5 Syntactic Metalogic Redux

5.1 Many-Sorted Logic

We now turn to a generalization of first-order logic – a generalization that has proven to be surprisingly controversial. This generalization proceeds by noting that in ordinary first-order logic, it is implicitly assumed that all syntactic objects are compatible. For example, for any two terms s, t, it makes sense to write $s = t$; and for any relation symbol r, and terms $t_1, \ldots, t_n$, it makes sense to write $r(t_1, \ldots, t_n)$. However, that assumption is not obviously warranted. Instead, one might insist that syntactic objects, such as terms, come with a **type** or **sort**, and that there are sort-based rules about how these objects can be combined.

This generalization can provoke two responses that pull in completely opposite directions. On the one hand, one might think that many-sorted logic is stronger than single-sorted logic, and hence that its theoretical commitments outrun those of single-sorted logic. (The obvious analogy here is with second-order logic.) On the other hand, some philosophers, such as Quine (1963, 267–268), argue that many-sorted logic is reducible to single-sorted logic, and hence is dispensable. If we give pride of place to classical (single-sorted) first-order logic, then both of these responses would undermine our motivation to study many-sorted logic. However, the presuppositions of these two responses cannot both be correct – i.e., many-sorted logic cannot both exceed the resources of single-sorted logic and also be reducible to it. So which view is the right one?

The view we will advance here is that many-sorted logic is, in one clear sense, reducible to single-sorted logic, but that this reduction does not mean that many-sorted logic is dispensible. Before we take up this argument, we need to explain how many-sorted logic works.

DEFINITION 5.1.1 A many-sorted **signature** Σ is a set of sort symbols, predicate symbols, function symbols, and constant symbols. Σ must have at least one sort symbol. Each predicate symbol $p \in \Sigma$ has an **arity** $\sigma_1 \times \ldots \times \sigma_n$, where $\sigma_1, \ldots, \sigma_n \in \Sigma$ are (not necessarily distinct) sort symbols. Likewise, each function symbol $f \in \Sigma$ has an **arity** $\sigma_1 \times \ldots \times \sigma_n \rightarrow \sigma$, where $\sigma_1, \ldots, \sigma_n, \sigma \in \Sigma$ are again (not necessarily distinct) sort symbols. Lastly, each constant symbol $c \in \Sigma$ is assigned a sort $\sigma \in \Sigma$. In addition to the elements of Σ, we also have a stock of variables. We use the letters x, y, and z to denote these variables, adding subscripts when necessary. Each variable has a sort $\sigma \in \Sigma$.

NOTE 5.1.2 The symbol $\sigma_1 \times \cdots \times \sigma_n$ has no intrinsic meaning. To say that "p has arity $\sigma_1 \times \cdots \times \sigma_n$" is simply an abbreviated way of saying that p can be combined with n terms, whose sorts must respectively be $\sigma_1, \ldots, \sigma_n$.

A Σ-term can be thought of as a "naming expression" in the signature Σ. Each Σ-term has a sort $\sigma \in \Sigma$.

DEFINITION 5.1.3 The **Σ-terms** of sort σ are recursively defined as follows. Every variable of sort σ is a Σ-term of sort σ, and every constant symbol $c \in \Sigma$ of sort σ is also a Σ-term of sort σ. Furthermore, if $f \in \Sigma$ is a function symbol with arity $\sigma_1 \times \ldots \times \sigma_n \rightarrow \sigma$ and $t_1, \ldots, t_n$ are Σ-terms of sorts $\sigma_1, \ldots, \sigma_n$, then $f(t_1, \ldots, t_n)$ is a Σ-term of sort σ. We will use the notation $t(\vec{x})$ to denote a Σ-term in which all of the variables that appear in t are in the sequence $\vec{x} \equiv x_1, \ldots, x_n$, but we leave open the possibility that some of the x_i do not appear in the term t.

A **Σ-atom** is an expression either of the form $s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n)$, where s and t are Σ-terms of the same sort $\sigma \in \Sigma$, or of the form $p(t_1, \ldots, t_n)$, where $t_1, \ldots, t_n$ are Σ-terms of sorts $\sigma_1, \ldots, \sigma_n$ and $p \in \Sigma$ is a predicate of arity $\sigma_1 \times \ldots \times \sigma_n$.

DEFINITION 5.1.4 The **Σ-formulas** are defined recursively as follows.

- Every Σ-atom is a Σ-formula.
- If ϕ is a Σ-formula, then $\neg\phi$ is a Σ-formula.
- If ϕ and ψ are Σ-formulas, then $\phi \rightarrow \psi$, $\phi \wedge \psi$, $\phi \vee \psi$, and $\phi \leftrightarrow \psi$ are Σ-formulas.
- If ϕ is a Σ-formula and x is a variable of sort $\sigma \in \Sigma$, then $\forall_\sigma x\phi$ and $\exists_\sigma x\phi$ are Σ-formulas.

In addition to the preceding formulas, we will use the notation $\exists_{\sigma=1} y\phi(x_1, \ldots, x_n, y)$ to abbreviate the formula

$$\exists_\sigma y(\phi(x_1, \ldots, x_n, y) \wedge \forall_\sigma z(\phi(x_1, \ldots, x_n, z) \rightarrow y = z)).$$

As before, the notation $\phi(\vec{x})$ will denote a Σ-formula ϕ in which all of the free variables appearing in ϕ are in the sequence $\vec{x} \equiv x_1, \ldots, x_n$, but we again leave open the possibility that some of the x_i do not appear as free variables in ϕ.

DEFINITION 5.1.5 A **Σ-sentence** is a Σ-formula that has no free variables.

We will not give an explicit listing of the derivation rules for many-sorted logic. Suffice it to say that they are direct generalizations of the derivation rules for single-sorted logic, provided that one observe all restrictions on syntactic compatibility. For example, in many sorted logic, we can infer $\forall x\phi(x)$ from $\phi(y)$ only if the variables x and y are of the same type. If they were not of the same type, then one of these two formulas would fail to be well-formed.

As a result of these restrictions, we need to exercise some caution about carrying over intuitions that we might have developed in using single-sorted logic. For example, in single-sorted logic, for any two terms s and t, we have a tautology

$$\vdash (s = t) \vee (s \neq t).$$

However, in many sorted logic, the expressions $s = t$ and $s \neq t$ are well-formed only when s and t are terms of the same sort. Thus, to the question "do s and t denote the same object?" many-sorted logic sometimes offers no answer.

One might be tempted, nonetheless, to think that if s and t are terms of different sorts, then we can just add $t \neq s$ as axiom. However, that suggestion can lead to disaster. For example, suppose that s denotes the number 0 and that t denotes the renowned actor David Hasselhoff. Because I accept Peano arithmetic, I assume that every natural number besides 0 is greater than 0. In other words, I assume that

$$\forall x(x \neq s \rightarrow (x > 0)),$$

where x is a variable ranging over natural numbers. If I now added $t \neq s$ to my total theory, then I would be committed to the claim that David Hasselhoff is greater than 0. These considerations show that we need to exercise caution when moving between many- and single-sorted frameworks.

Example 5.1.6 Let $\Sigma = \{\sigma_1, \sigma_2\}$, and let T be the empty theory in Σ. Note that both $\exists_{\sigma_1} x(x = x)$ and $\exists_{\sigma_2} y(y = y)$. This might seem like a strange consequence: T is the empty theory, and you might think that the empty theory should have no nontrivial consequences. But the combination of $\exists_{\sigma_1} x(x = x)$ and $\exists_{\sigma_2} y(y = y)$ seems like a nontrivial consequence, viz. that there are at least two things!

However, there is a mistake in our reasoning. Those two sentences together do not imply that there are at least two things. For there is no third quantifier $\exists$ such that $\exists v \exists w(v \neq w)$ is guaranteed to hold.

These considerations show that distinct sort symbols do not necessarily represent different kinds of things. Indeed, it is not generally valid to infer that there are $n + m$ objects from the fact that there are n objects of sort σ_1 and m objects of sort σ_2. ⌟

Example 5.1.7 Let $\Sigma = \{\sigma_1, \sigma_2, i\}$, where $i : \sigma_1 \rightarrow \sigma_2$. Let T be the theory that says that i is bijective; that is, i is injective:

$$(i(x) = i(y)) \rightarrow x = y,$$

and i is surjective:

$$\exists x(i(x) = z).$$

Then T defines a functional relation ϕ of sort $\sigma_2 \times \sigma_1$ by means of

$$\phi(z, x) \leftrightarrow (i(x) = z).$$

The function $j : \sigma_2 \rightarrow \sigma_1$ corresponding to ϕ is the inverse of i. ⌟

Example 5.1.8 The theory of categories can conveniently be formulated as a many-sorted theory. Let $\Sigma = \{O, A, d_0, d_1, i, \circ\}$, where O and A are sorts, $d_0 : A \rightarrow O$, $d_1 : A \rightarrow O$, $i : O \rightarrow A$, and $\circ$ is a relation of sort $A \times A \times A$. (The relation $\circ$ is used as the composition function on arrows – i.e., a partial function defined for compatible arrows.) We will leave it as an exercise for the reader to write down the axioms corresponding to the following ideas:

1 For each arrow f, $d_0 f$ is the domain object, and $d_1 f$ is the codomain object. Thus, we may write $f : d_0 f \to d_1 f$. More generally, we write $f : x \to y$ to indicate that $x = d_0 f$ and $y = d_1 f$. The function $\circ$ is defined on pairs of arrows where the first arrow's domain matches the second arrow's codomain.
2 The function $\circ$ is associative.
3 For each object x, $i(x) : x \to x$. Moreover, for any arrow f such that $d_1 f = x$, we have $i(x) \circ f = f$. And for any arrow g such that $d_0 g = x$, we have $g \circ i(x) = g$.

⌟

What can many-sorted logic do for us? In pure mathematics, it can certainly have pragmatic advantages to introduce sorts. For example, in axiomatizing category theory, it seems more intuitive to think about objects and arrows as different sorts of things, rather than introducing some predicate that is satisfied by objects but not by arrows. Similarly, in axiomatizing the theory of vector spaces, it is convenient to think of vectors and scalars as different sorts of things. Indeed, in this latter case, it's hard to imagine a mathematician investigating the question: "is c a scalar or a vector?" Instead, it seems that general words like "vector" and "scalar" function more like labels than they do as names of properties that mathematicians are interested in investigating.

But what about empirical theories? Could a many-sorted formulation of an empirical theory provide a more perspicuous representation of the structure of reality? Let's focus on a more specific question, that was central to twentieth-century philosophy of science: can the distinction between observable and unobservable be encoded into the syntax of a theory?

Suppose then that in formulating a theory T, we begin by introducing a sort symbol O for observable objects, and a sort symbol P for theoretical objects. Then, any relation symbol R must be explicitly sorted – i.e., each slot after R can be occupied only by terms of one particular sort. Similarly, formulas such as $t = t'$ and $t \neq t'$ are well-formed only if t and t' are terms of the same type. It should be clear now that this language does not have a predicate "is unobservable," nor does it have any well-formed expression corresponding to the sentence:

($*$) No theoretical entity (i.e., entity of type P) is an observable entity (i.e., entity of type O).

The grammatical malformity of ($*$) is sometimes brushed right over in criticisms of the syntactic view of theories (e.g., van Fraassen, 1980), and in criticisms of the observation–theory distinction (e.g., Dicken and Lipton, 2006).

5.2 Morita Extension and Equivalence

Glymour (1971) claims that definitional equivalence (see 4.6.15) is a necessary condition on the equivalence of scientific theories. However, there are several reasons to believe that this criterion is too strict.

First, it is frequently argued that many-sorted logic is reducible to single-sorted logic (see Schmidt, 1951; Manzano, 1996). What is actually shown in these arguments is that for any many-sorted theory T, a corresponding single-sorted theory T' can be constructed. But what is the relation between T and T'? Obviously, the two theories T and T' cannot be definitionally equivalent, since that criterion applies only to single-sorted theories. Therefore, to make sense of the claim that many-sorted logic can be reduced to single-sorted logic, we need a generalization of definitional equivalence.

Second, there are well-known examples of theories that could naturally be formulated either within a single-sorted framework or within a many-sorted framework – and we need a generalization of definitional equivalence to explain in what sense these two formulations are equivalent. For example, category theory can be formulated as a many-sorted theory, using both a sort of "objects" and a sort of "arrows" (Eilenberg and Mac Lane, 1942, 1945); and category theory can also be formulated as a single-sorted theory using only "arrows" (Mac Lane, 1948). (Freyd [1964, p. 5] and Mac Lane [1971 p. 9] also describe this alternate formulation.) These two formulations of category theory are in some sense equivalent, and we would like an account of this more general notion of equivalence.

Third, definitional equivalence is too restrictive even for single-sorted theories. For example, affine geometry can be formalized in a way that quantifies over points, or it can be formalized in a way that quantifies over lines (see Schwabhäuser et al., 1983). But saying that the point theory (T_p) and the line theory (T_ℓ) both are formulations of the same theory indicates again that T_p and T_ℓ are in some sense equivalent – although T_p and T_ℓ are *not* definitionally equivalent. Indeed, the smallest model of T_p has five elements, which we can think of as the four corners of a square and its center point. On the other hand, the smallest model of T_ℓ has six elements. But if T_p and T_ℓ were definitionally equivalent, then every model M of T_ℓ would be the reduct of an expansion of a model M' of T_p (de Bouvére, 1965). In particular, we would have $|M| = |M'|$, which entails that T_ℓ has a model of cardinality five – a contradiction. Therefore, T_p and T_ℓ are not definitionally equivalent.

Finally, even if we ignore the complications mentioned previously, and even if we assume that each many-sorted theory T can be replaced by a single-sorted variant T' (by the standard procedure of unifying sorts), definitional equivalence is still inadequate. Consider the following example.

Example 5.2.1 Let T_1 be the objects-and-arrows formulation of category theory, and let T_2 be the arrows-only formulation of category theory. Intuitively, T_1 and T_2 are equivalent theories; but their single-sorted versions T_1' and T_2' are not definitionally equivalent. Indeed, $T_2' = T_2$, since T_2 is single sorted. However, T_1' has a single sort that includes both objects and arrows. Thus, while T_2' has a model with one element (i.e., the category with a single arrow), T_1' has no models with one element (since every model of T_1' has at least one object and at least one arrow). Therefore, T_1' and T_2' are not definitionally equivalent. ⌟

These examples all show that definitional equivalence does not capture the sense in which some theories are equivalent. If one wants to capture this sense, one needs a more general criterion for theoretical equivalence than definitional equivalence. Our aim here is to introduce one such criterion. We will call it **Morita equivalence**. This criterion is a natural generalization of definitional equivalence. In fact, Morita equivalence is essentially the same as definitional equivalence, except that it allows one to define new sort symbols in addition to new predicate symbols, function symbols, and constant symbols. In order to state the criterion precisely, we again need to do some work. We begin by defining the concept of a Morita extension. In Chapter 7, we will show the sense in which Morita equivalence is a natural generalization of definitional equivalence.

As we did for predicates, functions, and constants, we need to say how to define new sorts. Let $\Sigma \subseteq \Sigma^+$ be signatures and consider a sort symbol $\sigma \in \Sigma^+ \backslash \Sigma$. One can define the sort σ as a product sort, a coproduct sort, a subsort, or a quotient sort. In each case, one defines σ using old sorts in Σ and new function symbols in $\Sigma^+ \backslash \Sigma$. These new function symbols specify how the new sort σ is related to the old sorts in Σ. We describe these four cases in detail.

product sort In order to define σ as a product sort, one needs two function symbols $\pi_1, \pi_2 \in \Sigma^+ \backslash \Sigma$ with π_1 of arity $\sigma \to \sigma_1$, π_2 of arity $\sigma \to \sigma_2$, and $\sigma_1, \sigma_2 \in \Sigma$. The function symbols π_1 and π_2 serve as the "canonical projections" associated with the product sort σ. A sort definition of the symbols σ, π_1, and π_2 as a **product sort** in terms of Σ is a Σ^+-sentence of the form

$$\forall_{\sigma_1} x \forall_{\sigma_2} y \exists_{\sigma=1} z (\pi_1(z) = x \wedge \pi_2(z) = y).$$

One should think of a product sort σ as the sort whose elements are ordered pairs, where the first element of each pair is of sort σ_1 and the second is of sort σ_2.

coproduct sort One can also define σ as a coproduct sort. One again needs two function symbols $\rho_1, \rho_2 \in \Sigma^+ \backslash \Sigma$ with ρ_1 of arity $\sigma_1 \to \sigma$, ρ_2 of arity $\sigma_2 \to \sigma$, and $\sigma_1, \sigma_2 \in \Sigma$. The function symbols ρ_1 and ρ_2 are the "canonical injections" associated with the coproduct sort σ. A sort definition of the symbols σ, ρ_1, and ρ_2 as a **coproduct sort** in terms of Σ is a Σ^+-sentence of the form

$$\forall_\sigma z \big(\exists_{\sigma_1=1} x(\rho_1(x) = z) \vee \exists_{\sigma_2=1} y(\rho_2(y) = z)\big) \wedge \forall_{\sigma_1} x \forall_{\sigma_2} y \neg \big(\rho_1(x) = \rho_2(y)\big)$$

One should think of a coproduct sort σ as the disjoint union of the elements of sorts σ_1 and σ_2.

When defining a new sort σ as a product sort or a coproduct sort, one uses two sort symbols in Σ and two function symbols in $\Sigma^+ \backslash \Sigma$. The next two ways of defining a new sort σ only require one sort symbol in Σ and one function symbol in $\Sigma^+ \backslash \Sigma$.

subsort In order to define σ as a subsort, one needs a function symbol $i \in \Sigma^+ \backslash \Sigma$ of arity $\sigma \to \sigma_1$ with $\sigma_1 \in \Sigma$. The function symbol i is the "canonical inclusion" associated with the subsort σ. A sort definition of the symbols σ and i as a **subsort** in terms of Σ is a Σ^+-sentence of the form

$$\forall_{\sigma_1} x\big(\phi(x) \leftrightarrow \exists_\sigma z(i(z) = x)\big) \wedge \forall_\sigma z_1 \forall_\sigma z_2\big(i(z_1) = i(z_2) \rightarrow z_1 = z_2\big), \quad (5.1)$$

where $\phi(x)$ is a Σ-formula. One can think of the subsort σ as consisting of "the elements of sort σ_1 that are ϕ." The sentence (5.1) entails the Σ-sentence $\exists_{\sigma_1} x\phi(x)$. As before, we will call this Σ-sentence the **admissibility condition** for the definition (5.1).

quotient sort Lastly, in order to define σ as a quotient sort, one needs a function symbol $\epsilon \in \Sigma^+\backslash\Sigma$ of arity $\sigma_1 \rightarrow \sigma$ with $\sigma_1 \in \Sigma$. A sort definition of the symbols σ and ϵ as a **quotient sort** in terms of Σ is a Σ^+-sentence of the form

$$\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2\big(\epsilon(x_1) = \epsilon(x_2) \leftrightarrow \phi(x_1, x_2)\big) \wedge \forall_\sigma z \exists_{\sigma_1} x(\epsilon(x) = z), \quad (5.2)$$

where $\phi(x_1, x_2)$ is a Σ-formula. This sentence defines σ as a quotient sort that is obtained by "quotienting out" the sort σ_1 with respect to the formula $\phi(x_1, x_2)$. The sort σ should be thought of as the set of "equivalence classes of elements of σ_1 with respect to the relation $\phi(x_1, x_2)$." The function symbol ϵ is the "canonical projection" that maps an element to its equivalence class. One can verify that the sentence (5.2) implies that $\phi(x_1, x_2)$ is an equivalence relation. In particular, it entails the following Σ-sentences:

$$\forall_{\sigma_1} x(\phi(x, x))$$
$$\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2(\phi(x_1, x_2) \rightarrow \phi(x_2, x_1))$$
$$\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2 \forall_{\sigma_1} x_3\big((\phi(x_1, x_2) \wedge \phi(x_2, x_3)) \rightarrow \phi(x_1, x_3)\big).$$

These Σ-sentences are the **admissibility conditions** for the definition (5.2).

Now that we have presented the four ways of defining new sort symbols, we can define the concept of a Morita extension. A Morita extension is a natural generalization of a definitional extension. The only difference is that now one is allowed to define new sort symbols.

