to the opposite of **Stone**, where the arrows have been flipped. Moreover, the functors relating these two categories do have a strongly constructive flavor. The functor $F : \mathbf{Bool} \to \mathbf{Stone}^{op}$ is the representable functor $\hom(-, 2)$, where 2 is the two-element Boolean algebra. The functor $G : \mathbf{Stone}^{op} \to \mathbf{Bool}$ takes the clopen subsets. In both cases, the functor involves construction of an object of one category out of an object of the second category, and possibly some reference object, such as 2.

Could these latter sorts of functors be taken as representing genuine theoretical equivalences? There are two clarifications we need to raise for that question. First, the question doesn't even make sense until we say something more about how a category, which may not be of the form $\mathrm{Mod}(T)$ for a first-order theory, can represent a theory. Second, for many physical theories – and *pace* Quine – the elements of a mathematical domain X are not necessary meant to represent objects in the physical world. Consider the following example, which – besides being extremely interesting in its own right – illustrates several of these points.

General relativity (GTR), qua mathematical object, can roughly be taken to be the category **Lor** of Lorentzian manifolds, equipped with an appropriate collection of smooth mappings between them. There has been a longstanding debate – stimulated, no doubt, by Quine's criterion of ontological commitment – about whether accepting GTR demands that one accept the existence of spacetime points. Perhaps partially in response to that claim, Earman noted that GTR could also be formulated in terms of mathematical objects called "Einstein algebras." The relationship between Lorentzian manifolds and Einstein algebras is suggestively parallel to the relationship between Stone spaces and Boolean algebras. This parallel was confirmed by Rosenstock et al. (2015), who showed that **Lor** is dual to the category **EAlg** of Einstein algebras.

If one takes categorical equivalence as the criterion for theoretical equivalence, then the Einstein algebra formulation of GTR is no better nor worse than the Lorentzian manifold formulation. However, one might also wish to draw a stronger conclusion: one might wish to say that Rosenstock et al.'s proof shows that accepting GTR does *not* involve ontological commitment to spacetime points.

However, that conclusion would be hasty. The implicit argument pattern here would run as follows:

Let T be a theory with a sort σ. If T is equivalent to T', and T' doesn't quantify over σ, then to accept T cannot involve ontological commitment to things of type σ.

To see that this inference pattern proves too much, we can consider some simple examples. First, consider the example of the theory T in sort $\Sigma = \{\sigma\}$ that says there are exactly two things, and consider the theory T' in sort $\Sigma' = \{\sigma'\}$ that says there are exactly two things. By the preceding inference rule we would have to conclude that accepting T does not demand ontological commitment to things of type σ, merely because there is another sort symbol σ'. This is silly. The difference between σ and σ' could be simply notational.

Perhaps then the argument pattern is meant to be a bit more nuanced.

Let T be a theory with sort σ. If T is equivalent to a theory T', and T' has no sort σ' that is "isomorphic" to σ, then accepting T does not involve ontological commitment to things of type σ.

The word "isomorphic" was put into quotes because we would still need to explicate what we mean by it. But that could be done; e.g., we might say that an equivalence $F : T \to T'$ shows that σ and σ' isomorphic if $F(\sigma) = \sigma'$ and $E_{x,y} \equiv (x =_\sigma y)$ for variables x, y of sort σ. But in this case, the proposed criterion simply begs the question against the idea that Morita equivalent theories can have the "same ontology." To take Morita equivalence seriously as a criterion of theoretical equivalence means simply that there is no cross-theoretical reference point for counting objects or quantifying over them.

8.5 Flat versus Structured Views of Theories

For the past fifty years, philosophers' discussions of the nature of scientific theories has been dominated by the dilemma: are theories sets of sentences, or are theories collections of models? But the point of this debate has become less and less clear. Most of us these days are non-essentialists about mathematical explications. For example, most of us don't think that scientific theories really are sets of axioms or collections of models. Instead, we think that different explications are good for different purposes. There is, nonetheless, a big question lurking in the background – viz. the question of whether we should conceive of theories as "flat," or whether we should conceive of them as "structured." And this question comes up whether one thinks that theories are made of sentences or whether one thinks that they are made of models.

The syntactic view of theories is usually formulated as follows:

A theory is a *set* of sentences.

This formulation provides a *flat* view: a theory consists of a collection of things, and not in any relations between those things or structure on those things. In contrast, a *structured* view of theories says that scientific theories are best represented by structured mathematical objects. For example, a structured syntactic view of theories might say that a theory consists of both sentences and inferential relations between those sentences.

A flat version of the semantic view might be formulated as

A theory is a *set* (or *class*) of models.

In contrast, a structured version of the semantic view will say that a theory consists of a structured collection of models. For example, a theory might consist of models with certain mappings between these models (such as elementary embeddings), or a theory might consist of models and certain "nearness" relations between those models.

Both the syntactic and the semantic views of theories are typically presented as flat views. In the latter case, I suspect that the flat point of view is accidental. That is, most proponents of the semantic view are not ideologically committed to the claim that a theory is a bare set (or class) of models. They may not have realized the implications of that claim or that there is an alternative to it.

In contrast, in the case of syntactically oriented views, some twentieth-century philosophers were ideologically committed to a flat view – perhaps due to their worries about intensional and/or normative concepts. The main culprit here is Quine, whose criticism of the analytic–synthetic distinction is directed precisely against a structured view of theories. On a structured syntactic view of theories, the essential structure of a theory includes not just some number of sentences, but also the logical relations between those sentences. In this case, commitment to a theory would involve claims about inferential relations – in particular, claims about which sentences are logical consequences of the empty set. In other words, a structured syntactic view of theories presupposes an analytic–synthetic distinction.

Quine's powerful criticisms of the analytic–synthetic distinction raise worries for a structured syntactic picture of theories. But is all well with the unstructured, or flat, syntactic view? I maintain that the unstructured view has *severe* problems that have never been addressed. First of all, if theories are sets of sentences, then what is the criterion of equivalence between theories? A mathematically minded person will be tempted to say that between two *sets*, there is only one relevant condition of equivalence, namely equinumerosity. But certainly we don't want to say that two theories are equivalent if they have the same number of sentences! Rather, if two theories are equivalent, then they should have some further structure in common. What structure should they have in common? I would suggest that, at the very least, equivalent theories ought to share the same inferential relations. But if that's the case, then the content of a theory includes its inferential relations.

8.6 Believing a Scientific Theory

The difference between scientific realists and antirealists is supposed to be that the former believe scientific theories, and the latter do not – or at least they don't believe everything that these theories say. For example, constructive empiricists like van Fraassen don't necessarily believe what scientific theories say about unobservable things. This classification is based on a presupposition, viz. that we understand what it means to "believe everything a scientific theory says." But there is something wrong with these

presupposition. On none of the reasonable analyses is a scientific theory nothing more than some claims about the world. If that's right, then the appropriate attitude to a successful scientific theory cannot be exactly the same thing as simple belief.

To see what's at issue here, it will be helpful to revisit an old objection to the semantic view of theories. According to the semantic view of theories, a scientific theory is a class of models. Now, the objector to the semantic view points out that there is a grammatical problem: in the phrase "S believes that X," the second argument X needs to be filled by something toward which a person can bear a propositional attitude. The argument X cannot be replaced by a name such as "Thor," or predicate such as "purple," much less by a name for a class of things, such as "the set of ..." In particular, it makes no grammatical sense to say that "S believes that $\mathscr{M}$," where $\mathscr{M}$ is a class of models.

The semanticists have a ready reply to this objection:

Semantic Analysis of Belief (SAB): When a theory T is given by means of a class $\mathscr{M}$ of models, then belief in T means belief that the world is isomorphic to one of the models in $\mathscr{M}$.

There are many problems with SAB, most notably the opacity of the notion of a model being isomorphic with the concrete world (see, e.g., Van Fraassen, 2008). However, there is another problem with SAB that we find even more serious, because it bears directly on questions of a normative nature, e.g., to what one commits oneself when one accepts a scientific theory. In particular, believing a theory involves further commitments beyond those that are expressed by SAB.