DEFINITION 5.2.2 Let $\Sigma \subset \Sigma^+$ be signatures and T a Σ-theory. A **Morita extension** of T to the signature Σ^+ is a Σ^+-theory

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

that satisfies the following conditions. First, for each symbol $s \in \Sigma^+\backslash\Sigma$, the sentence δ_s is an explicit definition of s in terms of Σ. Second, if $\sigma \in \Sigma^+\backslash\Sigma$ is a sort symbol and $f \in \Sigma^+\backslash\Sigma$ is a function symbol that is used in the sort definition of σ, then $\delta_f = \delta_\sigma$. (For example, if σ is defined as a product sort with projections π_1 and π_2, then $\delta_\sigma = \delta_{\pi_1} = \delta_{\pi_2}$.) And third, if α_s is an admissibility condition for a definition δ_s, then $T \vdash \alpha_s$.

Note that unlike a definitional extension of a theory, a Morita extension can have more sort symbols than the original theory.[1] The following is a particularly simple example of a Morita extension.

[1] Also note that if T^+ is a Morita extension of T to Σ^+, then there are restrictions on the arities of predicates, functions, and constants in $\Sigma^+\backslash\Sigma$. If $p \in \Sigma^+\backslash\Sigma$ is a predicate symbol of arity $\sigma_1 \times \ldots \times \sigma_n$, we immediately see that $\sigma_1, \ldots, \sigma_n \in \Sigma$. Taking a single Morita extension does not allow one to define predicate symbols that apply to sorts that are not in Σ. One must take multiple Morita extensions to do

Example 5.2.3 Let $\Sigma = \{\sigma, p\}$ and $\Sigma^+ = \{\sigma, \sigma^+, p, i\}$ be a signatures with σ and σ^+ sort symbols, p a predicate symbol of arity σ, and i a function symbol of arity $\sigma^+ \to \sigma$. Consider the Σ-theory $T = \{\exists_\sigma x p(x)\}$. The following Σ^+-sentence defines the sort symbol σ^+ as the subsort consisting of "the elements that are p."

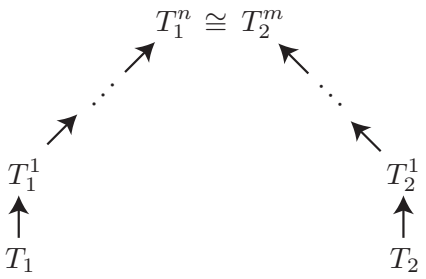
$$\forall_\sigma x\big(p(x) \leftrightarrow \exists_{\sigma^+} z(i(z) = x)\big) \wedge \forall_{\sigma^+} z_1 \forall_{\sigma^+} z_2\big(i(z_1) = i(z_2) \to z_1 = z_2\big). \qquad (\delta_{\sigma^+})$$

The Σ^+-theory $T^+ = T \cup \{\delta_{\sigma^+}\}$ is a Morita extension of T to the signature Σ^+. The theory T^+ adds to the theory T the ability to quantify over the set of "things that are p." ⌟

DEFINITION 5.2.4 Let T_1 be a Σ_1-theory and T_2 a Σ_2-theory. T_1 and T_2 are **Morita equivalent** if there are theories $T_1^1, \ldots, T_1^n$ and $T_2^1, \ldots, T_2^m$ that satisfy the following three conditions:

- Each theory T_1^{i+1} is a Morita extension of T_1^i,
- Each theory T_2^{i+1} is a Morita extension of T_2^i,
- T_1^n and T_2^m are logically equivalent Σ-theories with $\Sigma_1 \cup \Sigma_2 \subseteq \Sigma$.

Two theories are Morita equivalent if they have a "common Morita extension." The situation can be pictured as follows, where each arrow in the figure indicates a Morita extension.



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. By Theorem 4.6.17, if T' is a definitional extension of T, then T and T' are intertranslatable. Clearly intertranslatability is a transitive relation, and in Theorem 6.6.21, we will see that if two theories are intertranslatable, then they are definitionally equivalent. Therefore, 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

this. Likewise, any constant symbol $c \in \Sigma^+ \backslash \Sigma$ must be of sort $\sigma \in \Sigma$. And a function symbol $f \in \Sigma^+ \backslash \Sigma$ must either have arity $\sigma_1 \times \ldots \times \sigma_n \to \sigma$ with $\sigma_1, \ldots, \sigma_n, \sigma \in \Sigma$, or f must be one of the function symbols that appears in the definition of a new sort symbol $\sigma \in \Sigma^+ \backslash \Sigma$.

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 5.2.5 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 $\sigma_2 \to \sigma_1$ and $\sigma_3 \to \sigma_1$. Consider the following Σ-sentences.

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

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

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

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

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

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. ⌟

5.3 Quine on the Dispensability of Many-Sorted Logic

The notion of Morita equivalence bears directly on several central disputes in twentieth-century analytic philosophy. For example, in his debate with Carnap and the logical positivists, Quine claims that many-sorted logic is dispensable. Morita equivalence shows a precise sense in which Quine is right about that. Similarly, to motivate the rejection of metaphysical realism, Putnam claims that a geometric theory with points as primitives is equivalent to a theory with lines as primitives. (See, for example, Putnam, 1977,

489–491; Putnam, 1992, 109, 115–120; and Putnam, 2001.) Morita equivalence also shows a precise sense in which Putnam is right about that. We take up Quine's argument in the remainder of this section. We take up Putnam's argument in Section 7.4, after we have developed some semantic tools.

One proves Quine's claim by explicitly constructing a "corresponding" single-sorted theory $\widehat{T}$ for every many-sorted theory T. The basic idea behind the construction is intuitive. The theory $\widehat{T}$ simply replaces the sort symbols that the theory T uses with predicate symbols. This construction recalls the proof that every theory is definitionally equivalent to a theory that uses only predicate symbols (Barrett and Halvorson, 2016a, Prop. 2). Quine (1937, 1938, 1956, 1963) suggests the basic idea behind our proof, as do Burgess (2005, 12) and Manzano (1996, 221–222). However, the theorem that we prove here is more general than Quine's results because we make no assumption about what the theory T is, whereas Quine only considers Russell's theory of types and NBG set theory.

Let Σ be a signature with finitely many sort symbols $\sigma_1, \ldots, \sigma_n$. We begin by constructing a corresponding signature $\widehat{\Sigma}$ that contains one sort symbol σ. The symbols in $\widehat{\Sigma}$ are defined as follows. For every sort symbol $\sigma_j \in \Sigma$, we let q_{σ_j} be a predicate symbol of sort σ. For every predicate symbol $p \in \Sigma$ of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m}$, we let q_p be a predicate symbol of arity σ^m (the m-fold product of σ). Likewise, for every function symbol $f \in \Sigma$ of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m} \to \sigma_j$, we let q_f be a predicate symbol of arity σ^{m+1}. And, lastly, for every constant symbol $c \in \Sigma$ we let d_c be a constant symbol of sort σ. The single-sorted signature $\widehat{\Sigma}$ corresponding to Σ is then defined to be

$$\widehat{\Sigma} = \{\sigma\} \cup \{q_{\sigma_1}, \ldots, q_{\sigma_n}\} \cup \{q_p : p \in \Sigma\} \cup \{q_f : f \in \Sigma\} \cup \{d_c : c \in \Sigma\}.$$

We can now describe a method of "translating" Σ-theories into $\widehat{\Sigma}$-theories. Let T be an arbitrary Σ-theory. We define a corresponding $\widehat{\Sigma}$-theory $\widehat{T}$ and then show that $\widehat{T}$ is Morita equivalent to T.

We begin by translating the axioms of T into the signature $\widehat{\Sigma}$. This will take two steps. First, we describe a way to translate the Σ-terms into $\widehat{\Sigma}$-formulas. Given a Σ-term $t(x_1, \ldots, x_n)$, we define the $\widehat{\Sigma}$-formula $\widehat{\psi}_t(y_1, \ldots, y_n, y)$ recursively as follows.

- If $t(x_1, \ldots, x_n)$ is the variable x_i, then $\widehat{\psi}_t$ is the $\widehat{\Sigma}$-formula $y_i = y$.
- If $t(x_1, \ldots, x_n)$ is the constant c, then $\widehat{\psi}_t$ is the $\widehat{\Sigma}$-formula $d_c = y$.
- Suppose that $t(x_1, \ldots, x_n)$ is the term $f(t_1(x_1, \ldots, x_n), \ldots, t_k(x_1, \ldots, x_n))$ and that each of the $\widehat{\Sigma}$-formulas $\widehat{\psi}_{t_i}(y_1, \ldots, y_n, y)$ have been defined. Then $\widehat{\psi}_t(y_1, \ldots, y_n, y)$ is the $\widehat{\Sigma}$-formula

$$\exists_\sigma z_1 \ldots \exists_\sigma z_k \big(\widehat{\psi}_{t_1}(y_1, \ldots, y_n, z_1) \wedge \ldots \wedge \widehat{\psi}_{t_k}(y_1, \ldots, y_n, z_k) \wedge q_f(z_1, \ldots, z_k, y)\big).$$

One can think of the formula $\psi_t(y_1, \ldots, y_n, y)$ as the translation of the expression "$t(x_1, \ldots, x_n) = x$" into the signature $\widehat{\Sigma}$.

Second, we use this map from Σ-terms to $\widehat{\Sigma}$-formulas to describe a map from Σ-formulas to $\widehat{\Sigma}$-formulas. Given a Σ-formula $\psi(x_1, \ldots, x_n)$, we define the $\widehat{\Sigma}$-formula $\widehat{\psi}(y_1, \ldots, y_n)$ recursively as follows.

- If $\psi(x_1, \ldots, x_n)$ is $t(x_1, \ldots, x_n) = s(x_1, \ldots, x_n)$, where s and t are Σ-terms of sort σ_i, then $\widehat{\psi}(y_1, \ldots, y_n)$ is the $\widehat{\Sigma}$-formula

$$\exists_\sigma z\big(\widehat{\psi}_t(y_1, \ldots, y_n, z) \wedge \widehat{\psi}_s(y_1, \ldots, y_n, z) \wedge q_{\sigma_i}(z)\big).$$

- If $\psi(x_1, \ldots, x_n)$ is $p(t_1(x_1, \ldots, x_n), \ldots, t_k(x_1, \ldots, x_n))$, where $p \in \Sigma$ is a predicate symbol, then $\widehat{\psi}(y_1, \ldots, y_n)$ is the $\widehat{\Sigma}$-formula

$$\exists_\sigma z_1 \ldots \exists_\sigma z_k\big(\widehat{\psi}_{t_1}(y_1, \ldots, y_n, z_1) \wedge \ldots \wedge \widehat{\psi}_{t_k}(y_1, \ldots, y_n, z_k) \wedge q_p(z_1, \ldots, z_k)\big).$$

- This definition extends to all Σ-formulas in the standard way. We define the $\widehat{\Sigma}$-formulas $\widehat{\neg\psi} := \neg\widehat{\psi}$, $\widehat{\psi_1 \wedge \psi_2} := \widehat{\psi_1} \wedge \widehat{\psi_2}$, $\widehat{\psi_1 \vee \psi_2} := \widehat{\psi_1} \vee \widehat{\psi_2}$, and $\widehat{\psi_1 \to \psi_2} := \widehat{\psi_1} \to \widehat{\psi_2}$. Furthermore, if $\psi(x_1, \ldots, x_n, x)$ is a Σ-formula, then we define both of the following:

$$\widehat{\forall_{\sigma_i} x \psi} := \forall_\sigma y(q_{\sigma_i}(y) \to \widehat{\psi}(y_1, \ldots, y_n, y))$$
$$\widehat{\exists_{\sigma_i} x \psi} := \exists_\sigma y(q_{\sigma_i}(y) \wedge \widehat{\psi}(y_1, \ldots, y_n, y)).$$

One should think of the formula $\widehat{\psi}$ as the translation of the Σ-formula ψ into the signature $\widehat{\Sigma}$.

This allows us to consider the translations $\widehat{\alpha}$ of the axioms $\alpha \in T$. The single-sorted theory $\widehat{T}$ will have the $\widehat{\Sigma}$-sentences $\widehat{\alpha}$ as some of its axioms. But $\widehat{T}$ will have more axioms than just the sentences $\widehat{\alpha}$. It will also have some **auxiliary axioms**. These auxiliary axioms will guarantee that the symbols in $\widehat{\Sigma}$ "behave like" their counterparts in Σ. We define auxiliary axioms for the predicate symbols $q_{\sigma_1}, \ldots, q_{\sigma_n} \in \widehat{\Sigma}$, $q_p \in \widehat{\Sigma}$, and $q_f \in \widehat{\Sigma}$, and for the constant symbols $d_c \in \widehat{\Sigma}$. We discuss each of these four cases in detail.

We first define auxiliary axioms to guarantee that the symbols $q_{\sigma_1}, \ldots, q_{\sigma_n}$ behave like sort symbols. The $\widehat{\Sigma}$-sentence ϕ is defined to be $\forall_\sigma y(q_{\sigma_1}(y) \vee \ldots \vee q_{\sigma_n}(y))$.[2] Furthermore, for each sort symbol $\sigma_j \in \Sigma$ we define the $\widehat{\Sigma}$-sentence ϕ_{σ_j} to be

$$\exists_\sigma y(q_{\sigma_j}(y)) \wedge \forall_\sigma y\big(q_{\sigma_j}(y) \to (\neg q_{\sigma_1}(y) \wedge \ldots \wedge \neg q_{\sigma_{j-1}}(y) \wedge \neg q_{\sigma_{j+1}}(y) \wedge \ldots \wedge \neg q_{\sigma_n}(y))\big).$$

One can think of the sentences $\phi_{\sigma_1}, \ldots, \phi_{\sigma_n}$, and ϕ as saying that "everything is of some sort, nothing is of more than one sort, and every sort is nonempty."

Next we define auxiliary axioms to guarantee that the symbols q_p, q_f, and d_c behave like their counterparts p, f, and c in Σ. For each predicate symbol $p \in \Sigma$ of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m}$, we define the $\widehat{\Sigma}$-sentence ϕ_p to be

$$\forall_\sigma y_1 \ldots \forall_\sigma y_m \left(q_p(y_1, \ldots, y_m) \to \left(q_{\sigma_{j_1}}(y_1) \wedge \ldots \wedge q_{\sigma_{j_m}}(y_m)\right)\right).$$

This sentence restricts the extension of q_p to the subdomain of n-tuples satisfying $q_{\sigma_{j_1}}, \ldots, q_{\sigma_{j_m}}$, guaranteeing that the predicate q_p has "the appropriate arity." Consider, for example, the case of a unary predicate p of sort σ_i. In that case, ϕ_p says that

[2] Note that if there were infinitely many sort symbols in Σ, then we could not define the $\widehat{\Sigma}$-sentence ϕ in this way.

$$\forall_\sigma y(q_p(y) \to q_{\sigma_i}(y)),$$

which means that nothing outside the subdomain q_{σ_i} satisfies q_p. Note, however, that here we have made a conventional choice. We could just as well have stipulated that q_p applies to *everything* outside of the subdomain q_{σ_i}. All that matters here is that q_p is trivial (either trivially true or trivially false) except on the subdomain of objects satisfying q_{σ_i}.

For each function symbol $f \in \Sigma$ of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m} \to \sigma_j$, we define the $\widehat{\Sigma}$-sentence ϕ_f to be the conjunction

$$\forall_\sigma y_1 \ldots \forall_\sigma y_m \forall_\sigma y \big(q_f(y_1, \ldots, y_m, y) \to (q_{\sigma_{j_1}}(y_1) \wedge \ldots \wedge q_{\sigma_{j_m}}(y_m) \wedge q_{\sigma_j}(y))\big)$$
$$\wedge \forall_\sigma y_1 \ldots \forall_\sigma y_m \big((q_{\sigma_{j_1}}(y_1) \wedge \ldots \wedge q_{\sigma_{j_m}}(y_m)) \to \exists_{\sigma=1} y(q_f(y_1, \ldots, y_m, y))\big).$$

The first conjunct guarantees that the symbol q_f has "the appropriate arity," and the second conjunct guarantees that q_f behaves like a function. Lastly, if $c \in \Sigma$ is a constant symbol of arity σ_j, then we define the $\widehat{\Sigma}$-sentence ϕ_c to be $q_{\sigma_j}(d_c)$. This sentence guarantees that the constant symbol d_c also has "the appropriate arity."

We now have the resources to define a $\widehat{\Sigma}$-theory $\widehat{T}$ that is Morita equivalent to T:

$$\widehat{T} = \{\widehat{\alpha} : \alpha \in T\} \cup \{\phi, \phi_{\sigma_1}, \ldots, \phi_{\sigma_n}\} \cup \{\phi_p : p \in \Sigma\}$$
$$\cup \{\phi_f : f \in \Sigma\} \cup \{\phi_c : c \in \Sigma\}.$$

The theory $\widehat{T}$ has two kinds of axioms, the translated axioms of T and the auxiliary axioms. These axioms allow $\widehat{T}$ to imitate the theory T in the signature $\widehat{\Sigma}$. Indeed, one can prove the following result.

THEOREM 5.3.1 (Barrett) *The theories T and $\widehat{T}$ are Morita equivalent.*

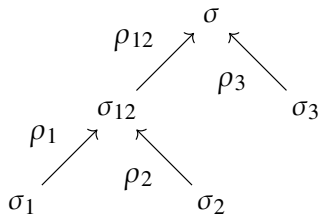
The proof of Theorem 5.3.1 requires some work, but the idea behind it is simple. The theory T needs to define symbols in $\widehat{\Sigma}$. It defines the sort symbol σ as a "universal sort" by taking the coproduct of the sorts $\sigma_1, \ldots, \sigma_n \in \Sigma$. The theory T then defines the symbols q_p, q_f, and d_c in $\widehat{\Sigma}$ simply by using the corresponding symbols p, f, and c in Σ. Likewise, the theory $\widehat{T}$ needs to define the symbols in Σ. It defines the sort symbol σ_j as the subsort of "things that are q_{σ_j}" for each $j = 1, \ldots, n$. And $\widehat{T}$ defines the symbols p, f, and c again by using the corresponding symbols q_p, q_f, and d_c.

We now proceed to the gory details. We prove a special case of the result for convenience. We will assume that Σ has only three sort symbols $\sigma_1, \sigma_2, \sigma_3$ and that Σ does not contain function or constant symbols. A perfectly analogous (though more tedious) proof goes through in the general case.

We prove the result by explicitly constructing a "common Morita extension" $T_4 \cong \widehat{T}_4$ of T and $\widehat{T}$ to the following signature:

$$\Sigma^+ = \Sigma \cup \widehat{\Sigma} \cup \{\sigma_{12}\} \cup \{\rho_1, \rho_2, \rho_{12}, \rho_3\} \cup \{i_1, i_2, i_3\}.$$

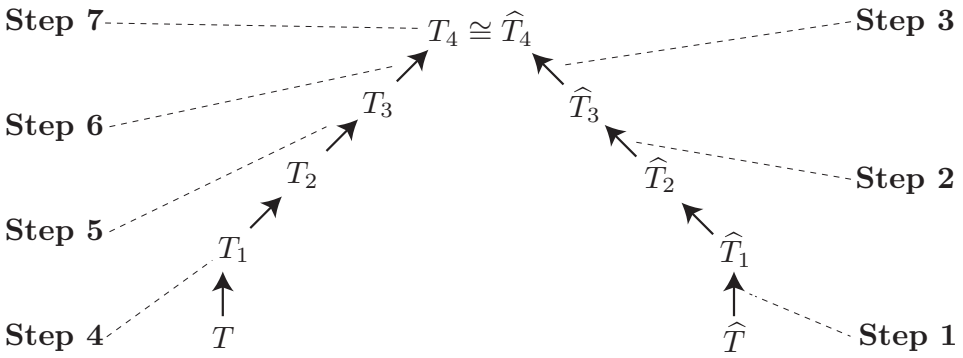
The symbol $\sigma_{12} \in \Sigma^+$ is a sort symbol. The symbols denoted by subscripted ρ are function symbols. Their arities are expressed in the following figure.



The symbols i_1, i_2, and i_3 are function symbols with arity $\sigma_1 \to \sigma$, $\sigma_2 \to \sigma$, and $\sigma_3 \to \sigma$, respectively.

We now turn to the proof.

Proof of Theorem 5.3.1 The following figure illustrates how our proof will be organized.



Steps 1–3 define the theories $\widehat{T}_1, \ldots, \widehat{T}_4$, Steps 4–6 define $T_1, \ldots, T_4$, and Step 7 shows that T_4 and $\widehat{T}_4$ are logically equivalent.