Consider a specific example. Let T be Einstein's general theory of relativity (GTR). According to SAB, a person believes GTR iff she believes that the world is isomorphic to one of the models of GTR. But that analysis is inaccurate in both directions: it captures both more and less than physicists actually believe when they accept GTR. First, it captures more, because it seems to commit physicists to the belief that there is some privileged model of GTR that gives the best overall picture of the physical world. If you know how GTR works, then you might laugh at that thought. Just imagine two relativists – say, a cosmologist and a black hole theorist – sitting down to argue over whose model gives a more perspicuous representation of reality. They won't do that, because they are well aware that these models are accurate representations for certain purposes and not for others. And what's more, it's we – the users of physics – who choose the intended application of the theory. Thus, SAB says more than physicists will actually want to say about their theories.

Second, the semantic analysis of belief (SAB) also omits some of the content that physicists pack into their theories. Indeed, SAB locates the content of a theory in one or other particular model, ignoring the fact that physicists routinely invoke the existence of other models, not to speak of a rich system of relations between models. Indeed, if a model M is removed from its context in $\mathrm{Mod}(T)$, then it can no longer do the representational and explanatory work that it's expected to do. Consider again the case of GTR. As we noted before, GTR is a powerful theory not because it is overly specific, but because it is widely applicable – offering different, but related, models for a wide variety of situations. GTR finds the unity between these situations, including counterfactual situations. (David Lewis said that a good theory balances informativeness and

simplicity. However, there are different ways of being informative: saying what is unique about your situation or saying what is common among many different situations.)

Furthermore, some of the most powerful explanations in GTR draw on facts about how a model sits inside the space of all models, some of which we know not to represent the actual world. For example, what explains the fact that our universe began in a singularity? According to GTR, singular spacetimes are generic, i.e., they densely pack the space of cosmological solutions to Einstein's field equations; hence, the reason our universe begins in a singularity is because most nomologically possible universes begin this way.

The fact that GTR uses all of its models, and the relations between them, is only reinforced by looking at simple examples from first-order logic. If we take a first-order theory T, then typically a single model M of T does not contain enough information to reconstruct T. In other words, if you give me a model M of T, I couldn't reliably reconstruct the theory T of which it was a model. What that means is that M contains less information than the theory T itself. The content of the syntactic object T is not contained in a single model M, but in the structured collection $\mathrm{Mod}(T)$ of all its models. What this means in turn is that accepting T cannot be reduced to a claim about one of the models in $\mathrm{Mod}(T)$; instead, accepting T must involve some sort of attitude toward the entire collection $\mathrm{Mod}(T)$.

The point we are making here ties all the way back to the preface of the book, where we tried to justify our omission of modal logic. There we claimed that accepting a first-order theory – with no explicit modal operators – involves modal commitments. We're making the same point here. To accept a theory T isn't just to take a stand on how the world *is*; it is also to take a stand on how the world *could be*. More is true. To accept a theory T involves choosing a language Σ, and this language determines how we parse the space of possibilities – e.g., which possibilities we consider to be isomorphic, and which we consider not to be isomorphic. (If you've read the previous chapters carefully, you're also aware that the language Σ determines the topological structure of $\mathrm{Mod}(T)$.) In short, the syntactic approach to theories had the advantage (largely unnoticed by its proponents) that the syntactic object T packs in a lot of information about what is possible and about how to classify possibilities. One of the dangers of the semantic view is forgetting how much scientific theories say.

The fault here doesn't lie completely with the semantic view of theories. In fact, there's an analogous problem for those, such as Quine, who accept a flat syntactic view of theories (see Section 8.5). According to the flat syntactic view, a theory T is a *set* of sentences. Indeed, Quine – among other flat syntacticists – sometimes equates belief in T with belief in a set of sentences. But that cannot be quite right, as we can see again from actual scientific theories, as well as from simple examples from first-order logic.