Step 1. We begin by defining the theory $\widehat{T}_1$. For each sort $\sigma_j \in \Sigma$ we consider the following sentence.

$$\begin{aligned} &\forall_\sigma y \left(q_{\sigma_j}(y) \leftrightarrow \exists_{\sigma_j} x(i_j(x) = y)\right) \\ &\qquad \wedge \forall_{\sigma_j} x_1 \forall_{\sigma_j} x_2 (i_j(x_1) = i_j(x_2) \to x_1 = x_2) \end{aligned} \tag{θ_{σ_j}}$$

The sentence θ_{σ_j} defines the symbols σ_j and i_j as the subsort of "things that are q_{σ_j}." The auxiliary axioms ϕ_{σ_j} of $\widehat{T}$ guarantee that the admissibility conditions for these definitions are satisfied. The theory $\widehat{T}_1 = \widehat{T} \cup \{\theta_{\sigma_1}, \theta_{\sigma_2}, \theta_{\sigma_3}\}$ is therefore a Morita extension of $\widehat{T}$ to the signature $\widehat{\Sigma} \cup \{\sigma_1, \sigma_2, \sigma_3, i_1, i_2, i_3\}$.

Step 2. We now define the theories $\widehat{T}_2$ and $\widehat{T}_3$. Let $\theta_{\sigma_{12}}$ be a sentence that defines the symbols $\sigma_{12}, \rho_1, \rho_2$ as a coproduct sort. The theory $\widehat{T}_2 = \widehat{T}_1 \cup \{\theta_{\sigma_{12}}\}$ is clearly a Morita extension of $\widehat{T}_1$. We have yet to define the function symbols ρ_{12} and ρ_3. The following two sentences define these symbols.

$$\forall_{\sigma_3} x \forall_\sigma y (\rho_3(x) = y \leftrightarrow i_3(x) = y) \tag{θ_{ρ_3}}$$

$$\forall_{\sigma_{12}} x \forall_\sigma y (\rho_{12}(x) = y \leftrightarrow \psi(x, y)) \tag{$\theta_{\rho_{12}}$}$$

The sentence θ_{ρ_3} simply defines ρ_3 to be equal to the function i_3. For the sentence $\theta_{\rho_{12}}$, we define the formula $\psi(x, y)$ to be

$$\exists_{\sigma_1} z_1 \big(\rho_1(z_1) = x \wedge i_1(z_1) = y\big) \vee \exists_{\sigma_2} z_2 \big(\rho_2(z_2) = x \wedge i_2(z_2) = y\big).$$

We should take a moment here to understand the definition $\theta_{\rho_{12}}$. We want to define what the function ρ_{12} does to an element a of sort σ_{12}. Since the sort σ_{12} is the coproduct of the sorts σ_1 and σ_2, the element a must "actually be" of one of the sorts σ_1 or σ_2. (The disjuncts in the formula $\psi(x, y)$ correspond to these possibilities.) The definition $\theta_{\rho_{12}}$ stipulates that if a is "actually" of sort σ_j, then the value of ρ_{12} at a is the same as the value of i_j at a. One can verify that $\widehat{T}_2$ satisfies the admissibility conditions for θ_{ρ_3} and $\theta_{\rho_{12}}$, so the theory $\widehat{T}_3 = \widehat{T}_2 \cup \{\theta_{\rho_3}, \theta_{\rho_{12}}\}$ is a Morita extension of $\widehat{T}_2$ to the signature

$$\widehat{\Sigma} \cup \{\sigma_1, \sigma_2, \sigma_3, \sigma_{12}, i_1, i_2, i_3, \rho_1, \rho_2, \rho_3, \rho_{12}\}.$$

Step 3. We now describe the Σ^+-theory $\widehat{T}_4$. This theory defines the predicates in the signature Σ. Let $p \in \Sigma$ be a predicate symbol of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m}$. We consider the following sentence.

$$\forall_{\sigma_{j_1}} x_1 \ldots \forall_{\sigma_{j_m}} x_m \left(p(x_1, \ldots, x_m) \leftrightarrow q_p(i_{j_1}(x_1), \ldots, i_{j_m}(x_m)) \right). \qquad (\theta_p)$$

The theory $\widehat{T}_4 = \widehat{T}_3 \cup \{\theta_p : p \in \Sigma\}$ is therefore a Morita extension of $\widehat{T}_3$ to the signature Σ^+.

Step 4. We turn to the left-hand side of our organizational figure and define the theories T_1 and T_2. We proceed in an analogous manner to the first part of Step 2. The theory $T_1 = T \cup \{\theta_{\sigma_{12}}\}$ is a Morita extension of T to the signature $\Sigma \cup \{\sigma_{12}, \rho_1, \rho_2\}$. Now let θ_σ be the sentence that defines the symbols $\sigma, \rho_{12}, \rho_3$ as a coproduct sort. The theory $T_2 = T_1 \cup \{\theta_\sigma\}$ is a Morita extension of T_1 to the signature $\Sigma \cup \{\sigma_{12}, \sigma, \rho_1, \rho_2, \rho_3, \rho_{12}\}$.

Step 5. This step defines the function symbols i_1, i_2, and i_3. We consider the following sentences.

$$\forall_{\sigma_3} x_3 \forall_\sigma y (i_3(x_3) = y \leftrightarrow \rho_3(x_3) = y) \qquad (\theta_{i_3})$$

$$\forall_{\sigma_2} x_2 \forall_\sigma y \big(i_2(x_2) = y \leftrightarrow \exists_{\sigma_{12}} z(\rho_2(x_2) = z \wedge \rho_{12}(z) = y)\big) \qquad (\theta_{i_2})$$

$$\forall_{\sigma_1} x_1 \forall_\sigma y \big(i_1(x_1) = y \leftrightarrow \exists_{\sigma_{12}} z(\rho_1(x_1) = z \wedge \rho_{12}(z) = y)\big) \qquad (\theta_{i_1})$$

The sentence θ_{i_3} defines the function symbol i_3 to be equal to ρ_3. The sentence θ_{i_2} defines the function symbol i_2 to be equal to the composition "$\rho_{12} \circ \rho_2$." Likewise, the sentence θ_{i_1} defines the function symbol i_1 to be "$\rho_{12} \circ \rho_1$." The theory $T_3 = T_2 \cup \{\theta_{i_1}, \theta_{i_2}, \theta_{i_3}\}$ is a Morita extension of T_2 to the signature $\Sigma \cup \{\sigma_{12}, \sigma, \rho_1, \rho_2, \rho_3, \rho_{12}, i_1, i_2, i_3\}$.

Step 6. We still need to define the predicate symbols in $\widehat{\Sigma}$. Let $\sigma_j \in \Sigma$ be a sort symbol and $p \in \Sigma$ a predicate symbol of arity $\sigma_{j_1} \times \ldots \times \sigma_{j_m}$. We consider the following sentences.

$$\forall_\sigma y (q_{\sigma_j}(y) \leftrightarrow \exists_{\sigma_j} x (i_j(x) = y)) \qquad (\theta_{q_{\sigma_j}})$$

$$\forall_\sigma y_1 \ldots \forall_\sigma y_m \big(q_p(y_1, \ldots, y_m) \leftrightarrow \exists_{\sigma_{j_1}} x_1 \ldots \exists_{\sigma_{j_m}} x_m (i_{j_1}(x_1) = y_1 \wedge \ldots \wedge i_{j_m}(x_m) = y_m \wedge p(x_1, \ldots, x_m))\big) \qquad (\theta_{q_p})$$

These sentences define the predicates $q_{\sigma_j} \in \widehat{\Sigma}$ and $q_p \in \widehat{\Sigma}$. One can verify that T_3 satisfies the admissibility conditions for the definitions $\theta_{q_{\sigma_j}}$. And, therefore, the theory

$T_4 = T_3 \cup \{\theta_{q_{\sigma_1}}, \theta_{q_{\sigma_2}}, \theta_{q_{\sigma_3}}\} \cup \{\theta_{q_p} : p \in \Sigma\}$ is a Morita extension of T_3 to the signature Σ^+.

Step 7. It only remains to show that the Σ^+-theories T_4 and $\widehat{T}_4$ are logically equivalent. One can verify by induction on the complexity of ψ that

$$T_4 \vdash \psi \leftrightarrow \widehat{\psi} \quad \text{and} \quad \widehat{T}_4 \vdash \psi \leftrightarrow \widehat{\psi}. \tag{5.3}$$

for every Σ-sentence ψ. One then uses (5.3) to show that T_4 and $\widehat{T}_4$ are logically equivalent. The argument involves a number of cases, but since each case is straightforward, we leave them to the reader to verify. The theories T_4 and $\widehat{T}_4$ are logically equivalent, which implies that T and $\widehat{T}$ are Morita equivalent. □

Theorem 5.3.1 validates Quine's claim that every many-sorted theory can be converted to a single-sorted theory. He concluded that many-sorted logic is dispensable. Whether Quine was right or wrong, his claims in this regard are probably the reason why many-sorted logic hasn't been part of the standard curriculum for analytic philosophers. We hope that our efforts here go some way toward remedying this unfortunate situation.

5.4 Translation Generalized

In the previous chapters, we've talked about various notions of a "translation" between theories. Of course, we did not find the definition of translation written on tablets of stone; nor did we have a Platonic vision of the one true form of a translation. No, we found Quine's definition in the literature, and it works quite well for some purposes, but it's also quite restrictive. In particular, Quine's notions of reconstrual and translation are not general enough to capture some well-known cases of translations between the theories of pure mathematics.

1. In the nineteenth century, the German mathematician Leopold Kronecker is reported to have said, "God made the integers, all else is the work of man." In more prosaic terms, talk about higher number systems – such as rational, real, and complex numbers – can be *reduced* to talk about integers. However, to effect such a reduction, one must treat each rational number as a pair of integers – or, more accurately, as an equivalence class of pairs of integers. Similarly, to reduce the complex numbers to the real numbers, one must treat a complex number as a pair of real numbers, viz. the real and imaginary parts of the complex number.
2. Now for a more controversial example, which we will take up at greater length in Section 7.4. There are different ways that one can write down axioms for Euclidean geometry. In one axiomatization, the basic objects are points; and in another axiomatization, the basic objects are lines. Is there a sense in which these two axiomatized theories could both be Euclidean geometry – in particular, that they could be equivalent? The answer is yes, but only if one allows translations that take a single variable of the first theory to a pair of variables of the second theory. In particular, a line needs to be treated as an equivalence class of pairs of points, and a point needs to be treated as a pair of intersecting lines.

In the previous chapter, we required that a formula $p(x)$ of Σ be translated to a formula $\phi(x)$ of Σ'. There's one particular part of this recipe that seems questionable: why would the same variable x occur in both formulas? In general, why suppose that two signatures Σ and Σ' should share the same variables in common? It's not like variables have some "trans-theoretical" meaning that must be preserved by any reasonable translation.

But how then can variables be reconstrued in moving from one theory to another? One natural proposal would be to include in a reconstrual a mapping from variables of Σ to variables of Σ' – i.e., a function that assigns a variable of Σ' to each variable of Σ. Even so, it's a nontrivial question whether there is an in-principle reason that a single variable in Σ must be reconstrued as a single variable in Σ'. Perhaps one theorist uses several variables to do the work that the other theorist manages to do with a single variable. Such cases are not hard to find in the sciences – for example, when the objects of one mathematical theory are reconstrued as "logical constructions" of objects in another mathematical theory.

Let's proceed then under the assumption that a single variable in one language could be reconstrued in terms of multiple variables in another language. Thus, a reconstrual, in the formal sense, should include a function that matches variables of the signature Σ to n-tuples of variables of the signature Σ'.

Consider again the case of reconstruing rational numbers (i.e., fractions) as pairs of integers. Of course, not *every* pair of integers gives a well-defined fraction. For example, there is no fraction of the form $\frac{1}{0}$. In that case, the "integer theorist" doesn't think of the domain of fractions as consisting of all pairs of integers; rather, she thinks of that domain as consisting of pairs of integers where the second entry is nonzero. To capture this nuance – the restriction of the domain of quantification – we stipulate that a reconstrual F includes a formula D of the target language Σ'. In the running example, the formula D could be given by

$$D(x, y) \equiv (x = x) \wedge (y \neq 0).$$

The integer theorist can then use the formula D to restrict her quantifiers to the domain of well-defined fractions.

Finally, and most controversially, let's consider how we might reconstrue the equality relation $=$ of the domain theory T as a relation of the target theory T'. (Our choice here will prove to be controversial when we show that it yields a positive verdict in favor of quantifier variance. See Example 5.4.16.) Recall that the single variables x and y will typically be reconstrued as n-tuples of variables $\vec{x}$ and $\vec{y}$. In that case, how should we reconstrue the formula $x = y$? One might naturally propose that $x = y$ be reconstrued as the formula

$$(x_1 = y_1) \wedge (x_2 = y_2) \wedge \cdots \wedge (x_n = y_n). \tag{5.4}$$

But here we need to think a bit harder about how and why variables of Σ are encoded as variables of Σ'. For this, let's consider again the example of rational numbers being reduced to integers.

Consider a formula $x = y$ in the theory of rational numbers. To the "integer theorist," the variables x and y really represent complex entities, namely fractions. What's more, to say that two fractions $\frac{x_1}{x_2}$ and $\frac{y_1}{y_2}$ are equal does not mean that $x_1 = y_1$ and $x_2 = y_2$. Rather, $\frac{x_1}{x_2} = \frac{y_1}{y_2}$ means that $x_1 \times y_2 = y_1 \times x_2$. In other words, the formula $x = y$ of the language of the rational numbers is reconstrued as the formula

$$x_1 \times y_2 = y_1 \times x_2, \tag{5.5}$$

in the language of the integers, where $\times$ is the multiplication operation.

This example suggests that we might not always want the formula $x = y$ to be reconstrued as Eqn. 5.4. Instead, we might prefer to reconstrue $x = y$ as some other Σ' formula $E(x_1, \ldots, x_n; y_1, \ldots, y_n)$. Of course, not everything goes: E will need to perform the same functions in the theory T' that the formula $x = y$ performs in the theory T. In particular, we will require that E be an equivalence relation relative to the theory T'.

We're now ready to consider ways in which the elements of one signature Σ can be reconstrued as syntactic structures built from a second signature Σ'. (We include here the case where $\Sigma' = \Sigma$. In that case, we will be considering substitutions and permutations of notation.) The case of relation symbols is relatively easy: an m-ary relation symbol r of Σ should correspond to a formula $F(r)$ of Σ' with mn free variables. To be even more precise, it's the relation symbol r and an n-tuple of variables $x_1, \ldots, x_n$ that corresponds to some particular formula $F(r)$ of Σ', and we require that $FV(F(r)) = \{\vec{x}_1, \ldots, \vec{x}_n\}$.

We will need to proceed with more caution for the function symbols in the signature Σ. The question at issue is: which syntactic structures over Σ' are the proper targets for a reconstrual of the function symbols in Σ? To say that the target must be another function symbol is too restrictive. Indeed, there's a well-known "theorem" that says that every first-order theory is equivalent to a theory that uses only relation symbols. (The reason that "theorem" is placed in quotes here is because the result cannot be proven with mathematical rigor until the word "equivalent" is defined with mathematical rigor.) The trick to proving that theorem is to reconstrue each function symbol f as a relation

$$p_f(x_1, \ldots, x_m, y) \equiv (f(x_1, \ldots, x_m) = y)$$

and then to add axioms saying that p_f relates each m-tuple $x_1, \ldots, x_m$ to a unique output y. If we are to be able to validate such a result (which is intuitively correct), then we ought to permit function symbols of Σ to be reconstrued as formulas of Σ'. We will deal with this issue by analogy with the way we dealt with relation symbols earlier: a function symbol f of Σ and $m + 1$ variables $x_1, \ldots, x_n, y$ of Σ ought to correspond to a formula $(Ff)(\vec{x}_1, \ldots, \vec{x}_n, \vec{y})$ of Σ'.

In order to define a more general notion of a translation, the key is to allow a single sort σ of Σ to be mapped to a sequence of sorts of Σ', including the case of repetitions of a single sort. The idea, in short, is to encode a single variable (or quantifier) in Σ by means of several variables (or quantifiers) in Σ'. In order to make this idea clearer, it will help to give a precise definition of the monoid of finite sequences from a set S.

DEFINITION 5.4.1 For a set S, we let S^* denote the **free monoid** on S, which is uniquely defined by the following universal property: there is a function $\eta_S : S \to S^*$, and for any monoid A, and function $f : S \to A$, there is a unique monoid morphism $f^* : S^* \to A$ such that $f^* \circ \eta_S = f$. Concretely speaking, S^* can be constructed as the set

$$S \amalg (S \times S) \amalg (S \times S \times S) \amalg \cdots,$$

where $\eta_S : S \to S^*$ is the first coprojection. In this case, given $f : S \to A$, $f^* : S^* \to A$ is the function

$$f^*(s_1, \ldots, s_n) = f(s_1) \circ \cdots \circ f(s_n),$$

where $\circ$ is the monoid operation on A.

DEFINITION 5.4.2 Let Σ and Σ' be many-sorted signatures with sets of sorts S and S' respectively. A generalized **reconstrual** $F : \Sigma \to \Sigma'$ consists of the following:

1. A function $F : S \to (S')^*$. That is, F maps the sorts of Σ to nonempty sequences of sorts of S'. For each $\sigma \in S$, let $d(\sigma)$ be the length of the sequence $F(\sigma)$. We call $d : S \to \mathbb{N}$ the **dimension function** of F.
2. A corresponding function $x \mapsto \vec{x} = x_1, \ldots, x_{d(\sigma)}$ from Σ-variables to sequences of Σ'-variables, such that $x_i : F(\sigma)_i$. We require that if $x \not\equiv y$, then the sequences $\vec{x}$ and $\vec{y}$ have no overlap.
3. A function D from Σ-variables to Σ'-formulas. We call D_x a **domain formula**. We require the map $x \mapsto D_x$ to be natural in the following sense: if y is of the same sort as x, then $D_y = D_x[\vec{y}/\vec{x}]$.
4. A function F that takes a relation symbol p of Σ, and a suitable context $x_1, \ldots, x_n$ of variables from Σ, and yields a formula $(Fp)(\vec{x}_1, \ldots, \vec{x}_n)$ of Σ'. We again require this map to be natural in the sense that

$$(Fp)(\vec{y}_1, \ldots, \vec{y}_n) = (Fp)(\vec{x}_1, \ldots, \vec{x}_n)[\vec{y}_1, \ldots, \vec{y}_n/\vec{x}_1, \ldots, \vec{x}_n].$$

A reconstrual F naturally extends to a map from Σ-formulas to Σ'-formulas. We define this extension, also called F, so that for any Σ-formula ϕ, with x free in ϕ, the following two constraints are satisfied:

$$F(\phi) \vdash D(\vec{x}), \qquad F(\phi[y/x]) = F(\phi)[\vec{y}/\vec{x}].$$

The first restriction is not technically necessary – it is simply a convenient way to ignore whatever the formula $F(\phi)$ says about things outside of the domain $D(\vec{x})$. (This apparently minor issue plays a significant role in Quine's argument for the dispensability of many-sorted logic. See 5.4.17.) Accordingly, for a relation symbol p of Σ, we first redefine $(Fp)(\vec{x}_1, \ldots, \vec{x}_n)$ by conjoining with $D(\vec{x}_1) \wedge \cdots \wedge D(\vec{x}_n)$. (We could have also have included this condition in the very definition of a reconstrual.) The extension of F proceeds as follows:

- Let $F(\phi \wedge \psi) = F(\phi) \wedge F(\psi)$, and let $F(\phi \vee \psi) = F(\phi) \vee F(\psi)$.
- Let $F(\neg\phi) = \neg F(\phi) \wedge D(\vec{x}_1) \wedge \cdots \wedge D(\vec{x}_n)$, where $x_1, \ldots, x_n$ are all the free variables that occur in ϕ.
- Let $F(\phi \rightarrow \psi) = F(\neg\phi) \vee F(\psi)$.
- Let $F(\exists x\phi) = \exists\vec{x}(D(\vec{x}) \wedge F(\phi))$.
- Let $F(\forall x\phi) = \forall\vec{x}(D(\vec{x}) \rightarrow F(\phi))$.

DEFINITION 5.4.3 Let $F : \Sigma \rightarrow \Sigma'$ be a reconstrual. We say that F is a **translation** of T into T' just in case, for every Σ-sentence ϕ, if $T \vdash \phi$ then $T' \vdash F(\phi)$. In this case, we write $F : T \rightarrow T'$. In the case that Σ has a single sort σ, we say that F is a $d(\sigma)$-dimensional translation.

The definition of a translation allows us to handle the case where the domain signature Σ has equality relations and function symbols. In particular, for each theory T in Σ, we explicitly include the following axioms:

- The equality introduction axioms: $\vdash x =_\sigma x$.
- The equality elimination axioms: $\phi(x), (x =_\sigma y) \vdash \phi(y)$, for each atomic or negated atomic formula ϕ of Σ.

As usual, these axioms together entail that $=_\sigma$ is an equivalence relation. Thus, if $F : T \rightarrow T'$ is a translation, then $F(=_\sigma)(\vec{x}, \vec{y})$ is an equivalence relation on domain $D(\vec{x})$. We abbreviate this relation by $E_\sigma(\vec{x}, \vec{y})$ or, when no confusion can result, simply as $E(\vec{x}, \vec{y})$. In this case, for each relation symbol p of Σ,

$$T', (Fp)(\vec{x}), E(\vec{x}, \vec{y}) \vdash (Fp)(\vec{y}).$$

Roughly speaking, the predicate Fp has to be compatible with the equivalence relation E: it holds of something iff it holds of everything E-equivalent to that thing. Equivalently, the extension of Fp is a union of E-equivalence classes.

Now suppose that Σ contains a constant symbol c. Then, choosing a variable x of the same sort, $c = x$ is a unary formula, and $F(c = x)$ is a formula $\phi(\vec{x})$. The theory T entails that the formula $c = x$ is uniquely satisfied. Hence, if $F : T \rightarrow T'$ is a translation, then T' entails that $\phi(\vec{x})$ is uniquely satisfied – relative to the equivalence relation E. In short, T' implies both $\exists\vec{x}(D_x \wedge \phi(\vec{x}))$ and $\phi(\vec{x}) \wedge \phi(\vec{y}) \rightarrow E(\vec{x}, \vec{y})$. Intuitively speaking, this means that the extension of $\phi(\vec{x})$ is a single E-equivalence class.

Similar reasoning applies to the case of any function symbol f of Σ. The Σ-formula $f(x_1, \ldots, x_n) = y$ is reconstrued as some Σ'-formula $\phi(\vec{x}_1, \ldots, \vec{x}_n, \vec{y})$. If $F : T \rightarrow T'$ is a translation, then T' entails that ϕ is a functional relation relative to E-equivalence. What this means intuitively is that ϕ is a function from E-equivalence classes to E-equivalence classes.

Example 5.4.4 (Quantifier variance) We now undertake an extended discussion of an example that is near and dear to metaphysicians: the debate between mereological universalism and nihilism. To keep the technicalities to a bare minimum, we will consider a

dispute over whether the composite of two things exists. Suppose that the parties to the dispute are are named Niels the Nihilist and Mette the Mereological Universalist. Niels says that there are exactly two things, whereas Mette says that there are exactly three things, one of which is composed of the other two.

Now, we press Niels and Mette to regiment their theories, and here's what they come up with. Niels has a signature Σ, which is empty, very much in line with his predilection for desert landscapes. Niels' theory has a single axiom, "there are exactly two things." Mette has a signature Σ' with a binary relation symbol p that she'll use to express the parthood relation. Mette's theory T' says that p is a strict partial order, that there are exactly two atoms, and exactly one thing above those two atoms. Note that Mette can define an open formula in Σ'

$$a(x) \equiv \neg\exists y\, p(y,x),$$

which intuitively expresses the claim that x is an atom.

At the turn of the twenty-first century, metaphysicians were engaged in a fierce debate about whether Niels or Mette has a better theory. Then some other philosophers, such as Eli Hirsch, said, "stop arguing – it's merely a verbal dispute, like an argument about whether there are six roses or half a dozen roses" (see Chalmers et al., 2009; Hirsch, 2011). These other philosophers espouse a position known as **quantifier variance**. One clear explication of quantifier variance would be to say that Niels and Mette's theories are **equivalent**. So are they equivalent or not? The answer to this question depends (unsurprisingly) on the standard of equivalence that we adopt. For example, it is easy to see that Niels and Mette's theories are not strictly intertranslatable in the sense of Defn. 4.5.15. However, we will now see that Niels and Mette's theories are intertranslatable in the weaker sense described in Defn. 5.4.14.

It seems clear that Mette can make sense of Niels' theory – in particular, that she can identify Niels' quantifier as a restriction of her own. The idea that Mette can "make sense of Niels' theory" can be cashed out formally as saying that Niels' theory can be translated into Mette's theory. Intuitively speaking, for any sentence ϕ asserted by Niels, there is a corresponding sentence ϕ^* asserted by Mette. For example, when Niels says,

There are exactly two things,

Mette can charitably interpret him as saying,

There are exactly two atoms.

Now we show that there is indeed a translation $F : T \to T'$, where T is Niels' theory, and T' is Mette's theory. Here Niels and Mette's theories are single-sorted, and we define F to be a one-dimensional reconstrual. We define the domain formula as $D_F(x) = a(x)$, and we translate Niels' equality relation as Mette's equality relation restricted to D_F.

Let's just check that F is indeed a translation. While a general argument is not difficult, let's focus on Niels' controversial claim ϕ: that there are *at most two* objects in the domain:

$$\phi \equiv \forall x \forall y \forall z((x = y) \vee (x = z) \vee (y = z)).$$

The reconstrual F takes $x = y$ to the formula $a(x') \wedge a(y') \wedge (x' = y')$, and hence $F(\phi)$ is the uncontroversially true statement that there are at most two atoms. Of course, Mette agrees with that claim, and so $F : T \rightarrow T'$ is a translation of Niels' theory into Mette's.

Indeed, F is a particularly nice translation: it's conservative, in the sense that if $T' \vdash F(\phi)$, then $T \vdash \phi$. Thus, not only does Mette affirm everything that Niels says about atoms; Niels also affirms everything that Mette says about atoms. Thus, there is a precise sense in which Niels' theory is simply a "sub-theory" of Mette's theory. They are in complete agreement relative to their shared language, and Mette simply has a larger vocabulary than Niels.

The existence of the translation $F : T \rightarrow T'$ comes as no surprise. But what about the other way around? Can Niels be as charitable to Mette as she has been to him? Can he find a way to affirm *everything* that she says? The answer to that question is far from clear. For example, Mette says things like, "x is a composite of y and z." How in the world could Niels make sense of that claim? How in the world could Niels say, "what Mette says here is perfectly correct, if only understood in the proper way"? Similarly, Mette says that "there are more than two things." How in the world could Niels validate such a claim?

We will now see that Niels can indeed charitably interpret, and endorse, all of Mette's assertions. Indeed, Niels needs only think of Mette's notion of "a thing" as corresponding to what he means by "a pair of things" – as long as two pairs are considered to be "the same" when they are permutations of each other.

More precisely, consider a two-dimensional reconstrual $G : \Sigma' \rightarrow \Sigma$ that encodes a Σ'-variable x as a pair x_1, x_2 of Σ-variables. Define $D_G(x_1, x_2)$ to be the formula $(x_1 = x_1) \wedge (x_2 = x_2)$ that holds for all pairs $\langle x_1, x_2 \rangle$. Define $E_G(x_1, x_2, y_1, y_2)$ to be the relation that holds between $\langle x_1, x_2 \rangle$ and $\langle y_1, y_2 \rangle$ just in case one is a permutation of the other. That is,

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

Clearly, T entails that E_G is an equivalence relation.

The signature Σ' consists of a single binary relation symbol p. Since G is two-dimensional, Gp must be defined to be a four-place relation in Σ. Here is the intuitive idea behind our definition of Gp: we will simulate atoms of Mette's theory by means of diagonal pairs, i.e., pairs of the form $\langle x, x \rangle$. We then say that Gp holds precisely between pairs when the first is diagonal, the second is not, and the first has a term in common with the second. More precisely,

$$(Gp)(x_1, x_2, y_1, y_2) \equiv (x_1 = x_2) \wedge (y_1 \neq y_2) \wedge (x_1 = y_1 \vee x_1 = y_2).$$

Recall that $a(x)$ is the formula of Σ' that says that x is an atom. We claim now that the translation $G(a(x))$ of $a(x)$ holds precisely for the pairs on the diagonal. That is,

$$T \vdash G(a)(x_1, x_2) \leftrightarrow (x_1 = x_2).$$

We argue by reductio ad absurdum. (Here we use the notion of a model, which will first be introduced in the next chapter. Hopefully, the intuition will be clear.) First, if

$$T \nvdash G(a)(x_1, x_2) \rightarrow (x_1 = x_2),$$

then there is a model M of T, and two distinct objects c, d of M such that $M \models G(a)(c,d)$. That means that

$$M \models \neg\exists y \exists z\, G(p)(y,z,c,d).$$

But, clearly, $M \models G(p)(c,c,c,d)$, a contradiction. To prove the other direction, it will suffice to show that for any model M of T, and for any $c \in M$, we have $M \models G(a)(c,c)$. Recalling that T only has one model, namely a model with two objects, the result easily follows.

When Mette the Mereologist says that there are more than two things, Niels the Nihilist understands her as saying that there are more than two pairs of things. Of course, Niels agrees with that claim. In fact, it's not hard to see that, under this interpretation, Niels affirms everything that Mette says. ⌟

DISCUSSION 5.4.5 We've shown that Niels' theory can be translated into Mette's, and vice versa. Granting that this is a good notion of "translation," does it follow that these two theories are equivalent? In short, no. Recall the simpler case of propositional theories. For example, let $\Sigma = \{p_0, p_1, \ldots\}$, let T be the empty theory in Σ, and let T' be the theory with axioms $p_0 \vdash p_1, p_0 \vdash p_2, \ldots$ Then there are translations $f : T \to T'$ and $g : T' \to T$, but T and T' are not equivalent theories. In general, mutual interpretability is not sufficient for equivalence. Nonetheless, we will soon see (Example 5.4.16) that there is a precise sense in which Niels' and Mette's theories are indeed equivalent. □

We are now ready to prove a generalized version of the **substitution theorem**. In its simplest form, the substitution theorem says a valid derivation $\phi_1, \ldots, \phi_n \vdash \psi$ is preserved under uniform substitution of the non-logical symbols in $\phi_1, \ldots, \phi_n$ and ψ. For example, from a valid derivation of $\exists x(p(x) \wedge q(x)) \vdash \exists x p(x)$, substitution of $\forall y r(y,z)$ for $p(x)$ yields a valid derivation of

$$\exists z(\forall y r(y,z) \wedge q(z)) \vdash \exists z \forall y r(y,z).$$

However, we need to be careful in describing what counts as a legitimate "substitution instance" of a formula. Let's test our intuitions against an example.

Example 5.4.6 Let Σ be a single-sorted signature with equality, but no other symbols. Let Σ' be a single-sorted signature with equality, and one other monadic predicate $D(x)$. We define a one-dimensional reconstrual $F : \Sigma \to \Sigma'$ by taking $D(x)$ to be the domain formula, and by taking $E(x,y)$ to be equality in Σ'. We will see now that the substitution theorem does *not* hold in the form: if $\phi \vdash \psi$ then $F(\phi) \vdash F(\psi)$.

In Σ, we have $x \neq y \vdash \exists z(x \neq z)$. Since F translates equality in Σ to equality in Σ', we have $F(x \neq y) \equiv (x \neq y)$. Furthermore, $F(\exists z(x \neq z))$ is the relativized formula $\exists z(D(z) \wedge x \neq z)$. But $x \neq y$ does not imply that there is a z such that $D(z)$ and $x \neq z$. For example, in the domain $\{a,b\}$, if the extension of D is $\{a\}$, then $a \neq b$, but not $\exists z(D(z) \wedge a \neq z)$. Thus, the substitution theorem does *not* hold in the form: if $\phi \vdash \psi$, then $F(\phi) \vdash F(\psi)$. So what's the problem here?

To speak figuratively, a reconstrual F maps a variable x of Σ to variables $\vec{x}$ that are relativized to the domain $D(\vec{x})$. However, the turnstile $\vdash$ for Σ' is not relativized in this fashion: a sequent $F(\phi) \vdash F(\psi)$ corresponds to a tautology $\vdash \forall\vec{x}(F(\phi) \to F(\psi))$. It wouldn't make sense to expect this last statement to hold, since the intention is for the variables in $F(\phi)$ and $F(\psi)$ to range over $D(\vec{x})$. Thus, the relevant question is whether $F(\phi) \vdash_{D(\vec{x})} F(\psi)$, where the latter is shorthand for

$$\vdash \forall\vec{x}(D(\vec{x}) \to (F(\phi) \to F(\psi))).$$

In the current example, then, the question is whether the following holds:

$$F(x \neq y), D(x), D(y) \vdash F(\exists y(x \neq y)).$$

And it obviously does. This example shows us how to formulate a substitution theorem for generalized reconstruals such as F. ⌟

THEOREM 5.4.7 (Substitution) *Let Σ be a signature without function symbols, and suppose that F is a reconstrual from Σ to Σ'. Then, for any formulas ϕ and ψ with free variables $x_1, \ldots, x_n$, if $\phi \vdash \psi$, then $F(\phi) \vdash_{D(\vec{x}_1,\ldots,\vec{x}_n)} F(\psi)$. In particular, if ϕ and ψ are Σ-sentences, then $F(\phi) \vdash F(\psi)$.*

Proof We will prove this result by induction on the construction of proofs. For the base case, the rule of assumptions justifies not only $\phi \vdash \phi$, but also $F(\phi) \vdash F(\phi)$, and hence $D(\vec{x}), F(\phi) \vdash F(\phi)$. The inductive cases for the Boolean connectives involve no special complications, and so we leave them to the reader.

Consider now the case of $\exists$-elim. Suppose that $\exists y\phi \vdash \psi$ results from application of $\exists$-elim to $\phi \vdash \psi$, in which case y is not free in ψ. We rewrite ϕ and ψ in the suggestive notation $\phi(x, y)$ and $\psi(x)$, indicating that x may be free in both ϕ and ψ, and that $y \not\equiv x$. (Note, however, that our argument doesn't depend on ϕ and ψ sharing exactly one free variable in common.) We want to show that $F(\exists y\phi(x, y)) \vdash_{D(\vec{x})} F(\psi(x))$, which expands to

$$D(\vec{x}), \exists\vec{y}(D(\vec{y}) \wedge F(\phi(x, y))) \vdash F(\psi(x)). \tag{5.6}$$

The inductive hypothesis here says that

$$D(\vec{x}), D(\vec{y}), F(\phi(x, y)) \vdash F(\psi(x)).$$

Since x and y are distinct variables, the sequences $\vec{x}$ and $\vec{y}$ have no overlap, and $\vec{y}$ does not occur free in $D(\vec{x})$. Thus, n-applications of $\exists$-intro yield the sequent (5.6).

Consider now the case of $\exists$-intro. Suppose that $\phi \vdash \exists y\psi$ follows from $\phi \vdash \psi$ by an application of $\exists$-intro. Again, we will rewrite the former sequent as

$$\phi(x, y) \vdash \exists y\psi(x, y).$$

We wish to show that

$$D(\vec{x}), D(\vec{y}), F(\phi(x, y)) \vdash \exists\vec{y}F(\psi(x, y)).$$

By the inductive hypothesis, we have

$$D(\vec{x}), D(\vec{y}), F(\phi(x,y)) \vdash F(\psi(x,y)).$$

Thus, the result follows by repeated application of ∃-intro. □

The previous version of the substitution theorem applies only to the case of signatures without function symbols. Intuitively, however, the formal validity of proofs should also be maintained through uniform substitution of terms.

Example 5.4.8 Suppose that $\phi(x) \vdash \psi(x)$, which is equivalent to $\vdash \forall x(\phi(x) \vdash \psi(x))$. Now let $t(\vec{y})$ be a term with free variables $\vec{y} \equiv y_1, \ldots, y_n$, and suppose that each of these variables is free for x in ϕ and ψ, but none of them are themselves free in either one of these formulas. (In the simplest case, these variables simply do not occur in either one of the formulas.) Then ∀-elim and intro yield

$$\vdash \forall \vec{y}(\phi(t(\vec{y})) \to \psi(t(\vec{y}))),$$

which is equivalent to $\phi(t(\vec{y}) \vdash \psi(t(\vec{y}))$. In other words, a valid proof remains valid if a variable x is uniformly replaced by a term $t(\vec{y})$, so long as the relevant restrictions are respected. ⌟

At this stage, we have a definition of a generalized translation, and we've shown that it yields a generalized substitution theorem. What we would like to do now is to look at specific sorts of translations – and most particularly, at which translations should count as giving an equivalence of theories. It turns out, however, that giving a good definition of equivalence is a bit complicated. As many examples will show, it won't suffice to say that a translation $F : T \to T'$ is an equivalence just in case it has an inverse $G : T' \to T$, and not even a quasi-inverse in the sense of 4.5.15. For a good definition of equivalence, we need a notion of a "homotopy" between translations, and we need a notion of the composition of translations. We turn first to the second of these.

DEFINITION 5.4.9 (Composition of reconstruals) Suppose that $F : \Sigma \to \Sigma_1$ and $G : \Sigma_1 \to \Sigma_2$ are reconstruals. Define a reconstrual $H : \Sigma \to \Sigma_2$ as follows:

- Since $G : S_1 \to S_2^*$, there is a unique morphism $G^* : S_1^* \to S_2^*$ such that $G = G^* \circ \eta_{S_1}$. In other words, G^* acts on a sequence of S_1 sorts by applying G to each element and then concatenating. We then define $H = G^* \circ F : S \to S_2^*$.
- We use the same idea to associate each variable x of Σ with a (double) sequence $X_1, \ldots, X_n$ of variables of Σ_2. In short,

$$\begin{aligned} H(x) &= G^*(F(x)) \\ &= X_1, \ldots, X_n \\ &= (x_{11}, \ldots, x_{1m_1}), \ldots, (x_{n1}, \ldots, x_{nm_n}), \end{aligned}$$

where $F(x) = x_1, \ldots, x_n$, and $G(x_i) = X_i = (x_{i1}, \ldots, x_{im_i})$.

- Let $D_F(\vec{x})$ denote the domain formula of F corresponding to the Σ-variable x. Let $D_G(X_i)$ denote the domain formula of G corresponding to the Σ_1-variable x_i. Then we define

$$D_H(X_1, \ldots, X_n) := G(D_F(\vec{x})).$$

  Recall that any free variable in $G(D_F(\vec{x}))$ occurs in the double sequence $X_1, \ldots, X_n$, and that $G(\phi) \vdash D_G(Y)$ if y is free in ϕ. Thus, $D_H(X_1, \ldots, X_n) \vdash D_G(X_i)$ for each $i = 1, \ldots, n$.
- For a relation symbol p of Σ, we define

$$(Hp)(X_1, \ldots, X_n) = G((Fp)(\vec{x}_1, \ldots, \vec{x}_n)).$$

PROPOSITION 5.4.10 *If F and G are translations, then $G \circ F$ is a translation.*

Proof This result follows trivially once we recognize that $G \circ F$ is a legitimate reconstrual. □

For some philosophers, it may seem that we have already greatly overcomplicated matters by using category theory to frame our discussion of theories. I'm sorry to say that matters are worse than that. The collection of theories really has more interesting structure than a category has; in fact, it's most naturally thought of as a **2-category**, where there are 0-cells (objects), 1-cells (arrows), and 2-cells (arrows between arrows). In particular, our 2-category of theories, **Th**, has first-order theories as the 0-cells, and translations as the 1-cells. We now define the 2-cells, which we call **t-maps**.

Let F and G be translations from T to T'. Since the definition of a t-map is heavily syntactic, we begin with an intuitive gloss in the special case where Σ has a single sort σ. In this case $F(\sigma)$ is a sequence $\sigma_1, \ldots, \sigma_m$ of Σ'-sorts, and $G(\sigma)$ is a sequence $\sigma'_1, \ldots, \sigma'_n$ of Σ'-sorts. Then a t-map $\chi : F \Rightarrow G$ consists of a formula $\chi(\vec{x}, \vec{y})$ that links m-tuples to n-tuples. This formula $\chi(\vec{x}, \vec{y})$ should have the following features:

1. The theory T' implies that $\chi(\vec{x}, \vec{y})$ is a functional relation from D_F to D_G, relative to the notion of equality given by the equivalence relations E_F and E_G.
2. For each formula ϕ of Σ, χ maps the extension of $F(\phi)$ into the extension of $G(\phi)$.

We now turn to the details of the definition.