As for examples from first-order logic, let Σ_1 be the empty signature, and let T_1 be the theory in Σ_1 that says there are exactly two things. Let $\Sigma_2 = \{c\}$, where c is a constant symbol, and let T_2 be the theory in Σ_2 that says there are exactly two things. Here T_1 and T_2 share the same axiom, but they aren't equivalent theories by any reasonable standard – not even by categorical equivalence. The first theory's model has automorphism group $\mathbb{Z}_2$, whereas the second theory's model has trivial automorphism

group (since the denotation of c is fixed). Nonetheless, T_1 and T_2 agree on the statements that they make about any particular model: they both say that there are two things. The user of T_2 has an extra name c, but her using this name does not amount to any claim about how things are. Thus, we have a puzzle: on a world-by-world basis, T_1 and T_2 say the same thing; and yet, it's not reasonable to think that T_1 and T_2 say the same thing.

The solution to this little puzzle is to recognize that believing a theory cannot be reduced to believing that a certain collection of sentences is true. At the very least, believing a theory also requires that we adopt a language – or an "ideology," as Quine liked to call it. However, Quine wasn't completely clear on what the reasons might be for accepting an ideology. The issues became slightly clearer when Lewis suggests that our choice of ideology corresponds to our beliefs about which properties are "natural," and when Sider (2013) suggests that choice of ideology is tantamount to assertion that the world has a certain structure. While we don't necessarily agree with this way of describing the situation, we agree that ideology plays a theoretical role.

If we claim that a theory is a collection of sentences, then we ought also to accept the claim that theories are equivalent only if they contain the same sets of sentences. Or, to be more accurate, two theories are equivalent just in case each sentence in the first is equivalent to a sentence in the second, and vice versa. But now, what standard of equivalence should we use for the sentences? The only reasonable standard – two sentences are equivalent if they express the same proposition – is of no use in comparing actual scientific theories. Thus, the only reasonable account of the identity of scientific theories treats theories as a *structured* objects, in which case equivalence means having the same structure. And then we have a challenge question: what does it mean to believe or accept a structured object?

It might be illuminating to compare a scientific theory with the kinds of beliefs for which people live and die – e.g., religious beliefs. As you know, many western religions have creeds that are supposed to capture the key tenets of the system of belief. Now, suppose that you were to try to write down the central tenets of a scientific theory as a creed. For example, you might take a copy of Robert Wald's *General Relativity* and start searching through it for the basic "truth claims" of the theory. However, you'll quickly grow frustrated, as it doesn't seem to make any specific claims. GTR doesn't say what happened on December 7, 1941, nor does it say how many planets are in our solar system, nor does it say (before one selects a particular model for application) how old the universe is. Instead, GTR consists of some mathematics and some recommendations about how to apply this mathematics to various situations. And yet, there is never any hint that GTR is a bad theory because it's not specific enough. Quite to the contrary, GTR is a good theory precisely because it is so general.

One might be tempted to think that the creed of GTR is summed up in its basic equation, Einstein's field equation (EFE). In this case, to accept GTR would be to say:

(†) I believe that $R_{ab} - \frac{1}{2} g_{ab} R = T_{ab}$.

This is an interesting possibility to consider, and there are two attitudes we could take to it. I will call these two attitudes the physicist's attitude, and the metaphysician's attitude (almost certainly caricaturing both). In my experience, physicists don't say

things like (†). Certainly, they write down EFE, and they use it to generate descriptions of situations that they take to be accurate. But I've never heard a physicist say, "I believe that Einstein's field equation is true." These physicists seem to have a positive attitude toward EFE – perhaps we should call it "acceptance," but I don't think we could call it "belief."

In contrast, the naive scientific realist might say something like: "The success of GTR gives us reason to think that EFE is true." In order to make sense of EFE being true, these realists will then cast about for referents for the terms that occur in it. For example, in the spirit of David Armstrong, they might say something like, "The symbols R_{ab} and T_{ab} refer to natural properties, and EFE is the statement that a second-order relation holds between these properties." This kind of realist seems to think that there aren't enough mundane physical objects to account for the meaning of the abstract statements of science. Accordingly, he makes up names for new things that can serve as the referents for the symbols in scientific statements such as EFE. After some subtle wordplay, we're supposed to be able to feel what it means to really believe that EFE is true.