DEFINITION 5.4.11 A **t-map** $\chi : F \Rightarrow G$ is a family of Σ'-formulas $\{\chi_\sigma\}$, where σ runs over the sorts of Σ, where each χ_σ has $d_K(\sigma) + d_L(\sigma)$ free variables, and such that T' entails the following (which we label with suggestive acronyms):

$$\chi_\sigma(\vec{x}, \vec{y}) \rightarrow (D_F(\vec{x}) \wedge D_G(\vec{y})) \qquad \text{(dom-ran)}$$

$$(E_F(\vec{x}, \vec{w}) \wedge E_G(\vec{y}, \vec{z}) \wedge \chi_\sigma(\vec{w}, \vec{z})) \rightarrow \chi_\sigma(\vec{x}, \vec{y}) \qquad \text{(well-def)}$$

$$D_F(\vec{x}) \rightarrow \exists \vec{y}(D_G(\vec{y}) \wedge \chi_\sigma(\vec{x}, \vec{y})) \qquad \text{(exist)}$$

$$(\chi_\sigma(\vec{x}, \vec{y}) \wedge \chi_\sigma(\vec{x}, \vec{z})) \rightarrow E_G(\vec{y}, \vec{z}) \qquad \text{(unique)}$$

Furthermore, for any Σ-formula $\phi(x_1, \ldots, x_n)$ with $x_1 : \sigma_1, \ldots, x_n : \sigma_n$, the theory must T' entail that

$$\chi_{\vec{\sigma}}(X, Y) \to (F(\phi)(X) \to G(\phi)(Y)),$$

where we abbreviate $X = \vec{x}_1, \ldots, \vec{x}_n$, $Y = \vec{y}_1, \ldots, \vec{y}_n$, and $\chi_{\vec{\sigma}}(X, Y) = \chi_{\sigma_1}(\vec{x}_1, \vec{y}_1) \wedge \ldots \wedge \chi_{\sigma_n}(\vec{x}_n, \vec{y}_n)$.

We are especially interested in what it might mean to say that two translations $F : T \to T'$ and $G : T \to T'$ are isomorphic – i.e., the conditions under which a t-map $\chi : F \Rightarrow G$ is an isomorphism.

DEFINITION 5.4.12 We say that a t-map $\chi : F \Rightarrow G$ is a **homotopy** (or an **isomorphism of translations**) if each of the functions χ establishes a bijective correspondence, relative to the equivalence relations E_F and E_G. More precisely, the theory T' entails

$$D_G(\vec{y}) \to \exists \vec{x}(D_F(\vec{x}) \wedge \chi(\vec{x}, \vec{y})) \qquad \text{(onto)}$$

$$(\chi(\vec{x}, \vec{y}) \wedge \chi(\vec{w}, \vec{y})) \to E_F(\vec{x}, \vec{w}) \qquad \text{(one-to-one)}$$

Furthermore, for each formula ϕ of Σ, the theory T' entails that

$$\chi(X, Y) \to (G(\phi)(Y) \to F(\phi)(X)).$$

Here we have omitted the sort symbol σ from χ_σ merely in the interest of notational simplicity.

DISCUSSION 5.4.13 Note that F and G can be isomorphic translations even if they have different dimension functions – i.e., if they encode Σ-variables into different-length strings of Σ'-variables. We will see an example below of a single sorted theory T, and a two-dimensional translation $F : T \to T$ that is isomorphic to the identity translation $1_T : T \to T$. In this case, the theory T might be glossed as saying: "pairs of individuals correspond uniquely to individuals."

DEFINITION 5.4.14 We say that two theories T and T' are **weakly intertranslatable** (also **homotopy equivalent**) if there are translations $F : T \to T'$ and $G : T' \to T$, and homotopies $\chi : GF \Rightarrow 1_T$ and $\chi' : 1_{T'} \Rightarrow FG$.

NOTE 5.4.15 Here the word "weakly" in "weakly equivalent" shouldn't be taken to hold any deep philosophical meaning – as if it indicates that the theories aren't fully equivalent. Instead, the use of that word traces back to category theory and topology, where it has proven to be interesting to "weaken" notions of strict equality, isomorphism, or homeomorphism. In many such cases, the weakened notion is a more interesting and useful notion than its strict counterpart. One thing we like about this proposed notion of theoretical equivalence is precisely its connection with the sorts of notions that prove to be fruitful in contemporary mathematical practice. If we were to wax metaphysical, we might say that such notions carve mathematical reality at the joints.

Example 5.4.16 (Quantifier variance) We can now complete the discussion of Example 5.4.4 by showing that Mette the Mereologist and Niels the Nihilist have equivalent theories – at least by the standard of "weak intertranslatability." Recall that the translation $F : T \to T'$ includes Niels' theory as a subtheory of Mette's, restricted to the atoms. The translation $G : T' \to T$ maps Mette's variables to pairs of Niels' variables (up to permutation), and it translates the parthood relation as the relation that holds between a diagonal pair and non-diagonal pair that matches in one place.

We give an informal description of the homotopy maps $\varepsilon : GF \Rightarrow 1_T$ and $\eta : FG \Rightarrow 1_{T'}$. First, $GF\sigma = \sigma, \sigma$. That is, GF translates Niels' variables into pairs of Niels' variables, and the domain formula is the diagonal $x = y$. It's easy enough then to define a functional relation

$$\varepsilon(x, y; z) \leftrightarrow (x = y) \wedge (x = z),$$

from the diagonal of σ, σ to σ. For the homotopy map η, note that FG translates Mette's variables into pairs of Mette's variables, and the domain formula is $a(x) \wedge a(y)$ – i.e., both x and y are atoms. We then define $\eta(x, y; z)$ to be the functional relation such that if $x = y$ then $z = x$, and if $x \neq y$, then z is the composite of x and y. A tedious verification shows that ε and η satisfy the definition of homotopy maps, and therefore F, G form a homotopy equivalence.

Thus, there is a precise notion of theoretical equivalence that validates the claim of quantifier variance. However, this fact just pushes the debate back one level – to a debate over what we should take to be the "correct" notion of theoretical equivalence. Perhaps weak intertranslatability seems more mathematically natural than its strong counterpart. Or perhaps weak intertranslatability is closer to the notion that mathematicians use in practice. But these kinds of considerations could hardly be expected to move someone who antecedently rejects the claim of quantifier variance. ⌟

Example 5.4.17 Let's look now at an example that is relevant to the debate between Carnap and Quine.

Suppose that $\Sigma = \{\sigma_1, \sigma_2, p, q\}$, with p a unary predicate symbol of sort σ_1, and q a unary predicate symbol of sort σ_2. Let T be the empty theory in Σ. For simplicity, we will suppose that T implies that there are at least two things of sort σ_1, and at least two things of sort σ_2. In order to get a more intuitive grasp on this example, let's suppose that the T-theorist is intending to use σ_1 to model the domain of mathematical objects, and σ_2 to model the domain of physical objects. As Carnap might say, "mathematical object" and "physical object" are *Allwörter* to mark out domains of inquiry. Let's suppose also that $p(x)$ stands for "x is prime," and $q(x)$ stands for "x is massive" (i.e., has nonzero mass).

Now, Quine thinks that there's no reason to use sorts. Instead, he says, we should suppose that there is a single domain that can be divided by the predicates, "being a mathematical object" and "being a physical object." He says,

since the philosophers [viz. Carnap] who would build such categorial fences are not generally resolved to banish from language all falsehoods of mathematics and like absurdities, I fail to see much benefit in the partial exclusions that they do undertake; for the forms concerned would remain still quite under control if admitted rather, like self-contradictions, as false. (Quine, 1960, p. 229)

Quine's proposal seems to be the following:

1. Unify the sorts σ_1 and σ_2 into a single sort σ; and
2. For each formula ϕ with a type-mismatch, such as "There is a massive number," declare that ϕ is false.

For example, in the signature Σ, the predicate symbols p and q are of different sorts, hence they cannot be applied to the same variable, and $\phi \equiv \exists x(p(x) \wedge \neg q(x))$ is ill-formed. Quine suggests then that ϕ should be taken to be false. But what then are we to do about the fact that $\neg\phi \vdash \forall x(p(x) \rightarrow q(x))$? If ϕ is false, then it follows that all prime numbers are massive. Something has gone wrong here.

Of course, Quine is right to think that the many-sorted theory T is equivalent to a single-sorted theory T_1. Nonetheless, there are a couple of problems for Quine's suggestion that we simply throw away T in favor of T_1. First, there is another single-sorted theory T_2 that is equivalent to T, but T_1 and T_2 disagree on how to extend the ranges of predicates in T. Quine provides no guidance about whether to choose T_1 or T_2, and it seems that the choice would have to be *conventional*. The second problem is that T leaves open possibilities for specification that would be prematurely settled by passing to T_1 (or to T_2).

To be more specific, we will construct these theories T_1 and T_2. First let $\Sigma_i = \{\sigma, u, p', q'\}$, where σ is a sort symbol, and u, p', q' are unary predicate symbols. Let T_1 be the theory in Σ_1 with axioms:

$$T_1 \vdash \exists x u(x) \wedge \exists x \neg u(x)$$
$$T_1 \vdash \forall x(\neg u(x) \rightarrow \neg p'(x))$$
$$T_1 \vdash \forall x(u(x) \rightarrow \neg q'(x)).$$

The first axiom ensures that the domains u and $\neg u$ are nonempty. The second axiom implements Quine's requirement that physical objects are not prime, and the third axiom implements Quine's requirement that mathematical objects are not massive. It then follows that

$$T_1 \vdash \neg\exists x(p'(x) \wedge q'(x))$$
$$T_1 \vdash \forall x(p'(x) \rightarrow \neg q'(x)).$$

It's not difficult to see that T can be translated into T_1. Indeed, we can set $F(\sigma_1) = \sigma = F(\sigma_2)$, taking the domain formula for σ_1 variables to be u, and the domain formulas for σ_2 variables to be $\neg u$. We can then set $F(p) = p'$ and $F(q) = q'$. It is not difficult to see that F is a translation. In fact, there is also a translation G from T_1 to T, but it is more difficult to define. The problem here is determining how to translate a variable x of the signature Σ_1 into variables of the signature Σ. In particular, x ranges over things that satisfy $u(x)$ as well as things that satisfy $\neg u(x)$, but each variable of Σ is held fixed to one of the sorts, either σ_1 or σ_2.

Consider now the theory T_2 that is just like T_1 except that it replaces the axiom $\forall x(u(x) \rightarrow \neg q'(x))$ with the axiom $\forall x(u(x) \rightarrow q'(x))$. The theory T_2 differs from T_1 precisely in that it adopts a different convention for how to extend the predicates $q'(x)$ and $\neg q'(x)$ to the domain $u(x)$. T_1 says that q' should be restricted to $\neg u(x)$, and T_2 says that $\neg q'$ should be restricted to $\neg u(x)$. Quine's original proposal seems to say that we should restrict *all* predicates of sort σ_2 to $\neg u(x)$, but that proposal is simply incoherent.

Thus, the many-sorted theory T could be replaced with the single-sorted theory T_1, or it could be replaced with the single-sorted theory T_2. In one sense, it shouldn't make any difference which of these two single-sorted theories we choose. (In fact, T_1 and T_2 are intertranslatable in the strict, single-sorted sense.) But in another sense, either choice could block us from adding new truths to the theory T.

Suppose, for example, that we decided to hold on to T, instead of replacing it with T_1 or T_2. Suppose further that we come to discover that

$$\psi \equiv \exists x p(x) \wedge \neg \exists y \neg q(y).$$

But if we take the translation manual $p \mapsto p'$ and $q \mapsto q'$, then T_1 rules out ψ since $T_1 \vdash \forall x(p'(x) \rightarrow \neg q'(x))$. In this case, then, it would have been disastrous to follow Quine's recommendation to replace T by T_1, because we would have thereby stipulated as false something that T allows to be true. One of the important lessons of this example is that equivalent theories aren't equally good in all ways. ⌟

5.5 Symmetry

Philosophers of science, and especially philosophers of physics, are fascinated by the topic of symmetry. And why so? For one, because contemporary physics is chock full of symmetries and symmetry groups. Moreover, philosophers of physics have taken it upon themselves to *interpret* the theories of physics – by which they mean, among other things, to say what those theories *really mean*, and to lay bare their *ontological commitments*. In the famous words of Bas van Fraassen, the goal of interpreting a theory is to say how the world might be such that the theory is true.

Symmetry is now thought to play a special role in interpretation, in particular as a tool to winnow the ontological wheat from the formal chaff (sometimes affectionately called "descriptive fluff" or "surplus structure"). Here's how the process is supposed to work: we are given a theory T that says a bunch of things. Some of the things that T says, we should take seriously. But some of the other things that T says – or seems, prima facie, to say – should not be taken seriously. So what rule should we use to factorize T into the pure descriptive part T_0, and the superfluous part T_1? At this point, we're supposed to look to the symmetries of T. In rough-and-ready formulation, T_0 is the part of T that is invariant under symmetries, and T_1 is the part of T that is not invariant under symmetries.

Philosophers didn't make this idea out of thin air; instead, they abstracted it from well-known examples of theories in physics.

- If you describe space by a three-dimensional vector space V, then you must associate the origin $0 \in V$ with a particular point in space. But all points in space were created equal, so the representation via V says something misleading. We can then wash out this superfluous structure by demanding that translation $x \mapsto x + a$ be a symmetry, which amounts to replacing V with the affine space A over V.
- In classical electrodynamics, we can describe the electromagnetic field via potentials. However, the values of these potentials don't matter; only the gradients (rates of change) of the potentials matter. There is, in fact, a group G of symmetries that changes the values of the potentials but leaves their gradients (and, hence, the Maxwell tensor F_{ab}) invariant.
- In quantum field theory, there is an algebra $\mathscr{F}$ of field operators and a group G of symmetries. Not all field operators are invariant under the group G. Those field operators that are invariant under G are called *observables*, and it is a common opinion that only the observables are "real."

Based on these examples, and others like them, it's tempting for philosophers to propose methodological rules, such as: "if two things are related by a symmetry, then they are the same," or "a thing is real only if it is invariant under symmetries." Such principles are tendentious, but my goal here isn't to attack them directly. Even before we can discuss the merits of these principles, we need to be clearer about what symmetries are.

What is a symmetry of a theory? Sometimes we hear talk of permutations of models. Other times we hear talk of permutations of spacetime points. And yet other times we hear talk about transformations of coordinates. The goal of this section, stated bluntly, is to clear away some of the major sources of confusion. These confusions come from conflating things that ought to be kept distinct. The first thing to distinguish are theories and individual models. Even if one is a firm believer in the semantic view of theories, still a collection of models is a very different thing from an individual model; and a symmetry of an individual model is a very different thing from a symmetry on the class of models. The second thing to distinguish is, yet again, syntax and semantics. One can look at symmetries from either point of view, but confusion can arise when we aren't clear about which point of view we're taking.

In physics itself, one occasionally hears talk of symmetries of equations. Such talk is especially prominent in discussions of spacetime theories, where one says things like, "X transforms as a tensor." Nonetheless, in recent years, philosophers of science have tended to look at symmetries as transformations of models. Certainly, it is possible to develop a rigorous mathematical theory of symmetries of models – as we shall discuss in the following two chapters. However, transformations of models aren't the only kind of symmetries that can be defined in a mathematically rigorous fashion. In this section, we discuss **syntactic symmetries** – i.e., symmetries of a theory considered as a syntactic object.

Some examples of syntactic symmetries are quite obvious and trivial.

Example 5.5.1 Let $\Sigma = \{p, q\}$ be a propositional logic signature, and let T be the empty theory in Σ. It seems intuitively correct to say that T cannot distinguish between the propositions p and q. And, indeed, we can cash this intuition out in terms of a "self-translation" $F : T \to T$. In particular, let F be the translation given by $Fp = q$ and $Fq = p$. It's easy to see then that F is its own inverse. Thus, F is a "self-equivalence" of the theory T. ⌟

In the previous example, F is its own inverse, and it is an exact inverse – i.e., $FF\phi$ is literally the formula ϕ. To formulate a general definition of a syntactic symmetry, both of these conditions can be loosened. First, the inverse of F may be a different translation $G : T \to T$. Second, G need not be an inverse in the strict sense, but only an inverse up to provable equivalence. Thus, we require only that there is a $G : T \to T$ such that $GF \simeq 1_T$ and $FG \simeq 1_T$ – i.e., $F : T \to T$ is an equivalence of theories.

DEFINITION 5.5.2 Let $F : T \to T$ be a translation of a theory T to itself. We say that F is a **syntactic symmetry** just in case F is an equivalence of theories.

DISCUSSION 5.5.3 The previous definition can make one's head spin. Isn't T trivially equivalent to itself? What does it mean to say that $F : T \to T$ is an equivalence? Just remember that whenever we say that two theories are equivalent, that is shorthand for saying that there is at least one equivalence between them. There may be, and typically will be, many different equivalences between them.

Example 5.5.4 Let's slightly change the previous example. Suppose now that T' is the theory in Σ with the single axiom $\vdash p$. Then intuitively, there should not be a symmetry of T' that takes p to q and vice versa. And that intuition can indeed be validated, although we leave the details to the reader. ⌟

Example 5.5.5 Now for a predicate logic example. Let Σ consist of a single binary relation symbol r. As shorthand, let's write $\phi(x, y) \equiv r(y, x)$, which is the "opposite" relation r^{op} of r. Let T be the empty theory in Σ. Now we define a translation $F : T \to T$ by setting $Fr = \phi$. To be more precise,

$$(Fr)(x, y) \;=\; \phi(x, y) \;=\; r(y, x).$$

In effect, F flips the order of the variables in r. It is easy to see then that $F : T \to T$ is a syntactic symmetry. ⌟

Example 5.5.6 Let's slightly change the previous example. Suppose now that T' is the theory in Σ with the single axiom

$$\vdash \forall x \exists y\, r(x, y).$$

Then there is no syntactic symmetry $F : T' \to T'$ such that $Fr = r^{op}$. Indeed, if there were such a symmetry F, then we would have

$$\forall x \exists y\, r(x,y) \vdash \forall x \exists y\, r(y,x),$$

which is intuitively not the case (and which can indeed be shown not to be the case).

Incidentally, this example shows yet again why it's not always good to identify things that are related by a symmetry. In the previous example, the relations $r(x,y)$ and $r^{op}(x,y)$ are related by a symmetry. A person with Ockhamist leanings may be sorely tempted to say that there is redundancy in the description provided by T, and that a better theory would treat $r(x,y)$ and $r^{op}(x,y)$ as a single relation. However, treating $r(x,y)$ and $r^{op}(x,y)$ as the same relation would foreclose certain possibilities – e.g., the possibility that $\forall x \exists y\, r(x,y)$ holds but $\forall x \exists y\, r^{op}(x,y)$ does not. In summary, redundancy in ideology isn't directly analogous to redundancy in ontology, and we should think twice before applying Ockham's razor at the ideological level. (For discussion of an analogous concrete case, see Belot [1998].) ⌟

EXERCISE 5.5.7 Suppose now that T is the theory in Σ with the single axiom

$$r(x,y) \vdash \neg r(y,x),$$

which says that r is asymmetric. This axiom can be rewritten as

$$r(x,y) \vdash \neg r^{op}(x,y).$$

Show that $Fr = r^{op}$ defines a symmetry of T.

EXERCISE 5.5.8 Show that the theory of a partial order (Example 4.1.1) has a symmetry that maps $\leq$ to the converse relation $\geq$.

Example 5.5.9 In the nineteenth century, mathematicians discovered a neat feature of projective geometry: points and lines play a dual role in the theory. Thus, they realized, every theorem in projective geometry automatically has a dual theorem, where the roles of points and lines have been reversed. In terms of first-order logic, projective geometry is most conveniently formulated within a many-sorted framework. We shall describe it as such in Section 7.4. One can also present projective geometry as a single-sorted theory T, with predicates for "is a point" and "is a line." In this case, the duality of projective geometry is a syntactic symmetry F of T that exchanges these two predicates. The duality of theorems amounts to the fact that $T \vdash \phi$ iff $T \vdash F\phi$.

A similar duality holds for the first-order theory of categories (see 5.1.8). In that case, the symmetry permutes the domain and codomain functions on arrows. One speaks intuitively of "flipping the arrows." However, that way of speaking can be misleading, since it suggests an action on a model (i.e., on a category), and not an action on syntactic objects. As we will soon see (Section 7.2), every syntactic symmetry of a theory does induce a functor on the category of models of that theory. In the case of the theory of categories, this dual functor takes each category $\mathbf{C}$ to its opposite category $\mathbf{C}^{op}$. ⌟

We now consider a special type of syntactic symmetry – a type that we might want to call **inner symmetry** or **continuous symmetry**. (The analogy here is an element of a Lie group that is connected by a continuous path to the identity element.) Suppose that $F : T \to T$ is a self-translation with the feature that $F \simeq 1_T$. That last symbol means, intuitively and loosely, that there is a formula $\chi(x, y)$ of Σ that establishes a bijective correspondence between the original domain of the quantifiers and the restricted domain $D_F(y)$. This bijective correspondence also matches up the extension of ϕ with the extension of $F\phi$, for each formula ϕ of Σ. (All these statements are relative to the theory T.)

The reason we might want to call F an "inner symmetry" is because the theory T itself can "see" that the formulas ϕ and $F\phi$ are coextensive: $T \vdash \phi \leftrightarrow F\phi$. In the general case of a syntactic symmetry, ϕ and $F\phi$ need not be coextensive. (In the first example, we have $Fp = q$, but $T \nvdash p \leftrightarrow q$.)

We claim that whenever this condition holds, i.e., when $F \simeq 1_T$, then F is a syntactic symmetry. Indeed, it's easy to check that $FF \simeq 1_T$, and hence F is an equivalence.

Example 5.5.10 Let Σ be a signature with a single propositional constant p. Let T be the empty theory in Σ. Define a reconstrual F of Σ by setting $Fp = \neg p$. Since Σ is empty, F is a translation. Moreover, since $FFp = \neg\neg p$ and $T \vdash p \leftrightarrow \neg\neg p$, it follows that F is its own quasi-inverse. Therefore, F is a syntactic symmetry. This result is not at all surprising: from the point of view of the empty theory T, p and $\neg p$ play the same sort of role.

Indeed, recall from Chapter 3 that translations between propositional theories correspond to homomorphisms between the corresponding Lindenbaum algebras. In this case, $F : T \to T$ corresponds to an automorphism $f : B \to B$. Moreover, B is the four-element Boolean algebra, and f is the automorphism that flips the two middle elements.

Although F is a syntactic symmetry, it is not the case that $T \vdash p \leftrightarrow Fp$. Therefore, F is not inner. Using the correspondence with Lindenbaum algebras, it's easy to see that T has no nontrivial inner symmetries. Or, to be more precise, if G is an inner symmetry of T, then $G \simeq 1_T$. For example, for $G = FF$, we have $Gp = \neg\neg p$. Here G is not strictly equal to the identity translation 1_T. Rather, for each formula ϕ, we have $T \vdash \phi \leftrightarrow G\phi$. ⌟

Example 5.5.11 Let T be Mette the Mereologist's theory, and let T' be Niels the Nihilist's theory. Recall from 5.4.16 that there is a pair of translations $F : T \to T'$ and $G : T' \to T$ that forms an equivalence. Thus, $GF \simeq 1_T$ and GF is an inner symmetry of Mette's theory. Here GF is the mapping that (intuitively speaking) relativizes Mette's quantifier to the domain of atoms. ⌟

Example 5.5.12 Let $\Sigma = \{\sigma_1, \sigma_2\}$, and let T be the empty theory in Σ. Define a reconstrual $F : \Sigma \to \Sigma$ by setting $F(\sigma_1) = \sigma_2$ and $F(\sigma_2) = \sigma_1$. Then F is a symmetry of T. This symmetry F is the only nontrivial symmetry of T, and it is not deformable to the identity 1_T. (If F were deformable to 1_T, then T would define an isomorphism

between σ_1 and σ_2.) In contrast, the empty theory T' in signature $\Sigma' = \{\sigma\}$ has no nontrivial symmetries. It follows that T and T' are not equivalent in the category **Th**. Finally, let T'' be the theory in $\Sigma \cup \{f\}$, where f is a function symbol, and where T'' says that $f : \sigma_1 \to \sigma_2$ is an isomorphism. Then F is still a symmetry of T'', and it *is* is contractible to $1_{T''}$. In fact, it is not difficult to see that T' and T'' are equivalent. This equivalence will send the isomorphism f of T'' to the equality relation for T'. ⌟

The examples we have given were all drawn from first-order logic, and not even from the more complicated parts thereof (e.g., it would be interesting to investigate the syntactic symmetries of first-order axiomatizations of special relativity). The goal has been merely to illustrate the fact that it would be a mistake to consider syntactic symmetries as trivial symmetries; in fact, the syntactic symmetries of a theory tell us a lot about the structure of that theory, and even about the relations between theories. For example, if two theories are equivalent, then they have the same group of syntactic symmetries.

We have also been keen to emphasize that having "redundant syntactic structure" – in particular, having nontrivial syntactic symmetry – is by no means a defect of a theory. Indeed, one of the reasons to allow syntactic redundancy in a theory is to leave open the possibility of future developments of that theory.

5.6 Notes

- For more details on many-sorted logic, see Feferman (1974), Manzano (1993), and Manzano (1996). The last of these also discusses a sense in which second-order logic (with Henkin semantics) is reducible to many-sorted first-order logic. For an application of many-sorted logic in recent metaphysics, see Turner (2010, 2012).
- The concept of Morita equivalence – if not the name – is already familiar in certain circles of logicians. See Andréka et al. (2008) and Mere and Veloso (1992). The name "Morita equivalence" descends from Kiiti Morita's work on rings with equivalent categories of modules. Two rings R and S are said to be Morita equivalent just in case there is an equivalence $\mathrm{Mod}(R) \cong \mathrm{Mod}(S)$ between their categories of modules. The notion was generalized from rings to algebraic theories by Dukarm (1988). See also Adámek et al. (2006). There is also a notion of Morita equivalence for C^*-algebras, see Rieffel (1974). More recently, topos theorists have defined theories to be Morita equivalent just in case their classifying toposes are equivalent (Johnstone, 2003). See Tsementzis (2017b) for a comparison of the topos-theoretic notion of Morita equivalence with ours.
- Price (2009) discusses Quine's criticism of Carnap's *Allwörter*, coming to a similar conclusion as ours – but approaching it from a less technical angle. We agree with Price that in citing the technical result, Quine didn't settle the philosophical debate.

- The notion of a generalized translation between first-order theories seems to have been first described in van Benthem and Pearce (1984), who mention antecedent work by Szczerba (1977) and Gaifman. Our treatment is essentially a generalization of what can be found in Visser (2006); Friedman and Visser (2014); Rooduijn (2015). Our notion of homotopy is inspired by similar notions in Ahlbrandt and Ziegler (1986).
- The implementation of Morita equivalence to first-order logic comes from Barrett and Halvorson (2016b). We claim no originality for the notion of defining new sorts. For example, Burgess (1984) uses "extension by abstractions," which is the same thing as our quotient sorts. See also Mere and Veloso (1992); Andréka et al. (2008).
- Quine's argument for the dispensability of many-sorted logic is discussed by Barrett and Halvorson (2017b).
- For recent considerations on quantifier variance, see Warren (2014); Dorr (2014); Hirsch and Warren (2017).
- For more on symmetry, see Weatherall (2016b); Dewar (2017b); Barrett (2018b).

6 Semantic Metalogic

6.1 The Semantic Turn

Already in the nineteenth century, geometers were proving the relative consistency of theories by interpreting them into well-understood mathematical frameworks – e.g., other geometrical theories or the theory of real numbers. At roughly the same time, the theory of sets was under active development, and mathematicians were coming to realize that the things they were talking about (numbers, functions, etc.) could be seen to be constituted by sets. However, it was only in the middle of the twentieth century that Alfred Tarski gave a precise definition of an *interpretation* of a theory in the universe of sets.

Philosophers of science were not terribly quick to latch onto the new discipline of logical semantics. Early adopters included the Dutch philosopher Evert Beth and, to a lesser extent, Carnap himself. It required a generational change for the semantic approach to take root in philosophy of science. Here we are using "semantic approach" in the broadest sense – essentially for any approach to philosophy of science that is reactionary against Carnap's syntax program, but that wishes to use precise mathematical tools (set theory, model theory, etc.) in order to explicate the structure of scientific theories.

What's most interesting for us is how the shift to the semantic approach influenced shifts in philosophical perspective. Some of the cases are fairly clear. For example, with the rejection of the syntactic approach, many philosophers stopped worrying about the "problem of theoretical terms" – i.e., how scientific theories (with their abstract theoretical terms) connect to empirical reality. According to Putnam, among others, if you step outside the confines of Carnap's *Wissenschaftslogik* program, there is no problem of theoretical terms. (Interestingly, debates about the conventionality of geometry all but stopped around the 1970s, just when the move to the semantic view was in full swing.) Other philosophers diagnosed the situation differently. For example, van Fraassen saw the semantic approach as providing the salvation of empiricism – which, he thought, was incapable of an adequate articulation from a syntactic point of view.

In reading twentieth-century analytic philosophy, it can seem that logical semantics by itself is supposed to obviate many of the problems that exercised the previous generation of philosophers. For example, van Fraassen (1989, p. 222) says that "the semantic view of theories makes language largely irrelevant to the subject [philosophy of science]." Indeed, the picture typically presented to us is that logical semantics deals

with mind-independent things (viz. set-theoretic structures), which can stand in mind-independent relations to concrete reality, and to which we have unmediated epistemic access. Such a picture suggests that logical semantics provides a bridge over which we can safely cross the notorious mind–world gap.

But something is fishy with this picture. How could logical semantics get any closer to "the world" than any other bit of mathematics? And why think that set-theoretic structures play this privileged role as intermediaries in our relation to empirical reality? For that matter, why should our philosophical views on science be tied down to some rather controversial view of the nature of mathematical objects? Why the set-theoretic middleman?

In what follows, we will attempt to put logical semantics back in its place. The reconceptualization we're suggesting begins with noting that logical semantics is a particular version of a general mathematical strategy called "representation theory." There is a representation theory for groups, for rings, for C^*-algebras, etc., and the basic idea of all these representation theories is to study one category $\mathbf{C}$ of mathematical objects by studying the functors from $\mathbf{C}$ to some other mathematical category, say $\mathbf{S}$. It might seem strange that such an indirect approach could be helpful for understanding $\mathbf{C}$, and yet, it has proven to be very frutiful. For example, in the representation theory of groups, one studies the representations of a group on Hilbert spaces. Similarly, in the representation theory of rings, one studies the modules over a ring. In all such cases, there is no suggestion that a represented mathematical object is less linguistic than the original mathematical object. If anything, the represented mathematical object has superfluous structure that is not intrinsic to the original mathematical object.

To fully understand that logical semantics is representation theory, one needs to see theories as objects in a category, and to show that "interpretations" are functors from that category into some other one. We carried out that procedure for propositional theories in Chapter 3, where we represented each propositional theory as a Boolean algebra. We could carry out a similar construction for predicate logic theories, but the resulting mathematical objects would be something more complicated than Boolean algebras. (Tarski himself suggested representing predicate logic theories as cylindrical algebras, but a more elegant approach involves syntactic categories in the sense of Makkai and Reyes [1977].) Thus, we will proceed in a different manner and directly define the arrows (in this case, translations) between predicate logic theories. We begin, however, with a little crash course in traditional model theory.

Example 6.1.1 Let T be the theory, in empty signature, that says, "there are exactly two things." A **model** of T is simply a set with two elements. However, every model of T has "redundant information" that is not specified by T itself. To the question "how many models does T have?" there are two correct answers: (1) more than any cardinal number and (2) exactly one (up to isomorphism). ⌟

Example 6.1.2 Let T_1 be the theory of groups, as axiomatized in Example 4.5.3. Then a model M of T_1 is a set S with a binary function $\cdot^M : S \times S \to S$ and a preferred element $e^M \in S$ that satisfy the conditions laid out in the axioms. Once again, every such model

M carries all the structure that T_1 requires of it and then some more structure that T_1 doesn't care about. ⌟

In order to precisely define the concept of a model of a theory, we must first begin with the concept of a Σ-structure.

DEFINITION 6.1.3 A **Σ-structure** M is a mapping from Σ to appropriate structures in the category **Sets**. In particular, M fixes a particular set S, and then

- M maps each n-ary relation symbol $p \in \Sigma$ to a subset $M(p) \subseteq S^n = S \times \cdots \times S$.
- M maps each n-ary function symbol $f \in \Sigma$ to a function $M(f) : S^n \to S$.

A Σ-structure M extends naturally to all syntactic structures built out of Σ. In particular, for each Σ-term t, we define $M(t)$ to be a function, and for each Σ-formula ϕ, we define $M(\phi)$ to be a subset of S^n (where n is the number of free variables in ϕ). In order to do so, we need to introduce several auxiliary constructions.

DEFINITION 6.1.4 Let Γ be a finite set of Σ-formulas. We say that $\vec{x} = x_1, \ldots, x_n$ is a **context** for Γ just in case $\vec{x}$ is a duplicate-free sequence that contains all free variables that appear in any of the formulas in Γ. We say that $\vec{x}$ is a **minimal context** for Γ just in case every variable x_i in $\vec{x}$ occurs free in some formula in Γ. Note: we also include, as a context for sentences, the zero-length string of variables.

DEFINITION 6.1.5 Let $\vec{x}$ and $\vec{y}$ be duplicate-free sequences of variables. Then $\vec{x}.\vec{y}$ denotes the result of concatenating the sequences, then deleting repeated variables in order from left to right. Equivalently, $\vec{x}.\vec{y}$ results from deleting from $\vec{y}$ all variables that occur in $\vec{x}$, and then appending the resulting sequence to $\vec{x}$.

DEFINITION 6.1.6 For each term t, we define the **canonical context** $\vec{x}$ of t as follows. First, for a variable x, the canonical context is x. Second, suppose that for each term t_i, the canonical context $\vec{x}_i$ has been defined. Then the canonical context for $f(t_1, \ldots, t_n)$ is $(\cdots((\vec{x}_1.\vec{x}_2)\cdots).\vec{x}_n$.

EXERCISE 6.1.7 Suppose that $\vec{x} = x_1, \ldots, x_n$ is the canonical context for t. Show that $FV(t) = \{x_1, \ldots, x_n\}$.

DEFINITION 6.1.8 For each formula ϕ, we define the **canonical context** $\vec{x}$ of ϕ as follows. First, if $\vec{x}_i$ is the canonical context for t_i, then the canonical context for $t_1 = t_2$ is $\vec{x}_1.\vec{x}_2$, and the canonical context for $p(t_1, \ldots, t_n)$ is $(\cdots((\vec{x}_1.\vec{x}_2)\cdots).\vec{x}_n$. For the Boolean connnectives, we also use the operation $\vec{x}_1.\vec{x}_2$ to combine contexts. Finally, if $\vec{x}$ is the canonical context for ϕ, then the canonical context for $\forall x\phi$ is the result of deleting x from $\vec{x}$, if it occurs.

EXERCISE 6.1.9 Show that the canonical context for ϕ does, in fact, contain all and only those variables that are free in ϕ.

If a Σ-structure M has a domain set S, then it assigns relation symbols to subsets of the Cartesian products,

$$S, S \times S, S^3, \ldots$$

Of course, these sets are all connected to each other by projection maps, such as the projection $S \times S \to S$ onto the first coordinate. We will now develop some apparatus to handle these projection maps. To this end, let $[n]$ stand for the finite set $\{1, \ldots, n\}$.

LEMMA 6.1.10 *For each injective function $p : [m] \to [n]$, there is a unique projection $\pi_p : S^n \to S^m$ defined by*

$$\pi_p \langle x_1, \ldots, x_n \rangle = \langle x_{p(1)}, \ldots, x_{p(m)} \rangle.$$

Furthermore, if $q : [\ell] \to [m]$ is also injective, then $\pi_{p \circ q} = \pi_q \circ \pi_p$.

Proof The first claim is obvious. For the second claim, it's easier if we ignore the variables $x_1, \ldots, x_n$ and note that π_p is defined by the coordinate projections:

$$\pi_i \circ \pi_p \;=\; \pi_{p(i)},$$

for $i = 1, \ldots, m$. Thus, in particular,

$$\pi_i \circ \pi_q \circ \pi_p \;=\; \pi_{q(i)} \circ \pi_p \;=\; \pi_{p(q(i))} \;=\; \pi_i \circ \pi_{p \circ q},$$

which proves the second claim. □

DEFINITION 6.1.11 Let $\vec{x} = x_1, \ldots, x_m$ and $\vec{y} = y_1, \ldots, y_n$ be duplicate-free sequences of variables. We say that $\vec{x}$ is a **subcontext** of $\vec{y}$ just in case each element in $\vec{x}$ occurs in $\vec{y}$. In other words, for each $i \in [m]$, there is a unique $p(i) \in [n]$ such that $x_i = y_{p(i)}$. Since $i \mapsto y_i$ is injective, $p : [m] \to [n]$ is also injective. Thus, p determines a unique projection $\pi_p : S^n \to S^m$. We say that π_p is the **linking projection** for contexts $\vec{y}$ and $\vec{x}$. If $\vec{x}$ and $\vec{y}$ are canonical contexts of formulas or terms, then we say that π_p is the **linking projection** for these formulas or terms.

We are now ready to complete the extension of the Σ-structure M to all Σ-terms.

DEFINITION 6.1.12 For each term t with n-free variables, we define $M(t) : S^n \to S$.

1. Recall that a constant symbol $c \in \Sigma$ is really a special case of a function symbol, viz. a 0-ary function symbol. Thus, $M(c)$ should be a function from S^0 to S. Also recall that the 0-ary Cartesian product of any set is a one-point set $\{*\}$. Thus, $M(c) : \{*\} \to S$, which corresponds to a unique element $c^M \in S$.
2. For each variable x, we let $M(x) : S \to S$ be the identity function. This might seem like a strange choice, but its utility will soon be clear.
3. Let $t \equiv f(t_1, \ldots, t_n)$, where $M(t_i)$ has already been defined. Let n_i be the number of free variables in t_i. The context for t_i is a subcontext of the context for t. Thus, there is a linking projection $\pi_i : S^n \to S^{n_i}$. Whereas the $M(t_i)$ may have different domains (if $n_i \neq n_j$), precomposition with the linking projections makes them functions of a common domain S^n. Thus, we define

$$M[f(t_1, \ldots, t_n)] \;=\; M(f) \circ \langle M(t_1) \circ \pi_1, \ldots, M(t_n) \circ \pi_n \rangle,$$

which is a function from S^n to S.

We illustrate the definition of $M(t)$ with a couple of examples.

Example 6.1.13 Suppose that f is a binary function symbol, and consider the two terms $f(x,y)$ and $f(y,x)$. The canonical context for $f(x,y)$ is x,y, and the canonical context for $f(y,x)$ is y,x. Thus, the linking projection for $f(x,y)$ and x is the projection $\pi_0 : S \times S \to S$ onto the first coordinate; and the linking projection for $f(y,x)$ and x is $\pi_1 : S \times S \to S$ onto the second coordinate. Thus,

$$M(f(x,y)) = M(f) \circ \langle \pi_0, \pi_1 \rangle = M(f).$$

A similar calculation shows that $M(f(y,x)) = M(f)$, which is as it should be: $f(x,y)$ and $f(y,x)$ should correspond to the same function $M(f)$.

However, it does *not* follow that the formula $f(x,y) = f(y,x)$ should be regarded as a semantic tautology. Whenever we place both $f(x,y)$ and $f(y,x)$ into the *same* context, this context serves as a reference point by which the order of inputs can be distinguished. ⌟

DEFINITION 6.1.14 For each formula ϕ of Σ with n distinct free variables, we define $M(\phi)$ to be a subset of $S^n = S \times \cdots \times S$.

1. $M(\bot)$ is the empty set $\emptyset$, considered as a subset of the one-element set 1.
2. Suppose that $\phi \equiv (t_1 = t_2)$, where t_1 and t_2 are terms. Let n_i be the number of free variables in t_i. Since the context for t_i is a subcontext of that for $t_1 = t_2$, there is a linking projection $\pi_i : S^n \to S^{n_i}$. We define $M(t_1 = t_2)$ to be the equalizer of the functions $M(t_1) \circ \pi_1$ and $M(t_2) \circ \pi_2$.
3. Suppose that $\phi \equiv p(t_1, \ldots, t_m)$, where p is a relation symbol and $t_1, \ldots, t_m$ are terms. Let n be the number of distinct free variables in ϕ. Since the context of t_i is a subcontext of that of ϕ, there is a linking projection $\pi_i : S^n \to S^{n_i}$. Then $\langle \pi_1, \ldots, \pi_m \rangle$ is a function from S^n to $S^{n_1} \times \cdots \times S^{n_m}$. We define $M[p(t_1, \ldots, t_m)]$ to be the pullback of $M(p) \subseteq S^m$ along the function
$$\langle M(t_1) \circ \pi_1, \ldots, M(t_m) \circ \pi_m \rangle.$$
4. Suppose that M has already been defined for ϕ. Then we define $M(\neg\phi) = S^n \backslash M(\phi)$.
5. Suppose that ϕ is a Boolean combination of ϕ_1, ϕ_2, and that $M(\phi_1)$ and $M(\phi_2)$ have already been defined. Let π_i be the linking projection for ϕ_i and ϕ, and let π_i^* be the corresponding pullback (preimage) map that takes subsets to subsets. Then we define
$$\begin{aligned} M(\phi_1 \wedge \phi_2) &= \pi_1^*(M(\phi_1)) \cap \pi_2^*(M(\phi_2)), \\ M(\phi_1 \vee \phi_2) &= \pi_1^*(M(\phi_1)) \cup \pi_2^*(M(\phi_2)), \\ M(\phi_1 \to \phi_2) &= (S^n \backslash \pi_1^*(M(\phi_1))) \cup \pi_2^*(M(\phi_2)). \end{aligned}$$
6. Suppose that $M(\phi)$ is already defined as a subset of S^n. Suppose first that x is free in ϕ, and let $\pi : S^{n+1} \to S$ be the linking projection for ϕ and $\exists x \phi$. Then we define $M(\exists x \phi)$ to be the image of $M(\phi)$ under π, i.e.,
$$M(\exists x \phi) = \{\vec{a} \in S^n \mid \pi^{-1}(\vec{a}) \cap M(\phi) \neq \emptyset\}.$$

If x is not free in ϕ, then we define $M(\exists x\phi) = M(\phi)$.
Similarly, if x is free in ϕ, then we define

$$M(\forall x\phi) \;=\; \{\vec{a} \in S^n \mid \pi^{-1}(\vec{a}) \subseteq M(\phi)\}.$$

If x is not free in ϕ, then we define $M(\forall x\phi) = M(\phi)$.