However, this naive realist way of looking at EFE doesn't capture the way that these kinds of equations function in physics. As with any other differential equation in physics, EFE is used as a guide for differentiating between what is nomologically possible and what is not. A differential equation doesn't say how things are; it says how things could be.

It might sound like I'm simply endorsing instrumentalism, i.e., saying that the theoretical statements of science are mere instruments from which to derive predictions. But that accusation depends on a false dilemma between naive realism and instrumentalism – a dilemma that is sadly reinforced by formal semantics. In formal semantics we have a simple, black-and-white distinction between interpreted and uninterpreted terms. Accordingly, we're tempted to think that the terms of EFE are either interpreted (hence EFE is either true or false) or uninterpreted (hence EFE is just an instrument). But this is the wrong way to think about things. The symbols in EFE in themselves are neither interpreted nor uninterpreted. It is we, users of the theory, who endow these symbols with an interpretation. What's more, we might well want to interpret the symbols differently for different applications.

The existence of more than one model – or, to speak more accurately, of more than one application – is not a bug of scientific theories; it is a feature. What is lost in informativeness is gained in applicability. But the more flexible a theory is in its applications, the less sense it makes to think of our attitude toward that theory as simple "belief." Perhaps this is one reason why we need another word, such as "acceptance." As van Fraassen pointed out long ago, to accept a theory cannot be reduced to an attitude that the theory somehow mirrors the world. Acceptance of a theory involves a sort of appropriation, where the theory serves as a guide to future action.

I've been considering the question, "what does it mean to accept a scientific theory?" and have found ample reason to reject the idea that it's nothing more than a special case of belief. Accepting a scientific theory may involve believing that some things are true, but it also involves a more complex set of attitudes – such as adopting certain standards for explanation, certain rules for reasoning about counterfactual scenarios, etc.

8.7 Notes

- For more technical details on second-order logic, see Shapiro (1991); Manzano (1996). Philosophers have argued quite a bit about the advantages and disadvantages of second-order logic. For example, Quine argued that second-order logic is "set theory in sheep's clothing." See, e.g., Bueno (2010).
- Carnap gives his mature view of Ramsey sentences in Carnap (1966). For more on the role of Ramsey sentences in Carnap's philosophy of science, see Psillos (2000, 2006); Friedman (2011); Demopoulos (2013).
- For more on the Ramsey sentence functionalism, see Shoemaker (1981).
- For a detailed, but older, discussion of the technical issues surrounding Ramsey sentences, see Tuomela (1973, chapter 3). For a recent discussion of the prospects of Ramsey sentence structuralism, see Ketland (2004); Melia and Saatsi (2006); Ainsworth (2009); Dewar (2019).
- For general surveys of structural realism, see Frigg and Votsis (2011); Ladyman (2014). The idea behind structural realism goes much further back than the 1980s. Something similar had been proposed by Poincaré and Russell in the early 1900s, and then again by Grover Maxwell in the 1960s. What's new about the 1990s reincarnation of structural realism is (1) the explicit claim that it can solve the pessimistic metainduction and (2) the explication of structure in terms of Ramsey sentences. Needless to say, the idea behind structural realism could survive, even if – as we've argued – Ramsey sentences don't provide a useful explication of the structure of a theory.
- My view on counting possibilities was influenced by Weatherall (2016b).
- Putnam's model-theoretic argument first appeared in Putnam (1977, 1980), with antecedents in Quine's permutation arguments for ontological relativity. The most influential response to Putnam is Lewis' (1984), which is the *locus classicus* of his version of metaphysical realism which emphasizes the notion of *natural properties*. That torch has been taken up by Sider (2013). The response we gave to Putnam's argument follows the spirit of Van Fraassen (1997). For an excellent overview of Putnam's arguments, see Button (2013).