Example 6.1.15 Let's unpack the definitions of $M(x = y)$ and $M(x = x)$. For the former, the canonical context for $x = y$ is x, y. Thus, the linking projection for $x = y$ and x is $\pi_0 : S \times S \to S$ onto the first coordinate, and the linking projection for $x = y$ and y is $\pi_1 : S \times S \to S$ onto the second coordinate. By definition, $M(x) \equiv 1_S \equiv M(y)$, and $M(x = y)$ is the equalizer of $1_s \circ \pi_0$ and $1_S \circ \pi_1$. This equalizer is clearly the diagonal subset of $S \times S$:

$$M(x = y) \;\equiv\; \{\langle a, b\rangle \in S \times S \mid a = b\} \;\equiv\; \{\langle a, a\rangle \mid a \in S\}.$$

In contrast, the canonical context for $x = x$ is x, and the linking projection for $x = x$ and x is simply the identity. Thus, $M(x = x)$ is defined to be the equalizer of $M(x)$ and $M(x)$, which is the entire set S. That is, $M(x = x) \equiv S$. ⌟

EXERCISE 6.1.16 Describe $M(f(x, y) = f(y, x))$, and explain why it won't neccesarily be the entire set $S \times S$.

We are now going to define a relation $\phi \models_M \psi$ of semantic entailment in a structure M; and we will use that notion to define the absolute relation $\phi \models \psi$ of semantic entailment. (In short: $\phi \models \psi$ means that $\phi \models_M \psi$ in every structure M.) Here ϕ and ψ are formulas (not necessarily sentences), so we need to take a bit of care with their free variables. One thing we could do is to consider the sentence $\forall\vec{x}(\phi \to \psi)$, where $\vec{x}$ is any sequence that includes all variables free in ϕ or ψ. However, even in that case, we would have to raise a question about whether the definition depends on the choice of the sequence $\vec{x}$. Since we have to deal with that question in any case, we will instead look more directly at the relation between the formulas ϕ and ψ, which might share some free variables in common.

As a first proposal, we might try saying that $\phi \models_M \psi$ just in case $M(\phi) \subseteq M(\psi)$. But the problem with this proposal is that $M(\phi)$ and $M(\psi)$ are typically defined to be subsets of different sets. For example: the definition of $\models_M$ should imply that $p(x) \models_M (p(x) \vee q(y))$. However, for any Σ-structure M, $M(p(x))$ is a subset of S whereas $M(p(x) \vee q(y))$ is a subset of $S \times S$. The way to fix this problem is to realize that $M(p(x))$ can also be considered to be a subset of $S \times S$. In particular, $p(x)$ is equivalent to $p(x) \wedge (y = y)$, and intuitively $M(p(x) \wedge (y = y))$ should be the subset of $S \times S$ of things satisfying $p(x)$ and $y = y$. In other words, $M(p(x) \wedge (y = y))$ should be $M(p(x)) \times S$.

Here's what we will do next. First we will extend the definition of M so that it assigns a formula ϕ an extension $M_{\vec{x}}(\phi)$ relative to a context $\vec{x}$. Then we will define $\phi \models_M \psi$

to mean that $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$, where $\vec{x}$ is an arbitrarily chosen context for ϕ, ψ. Then we will show that this definition does not depend on which context we chose.

In order to define $M_{\vec{y}}(\phi)$ where $\vec{y}$ is an arbitrary context for ϕ, we will first fix the canonical context $\vec{x}$ for ϕ, and we will set $M_{\vec{x}}(\phi) = M(\phi)$. Then for any other context $\vec{y}$ of which $\vec{x}$ is a subcontext, we will use the linking projection π_p to define $M_{\vec{y}}(\phi)$ as a pullback of $M_{\vec{x}}(\phi)$.

DEFINITION 6.1.17 Let $\vec{y} = y_1, \ldots, y_n$ be a context for ϕ, let $\vec{x} = x_1, \ldots, x_m$ be the canonical context for ϕ, and let $p : [m] \to [n]$ be the corresponding injection. We define $M_{\vec{y}}(\phi)$ to be the pullback of $M(\phi)$ along π_p. In particular, when $\vec{y} = \vec{x}$, then $p : [n] \to [n]$ is the identity, and $M_{\vec{x}}(\phi) = M(\phi)$.

Now we are ready to define the relation $\phi \vDash_M \psi$.

DEFINITION 6.1.18 For each pair of formulas ϕ, ψ, let $\vec{x}$ be the canonical context for $\phi \to \psi$. We say that $\phi \vDash_M \psi$ just in case $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$.

We will now show that the definition of $\phi \vDash_M \psi$ is independent of the chosen context $\vec{x}$ for ϕ, ψ. In particular, we show that for any two contexts $\vec{x}$ and $\vec{y}$ for ϕ, ψ, we have $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$ if and only if $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$. As the details of this argument are a bit tedious, the impatient reader may wish to skip to Definition 6.1.23.

We'll first check the compatibility of the definitions $M_{\vec{y}}(\phi)$ and $M_{\vec{z}}(\phi)$, where $\vec{y}$ and $\vec{z}$ are contexts for ϕ.

LEMMA 6.1.19 *Suppose that $\vec{x} = x_1, \ldots, x_\ell$ is a subcontext of $\vec{y} = y_1, \ldots, y_m$, and that $\vec{y}$ is a subcontext of $\vec{z} = z_1, \ldots, z_n$. Suppose that $p : [\ell] \to [m]$, $q : [m] \to [n]$, and $r : [\ell] \to [n]$ are the corresponding injections. Then $r = q \circ p$.*

Proof By definition of p, $y_{p(i)} = x_i$ for $i \in [\ell]$. By definition of r, $z_{r(i)} = x_i$ for $i \in [\ell]$. Thus, $y_{p(i)} = z_{r(i)}$. Furthermore, by definition of q, $z_{q(p(i))} = y_{p(i)}$. Therefore, $z_{q(p(i))} = z_{r(i)}$, and $q(p(i)) = r(i)$. □

LEMMA 6.1.20 *Suppose that $\vec{x}$ is a context for ϕ, and that $\vec{x}$ is a subcontext of $\vec{y}$. Let $\pi^r : S^n \to S^m$ be the projection connecting the contexts $\vec{y}$ and $\vec{x}$. Then $M_{\vec{y}}(\phi)$ is the pullback of $M_{\vec{x}}(\phi)$ along π_r.*

Proof Let π_p be the projection connecting $\vec{x}$ to the canonical context for ϕ, and let π_q be the projection connecting $\vec{y}$ to the canonical context for ϕ. Thus, $M_{\vec{x}}(\phi) = \pi_p^*[M(\phi)]$, where π_p^* denotes the operation of pulling back along π_p. Similarly, $M_{\vec{y}}(\phi) = \pi_q^*[M(\phi)]$. Furthermore, $\pi_q = \pi_p \circ \pi_r$, and since pullbacks commute, we have

$$M_{\vec{y}}(\phi) = \pi_q^*[M(\phi)] = \pi_r^*[\pi_p^*[M(\phi)]] = \pi_r^*[M_{\vec{x}}(\phi)],$$

as was to be shown. □

PROPOSITION 6.1.21 *Suppose that $\vec{x}$ is a context for ϕ, ψ, and that $\vec{x}$ is a subcontext of $\vec{y}$. If $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$ then $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$.*

Proof Suppose that $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$. Let $\pi_r : S^n \to S^m$ be the projection connecting the contexts $\vec{y}$ and $\vec{x}$. By the previous lemma, $M_{\vec{y}}(\phi) = \pi_r^*[M_{\vec{x}}(\phi)]$ and $M_{\vec{y}}(\psi) = \pi_r^*[M_{\vec{x}}(\psi)]$. Since pullbacks preserve set inclusion, $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$. □

Since we defined $\phi \models_M \psi$ using a minimal context $\vec{x}$ for ϕ, ψ, we now have the first half of our result: if $\phi \models_M \psi$, then $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$ for any context $\vec{y}$ for ϕ, ψ. To complete the result, we now show that redundant variables can be deleted from contexts.

LEMMA 6.1.22 *Let $\vec{x}$ be a context for ϕ, and suppose that y does not occur in $\vec{x}$. Then $M_{\vec{x}.y}(\phi) = M_{\vec{x}}(\phi) \times S$.*

Proof Let $\vec{x} = x_1, \ldots, x_n$, and let $p : [n] \to [n+1]$ be the injection corresponding to the inclusion of $\vec{x}$ in $\vec{x}.y$. In this case, $p(i) = i$ for $i = 1, \ldots, n$, and $\pi_p : S^{n+1} \to S^n$ projects out the last coordinate. By Lemma 6.1.20, $M_{\vec{x}.y}(\phi)$ is the pullback of $M_{\vec{x}}(\phi)$ along π_p. However, the pullback of any set A along π_p is simply $A \times S$. □

Now suppose that $M_{\vec{x}.y}(\phi) \subseteq M_{\vec{x}.y}(\psi)$, where $\vec{x}$ is a context for ϕ, ψ, and y does not occur in $\vec{x}$. Then the previous lemma shows that $M_{\vec{x}.y}(\phi) = M_{\vec{x}}(\phi) \times S$ and $M_{\vec{x}.y}(\psi) = M_{\vec{x}}(\psi) \times S$. Thus, $M_{\vec{x}.y}(\phi) \subseteq M_{\vec{x}.y}(\psi)$ if and only if $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$. A quick inductive argument then shows that any number of appended empty variables makes no difference.

We can now conclude the argument that $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$ if and only if $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$, where $\vec{x}$ is a subcontext of $\vec{y}$. The "if" direction was already shown in Prop. 6.1.21. For the "only if" direction, suppose that $M_{\vec{y}}(\phi) \subseteq M_{\vec{y}}(\psi)$. First use Prop. 6.1.21 again to move any variables not in $\vec{x}$ to the end of the sequence $\vec{y}$. (Recall that $\vec{y}$ is a subcontext of any permutation of $\vec{y}$.) Then use the previous lemma to eliminate these variables. The resulting sequence is a permutation of $\vec{x}$, hence a subcontext of $\vec{x}$. Finally, use Prop. 6.1.21 one more time to show that $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$. Thus, we have shown that the definition of $\phi \models_M \psi$ is independent of the context chosen for ϕ, ψ.

DEFINITION 6.1.23 We say that ϕ **semantically entails** ψ, written $\phi \models \psi$, just in case $\phi \models_M \psi$ for every Σ-structure M. We write $\models \psi$ as shorthand for $\top \models \psi$.

NOTE 6.1.24 The canonical context $\vec{x}$ for the pair $\{\top, \phi\}$ is simply the context for ϕ. By definition, $M_{\vec{x}}(\top)$ is the pullback of 1 along the unique map $\pi : S^n \to 1$. Thus, $M_{\vec{x}}(\top) = S^n$, and $\top \models_M \phi$ if and only if $M(\phi) = S^n$.

We're now ready for two of the most famous definitions in mathematical philosophy.

Truth in a Structure

A sentence ϕ has zero free variables. In this case, $M(\phi)$ is defined to be a subset of $S^0 = 1$, a one-element set. We say that ϕ is **true** in M if $M(\phi) = 1$, and we say that ϕ is **false** in M if $M(\phi) = \emptyset$.

Model

Let T be a theory in signature Σ, and let M be a Σ-structure. We say that M is a **model** of T just in case: for any sentence ϕ of Σ, if $T \vdash \phi$, then $M(\phi) = 1$.

6.2 The Semantic View of Theories

In Chapter 4, we talked about how Rudolf Carnap used syntactic metalogic to explicate the notion of a scientific theory. By the 1960s, people were calling Carnap's picture the "syntactic view of theories," and they were saying that something was fundamentally wrong with it. According to Suppe (2000), the syntactic view of theories died in the late 1960s (March 26, 1969, to be precise) after having met with an overwhelming number of objections in the previous two decades. Upon the death of the syntactic view, it was unclear where philosophy of science would go. Several notable philosophers – such as Feyerabend and Hanson – wanted to push philosophy of science away from formal analyses of theories. However, others such as Patrick Suppes, Bas van Fraassen, and Fred Suppe saw formal resources for philosophy of science in other branches of mathematics, most particularly set theory and model theory. Roughly speaking, the "semantic view of theories" designates proposals to explicate theory-hood by means of semantic metalogic.

We now have the technical resources in place to state a preliminary version of the semantic view of theories:

(SV) A scientific theory is a class of Σ-structures for some signature Σ.

Now, proponents of the semantic view will balk at SV for a couple of reasons. First, semanticists stress that a scientific theory has two components:

1. A theoretical definition and
2. A theoretical hypothesis.

The theoretical definition, roughly speaking, is intended to replace the first component of Carnap's view of theories. That is, the theoretical definition is intended to specify some abstract mathematical object – the thing that will be used to do the representing. Then the theoretical hypothesis is some claim to the effect that some part of the world can be represented by the mathematical object given by the theoretical definition. So, to be clear, SV here is only intended to give one-half of a theory, viz. the theoretical definition. I am not speaking yet about the theoretical hypothesis.

But proponents of the semantic view will balk for a second reason: SV makes reference to a signature Σ. And one of the supposed benefits of the semantic view was to free us from the language dependence implied by the syntactic view. So, how are we to modify SV in order to maintain the insight that a scientific theory is independent of the language in which it is formulated?

I will give two suggestions, the first of which I think cannot possibly succeed. The second suggestion works, but it shows that the semantic view actually has no advantage over the syntactic view in being "free from language dependence."

How then to modify SV? The first suggestion is to formulate a notion of mathematical structure that makes no reference to language. At first glance, it seems simple enough to do so. The paradigm case of a mathematical structure is supposed to be an ordered n-tuple $\langle X, R_1, \ldots R_n \rangle$, where X is a set, and $R_1, \ldots, R_n$ are relations on X. (This notion of mathematical structure follows in the footsteps of Bourbaki [1970], which, incidentally, has been rendered obsolete by category theory.) Consider, for example, the proposal made by Lisa Lloyd:

> In our discussion, a *model* is not such an interpretation [i.e., not an Σ-structure], matching statements to a set of objects which bear certain relations among themselves, but the set of objects itself. That is, models should be understood as structures; in the cases we shall be discussing, they are mathematical structures, i.e., a set of mathematical objects standing in certain mathematically representable relations. (Lloyd, 1984, p. 30)

However, it's difficult to make sense of this proposal. Consider the following example.

Example 6.2.1 Let a be an arbitrary set, and consider the following purported example of a mathematical structure:

$$M = \big\langle \{a, b, \langle a, a\rangle\}, \{\langle a, a\rangle\} \big\rangle.$$

That is, the domain X consists of three elements $a, b, \langle a, a\rangle$, and the indicated structure is the singleton set containing $\langle a, a\rangle$. But how are we supposed to understand this structure? Are we supposed to consider $\{\langle a, a\rangle\}$ to be a subset of X or as a subset of $X \times X$? The former is a structure for a signature Σ with a single unary predicate symbol; the latter is a structure for a signature Σ' with a single binary relation symbol. In writing down M as an ordered n-tuple, we haven't yet fully specified an intended mathematical structure.

We conclude then that a mathematical structure cannot simply be, "a set of mathematical objects standing in certain mathematically representable relations." To press the point further, consider another purported example of a mathematical structure:

$$N = \big\langle \{a, b, \langle a, b\rangle\}, \{\langle a, b\rangle\} \big\rangle.$$

Are M and N isomorphic structures? Once again, the answer is underdetermined. If M and N are supposed to be structures for a signature Σ with a single unary predicate symbol, then the answer is yes. If M and N are supposed to be structures for a signature Σ' with a single binary relation symbol, then the answer is no. ⌟

Thus, it's doubtful that there is any "language-free" account of mathematical structures, and hence no plausible language-free semantic view of theories. I propose then that we embrace the fact that we are "suspended in language," to borrow a phrase from Niels Bohr. To deal with our language dependence, we need to consider notions of equivalence of theory-formulations – so that the same theory can be formulated in

different languages. And note that this stratagem is available for both semantic and syntactic views of theories. Thus, "language independence" is not a genuine advantage of the semantic view of theories as against the syntactic view of theories.

Philosophical Moral

It is of crucial importance that we do not think of a Σ-structure M as representing the world. To say that the world is isomorphic to, or even partially isomorphic to, or even similar to, M, would be to fall into a profound confusion.

A Σ-structure M is *not* a "set-theoretic structure" in any direct sense of that phrase. Rather, M is a function whose domain is Σ and whose range consists of some sets, subsets, and functions between them. If one said that "M represents the world," then one would be saying that the world is represented by a mathematical object of type $\Sigma \to \mathbf{Sets}$. Notice, in particular, that M has "language" built into its very definition.

6.3 Soundness, Completeness, Compactness

We now prove versions of four central metalogical results: soundness, completeness, compactness, and Löwnheim–Skölem theorems. For these results, we will make a couple of simplifying assumptions, merely for the sake of mathematical elegance. We will assume that Σ is fixed signature that is countable and that has no function symbols. This assumption will permit us to use the topological techniques introduced by Rasiowa and Sikorski (1950).

Soundness

In its simplest form, the soundness theorem shows that for any sentence ϕ, if ϕ is provable ($\top \vdash \phi$), then ϕ is true in all Σ-structures ($\top \vDash \phi$). Inspired by categorical logic, we derive this version of soundness as a special case of a more general result for Σ-formulas. We show that: for any Σ-formulas ϕ and ψ, and for any context $\vec{x}$ for $\{\phi, \psi\}$, if $\phi \vdash_{\vec{x}} \psi$, then $M_{\vec{x}}(\phi) \subseteq M_{\vec{x}}(\psi)$.

The proof proceeds by induction on the construction of proofs – i.e., over the definition of the relation $\vdash$. Most cases are trivial verifications, and we leave them to the reader. We will just consider the case of the existential elimination rule, which we consider in the simple form:

$$\frac{\phi \vdash_{x,y} \psi}{\exists y \phi \vdash_x \psi}$$

assuming that y is not free in ψ. We assume then that the result holds for the top line – i.e., $M_{x,y}(\phi) \subseteq M_{x,y}(\psi)$. By definition, $M_x(\exists y \phi)$ is the image of $M_{x,y}(\phi)$ under the projection $X \times Y \to X$. And since y is not free in ψ, $M_{x,y}(\psi) = M_x(\psi) \times Y$.

To complete the argument, it will suffice to make the following general observation about sets: if $A \subseteq X \times Y$ and $B \subseteq X$, then the following inference is valid:

$$\frac{A \subseteq \pi^{-1}(B)}{\pi(A) \subseteq B}.$$

Indeed, suppose that $z \in \pi(A)$, which means that there is a $y \in Y$ such that $\langle z, y\rangle \in A$. By the top line, $\langle z, y\rangle \in \pi^{-1}(B)$, which means that $z = \pi\langle z, y\rangle \in B$. Now set $A = M_{x,y}(\phi)$ and $B = M_x(\psi)$, and it follows that existential elimination is sound.

We leave the remaining steps of this proof to the reader, and briefly comment on the philosophical significance (or lack thereof) of the soundness theorem. (The discussion here borrows from the ideas of Michaela McSweeney. See McSweeney [2016b].) Philosophers often gloss this theorem as showing that the derivation rules are "safe" – i.e., that they don't permit derivations which are not valid, or even more strongly, that the rules won't permit us to derive a false conclusion from true premises. But now we have a bit of a philosophical conundrum. What is this standard of validity against which we are supposed to measure $\vdash$? Moreover, why think that this other standard of validity is epistemologically prior to the standard of validity we have specified with the relation $\vdash$?

Philosophers often gloss the relation $\models$ in terms of "truth preservation." They say that $\phi \models \psi$ *means that* whenever ϕ is true, then ψ is true. Such statements can be highly misleading, if they cause the reader to think that $\models$ is the intuitive notion of truth preservation. No, the relation $\models$ is yet another attempt to capture, in a mathemtically precise fashion, our intuitive notion of logical consequence. We have two distinct ways of representing this intuitive notion: the relation $\vdash$ and the relation $\models$. The soundness and completeness theorems happily show that we've captured the same notion with two different definitions.

The important point here is that *logical syntax and logical semantics are enterprises of the same kind*. The soundness and completeness theorems are not theorems about how mathematics relates to the world, nor are they theorems about how a mathematical notion relates to an intuitive notion. No, these theorems demonstrate a relationship between mathematical things.

The soundness theorem has sometimes been presented as an "absolute consistency" result – i.e., that the predicate calculus is consistent *tout court*. But such presentations are misleading: The soundness theorem shows only that the predicate calculus is consistent relative to the relation $\models$, i.e., that the relation $\vdash$ doesn't exceed the relation $\models$. It doesn't prove that there is no sentence ϕ such that $\models \phi$ and $\models \neg\phi$. We agree, then, with David Hilbert: the only kind of formal consistency is relative consistency.

Completeness

In Chapter 3, we saw that the completeness theorem for propositional logic is equivalent to the Boolean ultrafilter axiom (i.e., every nonzero element in a Boolean algebra is contained in an ultrafilter). In many textbooks of logical metatheory, the completeness theorem for predicate logic uses Zorn's lemma, which is a variant of the axiom of choice (AC). It is known, however, that the completeness theorem does not require the

full strength of AC. The proof we give here uses the Baire category theorem, which is derivable in ZF with the addition of the axiom of dependent choices, a slightly weaker choice principle. (Exercise: can you see where in the proof we make use of a choice principle?)

THEOREM 6.3.1 (Baire category theorem) *Let X be a compact Hausdorff space, and let $U_1, U_2, \ldots$ be a countable family of sets, all of which are open and dense in X. Then $\bigcap_{i=1}^{\infty} U_i$ is dense in X.*

Proof Let $U = \bigcap_{i=1}^{\infty} U_i$, and let O be a nonempty open subset of X. We need only show that $O \cap U$ is nonempty. To this end, we inductively define a family O_i of open subsets of X as follows:

- $O_1 = O \cap U_1$, which is open, and nonempty since U_1 is dense;
- Assuming that O_n is open and nonempty, it has nonempty intersection with U_{n+1}, since the latter is dense. But any point $x \in O_n \cap U_{n+1}$ is contained in a neighborhood O_{n+1} such that $O_{n+1} \subseteq U_{n+1}$, and $\overline{O}_{n+1} \subseteq O_n$, using the regularity of X.

It follows then that the collection $\{\overline{O}_i : i \in \mathbb{N}\}$ satisfies the finite intersection property. Since X is compact, there is a p in $\bigcap_{i=1}^{\infty} \overline{O}_i$. Since $\overline{O}_{i+1} \subseteq O_i$, it also follows that $p \in O_i \subseteq U_i$, for all i. Therefore, $O \cap U$ is nonempty. □

Our proof of the completess theorem for predicate logic is similar in conception to the proof for propositional logic. First we construct a Boolean algebra B of provably-equivalent formulas. Using the definition of $\vdash$, it is not difficult to see that the equivalence relation is compatible with the Boolean operations. Thus, we may define Boolean operations as follows:

$$[\phi] \cap [\psi] = [\phi \wedge \psi], \qquad [\phi] \cup [\psi] = [\phi \vee \psi], \qquad -[\phi] = [\neg\phi].$$

If we let $0 = [\bot]$ and $1 = [\top]$, then it's easy to see that $\langle B, 0, 1, \cap, \cup, - \rangle$ is a Boolean algebra.

Now we want to show that if ϕ is not provably equivalent to a contradition, then there is a Σ-structure M such that $M(\phi)$ is not empty. In the case of propositional logic, it was enough to show that there is a homomorphism $f : B \to 2$ such that $f(\phi) = 1$. But that won't suffice for predicate logic, because once we have this homomorphism $f : B \to 2$, we need to use it to build a Σ-structure M, and to show that $M(\phi)$ is not empty. As we will now see, to ensure that $M(\phi)$ is not empty, we must choose a homomorphism $f : B \to 2$ that is "smooth on existentials."

DEFINITION 6.3.2 Let $f : B \to 2$ be a homomorphism. We say that f is **smooth on existentials** just in case for each formula ψ, if $f(\exists x \psi) = 1$, then $f(\psi[x_i/x]) = 1$ for some $i \in \mathbb{N}$.

We will see now that these "smooth on existentials" homomorphisms are dense in the Stone space X of B. In fact, the argument here is quite general. We first show that for any particular convergent family $a_i \to a$ in a Boolean algebra, the set of non-smooth

homomorphisms is closed and has empty interior. By saying that $a_i \to a$ is convergent, we mean that $a_i \leq a$ for all i, and for any $b \in B$, if $a_i \leq b$ for all i, then $a \leq b$. That is, a is the least upper bound of the a_i.

Let's say that a homomorphism $f : B \to 2$ is **smooth** relative to the convergent family $a_i \to a$ just in case $f(a_i) \to f(a)$ in the Boolean algebra 2. Now let D be the set of homomorphisms $f : B \to 2$ such that f is *not* smooth on $a_i \to a$. Any homomorphism $f : B \to 2$ preserves order, and hence $f(a_i) \leq f(a)$ for all i. Thus, if $f(a_i) = 1$ for any i, then f is smooth on $a_i \to a$. It follows that

$$D = E_a \cap \left[\bigcap_{i \in I} E_{\neg a_i} \right].$$

As an intersecton of closed sets, D is closed. To see that D has empty interior, suppose that $f \in E_b \subseteq D$, where E_b is a basic open subset of X. Then we have $E_b \subseteq E_{\neg a_i}$, which implies that $a_i \leq \neg b$; and since $a_i \leq a$, we have $a_i \leq a \wedge \neg b$. Thus, $a \wedge \neg b$ is an upper bound for the family $\{a_i\}$. Moreover, if $a = a \wedge \neg b$, then $a \wedge b = 0$ in contradiction with the fact that $f(a \wedge b) = 1$. Therefore, a is not the upper bound of $\{a_i\}$, a contradiction. We conclude that D contains no basic open subsets, and hence it has empty interior.

Now, this general result about smooth homomorphisms is of special importance for the Boolean algebra of equivalence classes of formulas. For in this case, existential formulas are the least upper bound of their instances.

LEMMA 6.3.3 *Let ϕ be a Σ-formula, and let I be the set of indices such that x_i does not occur free in ϕ. Then in the Lindenbaum algebra, $E_{(\exists x \phi)}$ is the least upper bound of $\{E_{(\phi[x_i/x])} \mid i \in I\}$.*

Proof For simplicity, set $E = E_{(\exists x \phi)}$ and $E_i = E_{(\phi[x_i/x])}$. The $\exists$-intro rule shows that $E_i \leq E$. Now suppose that $E_\psi \in B$ such that $E_i \leq E_\psi$ for all $i \in \mathbb{N}$. That is, $\phi[x_i/x] \vdash \psi$ for all $i \in I$. Since ϕ and ψ have a finite number of free variables, there is some $i \in I$ such that x_i does not occur free in ψ. By the $\exists$-elim rule, $\exists x_i \phi[x_i/x] \vdash \psi$. Since x_i does not occur free in ϕ, $\exists x_i \phi[x_i/x]$ is equivalent to $\exists x \phi$. Thus, $\exists x \phi \vdash \psi$, and $E \leq E_\psi$. Therefore, E is the least upper bound of $\{E_i \mid i \in I\}$. □

Thus, for each existential Σ-formula ϕ, the clopen set E_ϕ is the union of the clopen subsets corresponding to the instances of ϕ, plus the meager set D_ϕ of homomorphisms that are not smooth relative to ϕ. Since the signature Σ is countable, there are countably many such existential formulas, and countably many of these sets D_ϕ of non-smooth homomorphisms. Since each D_ϕ is meager, the Baire category theorem entails that their union also is meager. Thus, the set U of homomorphisms that are smooth on *all* existentials is open and dense in the Stone space X.

We are now ready to continue with the completeness theorem. Let ϕ be our arbitrary formula that is not provably equivalent to a contradiction. We know that the set E_ϕ of homomorphisms $f : B \to 2$ such that $f([\phi]) = 1$ is open and nonempty. Hence, E_ϕ has nonempty intersection with U. Let $f \in E_\phi \cap U$. That is, $f([\phi]) = 1$, and f is smooth on all existentials. We now use f to define a Σ-structure M.

- Let the domain S of M be the set of natural numbers.
- For an n-ary relation symbol $R \in \Sigma$, let $\vec{a} \in M(R)$ if and only if $f(R(x_{a_1}, \ldots, x_{a_n})) = 1$.

LEMMA 6.3.4 *For any Σ-formula ϕ with canonical context $x_{c_1}, \ldots, x_{c_n}$, if $f(\phi) = 1$, then $\vec{c} \in M(\phi)$.*

Proof We prove this result by induction on the construction of ϕ. Note that an n-tuple $\vec{c}$ of natural numbers corresponds to a unique function $c : [n] \to \mathbb{N}$. Supposing that we are given a fixed enumeration $x_1, x_2, \ldots$ of the variables of Σ, each such function c also corresponds to an n-tuple $x_{c_1}, \ldots, x_{c_n}$, possibly with duplicate variables. Since each formula ϕ determines a canonical context (without duplicates), ϕ also determines an injection $a : [n] \to \mathbb{N}$. For any other function $c : [n] \to \mathbb{N}$, we let ϕ_c denote the result of replacing all free occurences of x_{a_i} in ϕ with x_{c_i}.

1. Suppose that $\phi \equiv R(x_{a_1}, \ldots, x_{a_m})$, and let $x_{c_1}, \ldots, x_{c_n}$ be the canonical context of ϕ. Thus, for each $i = 1, \ldots, m$, there is a $p(i)$ such that $x_{a_i} = x_{c_{p(i)}}$. Now, $M(\phi)$ is defined to be the pullback of $M(R)$ along π_p. Since $\pi_i \pi_p(\vec{c}) = c_{p(i)} = a_i$ and $\vec{a} \in M(R)$, it follows that $\vec{c} \in M(\phi)$.
2. Suppose that the result is true for ϕ and ψ, and suppose that $f(\phi \wedge \psi) = 1$. Let $\vec{x} = x_{c_1}, \ldots, x_{c_n}$ be the canonical context of $\phi \wedge \psi$. The context of ϕ is a subsequence of $\vec{x}$, i.e., it is of the form $x_{c_{p(1)}}, \ldots, x_{c_{p(m)}}$ where $p : [m] \to [n]$ is an injection. If $\pi_p : S^n \to S^m$ is the corresponding projection, then

$$\pi_p(\vec{c}) = \langle c_{p(1)}, \ldots, c_{p(m)} \rangle.$$

   Similarly, if $x_{c_{q(1)}}, \ldots, x_{c_{q(\ell)}}$ is the context of ψ, then

$$\pi_q(\vec{c}) = \langle c_{q(1)}, \ldots, c_{q(\ell)} \rangle.$$

   Since $f(\phi) = 1 = f(\psi)$, the inductive hypothesis entails that $\pi_p(\vec{c}) \in M(\phi)$ and $\pi_q(\vec{c}) \in M(\psi)$. By definition, $M(\phi \wedge \psi) = \pi_p^*(M(\phi)) \cap \pi_q^*(M(\psi))$, hence $\vec{c} \in M(\phi \wedge \psi)$ iff $\pi_p(\vec{c}) \in M(\phi)$ and $\pi_q(\vec{c}) \in M(\psi)$.
3. Suppose that $\phi \equiv \exists x_k \psi$, and that the result is true for ψ, as well as for any ψ' that results from uniform replacement of free variables in ψ. Suppose first that x_k is free in ψ. For notational simplicity, we will assume that x_k is the last variable in the canonical context for ψ. Thus, if the context for ϕ is $x_{c_1}, \ldots, x_{c_n}$, then the context for ψ is $x_{c_1}, \ldots, x_{c_n}, x_k$. (In the case where ϕ is a sentence, i.e., $n = 0$, the string $\vec{c}$ is empty.)

   Now suppose that $f(\exists x_k \psi) = 1$. Since f is smooth on existentials, there is a $j \in \mathbb{N}$ such that x_j is not free in ψ, and $f(\psi[x_j/x_k]) = 1$. The context of $\psi[x_j/x_k]$ is $x_{c_1}, \ldots, x_{c_n}, x_j$, and the inductive hypothesis entails that $\vec{c}, j \in M(\psi[x_j/x_k])$. By the definition of $M(\exists x_k \psi)$, if $\vec{c}, j \in M(\psi[x_j/x_k])$, then $\vec{c} = \pi(\vec{c}, j) \in M(\exists x_k \psi)$.

The remaining inductive steps are similar to the preceding steps, and are left to the reader. □

This lemma concludes the proof of the completeness theorem, and immediately yields two other important model-theoretic results.

THEOREM 6.3.5 (Downward Löwenheim–Skølem) *Let Σ be an countable signature, and let ϕ be a Σ-sentence. If ϕ has a model, then ϕ also has a countable model.*

Proof If ϕ has a model, then, by the soundness theorem, ϕ is not provably equivalent to a contradiction. Thus, by the completeness theorem, ϕ has a model whose domain is the natural numbers. □

DISCUSSION 6.3.6 The downward Löwenheim–Skølem theorem does not hold for arbitrary sets of sentences in uncountable signatures. Indeed, let $\Sigma = \{c_r \mid r \in \mathbb{R}\}$, and let T be the theory with axioms $c_r \neq c_s$ when $r \neq s$. Then T has a model (for example, the real numbers $\mathbb{R}$) but no countable model.

The Löwenheim–Skølem theorem has sometimes been thought to be paradoxical, particularly in application to the case where T is the theory of sets. The theory of sets implies a sentence ϕ whose intended interpretation is, "there is an uncountable set." The LS theorem implies that if T has any model, then it has a countable model M, and hence that $\models_M \phi$. In other words, there is a countable model M that makes true the sentence, "there is an uncountable set."

THEOREM 6.3.7 (Compactness) *Suppose that T is a set of Σ-sentences. If each finite subset of T has a model, then T has a model.*

It would be nice to be able to understand the compactness theorem for predicate logic directly in terms of the compactness of the Stone space of the Lindenbaum algebra. However, this Stone space isn't exactly the space of Σ-structures, and so its compactness isn't the same thing as compactness in the logical sense. We could indeed use each point $f \in X$ to define a Σ-structure M; but, in general, $f(\phi) = 1$ wouldn't entail that $M(\phi) = 1$. What's more, there are additional Σ-structures that are not represented by points in X, in particular, Σ-structures with uncountably infinite domains. Thus, we are forced to turn to a less direct proof of the compactness theorem.

Proof We first modify the proof of the completeness theorem by constructing the Boolean algebra B_T of equivalence classes of formulas modulo T-provable equivalence. This strengthened completeness theorem shows that if $T \models \phi$, then $T \vdash \phi$. However, if $T \vdash \phi$, then $T_0 \vdash \phi$ for some finite subset T_0 of T. □

DISCUSSION 6.3.8 The compactness theorem yields all sorts of surprises. For example, it shows that there is a model that satisfies all of the axioms of the natural numbers, but which has a number greater than all natural numbers. Let Σ consist of a signature for arithmetic and one additional constant symbol c. We assume that Σ has a name n for each natural number. Now let

$$T = Th(\mathbb{N}) \cup \{n < c \mid n \in \mathbb{N}\},$$

where $Th(\mathbb{N})$ consists of all Σ-sentences true in $\mathbb{N}$. It's easy to see that each finite subset of T of consistent. Therefore, by compactness, T has a model M. In the model M, $n^M < c^M$ for all $n \in \mathbb{N}$.

6.4 Categories of Models

There are many interesting categories of mathematical objects such as sets, groups, topological spaces, smooth manifolds, rings, etc. Some of these categories are of special interest for the empirical sciences, as the objects in thos categories are the "models" of a scientific theory. For example, a model of Einstein's general theory of relativity (GTR) is a smooth manifold with Lorentzian metric. Hence, the mathematical part of GTR can be considered to be some particular category of manifolds. (The choice of arrows for this category of models raises interesting theoretical questions. See, e.g., Fewster [2015].) Similarly, since a model of quantum theory is a complex vector space equipped with some particular dynamical evolution, the mathematical part of quantum theory can be considered to be some category of vector spaces.

Philosophers of science want to talk about real-life scientific theories – not imaginary theories that can be axiomatized in first-order logic. Nonetheless, we can benefit tremendously from considering tractable formal analogies, what scientists themselves would call "toy models." In this section, we pursue an analogy between models of a scientific theory and models of a first-order theory T. In particular, we show that any first-order theory T has a category $\mathrm{Mod}(T)$ of models, and intertranslatable theories have equivalent categories of models. Thus, we can think of the 2-category of all categories of models of first-order theories as a formal analogy to the universe of all scientific theories.

There are two natural definitions of arrows in the category $\mathrm{Mod}(T)$, one more liberal (homomorphism) and another more conservative (elementary embedding).

DEFINITION 6.4.1 Let Σ be a fixed signature, and let M and N be Σ-structures. We will use X and Y to denote their respective domain sets. A Σ-**homomorphism** $h : M \to N$ consists of a function $h : X \to Y$ that satisfies the following:

1. For each relation symbol $R \in \Sigma$, there is a commutative diagram:

$$\begin{array}{ccc} MR & \longrightarrow & NR \\ \downarrow & & \downarrow \\ X^n & \xrightarrow{h^n} & Y^n \end{array}$$

   Here the arrows $MR \rightarrowtail X^n$ and $NR \rightarrowtail Y^n$ are the subset inclusions, and $h^n : X^n \to Y^n$ is the map defined by $h^n\langle a_1, \ldots, a_n\rangle = \langle h(a_1), \ldots, h(a_n)\rangle$. The fact that the diagram commutes says that for any $\langle a_1, \ldots, a_n\rangle \in MR$, we have $\langle h(a_1), \ldots, h(a_n)\rangle \in NR$.
2. For each function symbol $f \in \Sigma$, the following diagram commutes:

$$\begin{array}{ccc} X^n & \xrightarrow{h^n} & Y^n \\ \downarrow{\scriptstyle Mf} & & \downarrow{\scriptstyle Nf} \\ X & \xrightarrow{h} & Y \end{array}$$

In other words, for each $\langle a_1, \ldots, a_n \rangle \in X^n$, we have $h(M(f)\langle a_1, \ldots, a_n \rangle) = N(f)\langle h(a_1), \ldots, h(a_n) \rangle$. When c is a constant symbol, this condition implies that $h(c^M) = c^N$.

DEFINITION 6.4.2 Let M and N be Σ-structures, and let $h : M \to N$ be a homomorphism. We say that h is a **Σ-elementary embedding** just in case for each Σ-formula ϕ, the following is a pullback diagram:

$$\begin{array}{ccc} M(\phi) & \longrightarrow & N(\phi) \\ \downarrow & & \downarrow \\ X^n & \xrightarrow{h^n} & Y^n \end{array}$$

In other words, for all $\vec{a} \in X^n$, $\vec{a} \in M(\phi)$ iff $h(\vec{a}) \in N(\phi)$. In particular, for the case where ϕ is a sentence, the following is a pullback:

$$\begin{array}{ccc} M(\phi) & \longrightarrow & N(\phi) \\ \downarrow & & \downarrow \\ 1 & \longrightarrow & 1 \end{array}$$

which means that $M \models \phi$ iff $N \models \phi$.

EXERCISE 6.4.3 Show that the composite of elementary embeddings is an elementary embedding.

Note that the conditions for being an elementary embedding are quite strict. For example, let ϕ be the sentence that says there are exactly n things. If $h : M \to N$ is an elementary embedding, then $M \models \phi$ iff $N \models \phi$. Thus, if the domain X of M has cardinality $n < \infty$, then Y also has cardinality $n < \infty$. Suppose, for example, that T is the theory of groups. Then for any two finite groups G, H, there is an elementary embedding $h : G \to H$ only if $|G| = |H|$. Therefore, the notion of elementary embedding is stricter than the notion of a group homomorphism.

Similarly, let Σ be the empty signature. Let M be a Σ-structure with one element, and let N be a Σ-structure with two elements. Then any mapping $h : M \to N$ is a homomorphism, since Σ is empty. However, $\models_M x = y$ but $\not\models_N x = y$. Therefore, there is no elementary embedding $h : M \to N$.

The strictness of elementary embeddings leads to a little dilemma in choosing arrows in our definition of the category $\mathrm{Mod}(T)$ of models of a theory T. Do we choose homomorphisms between models, of which there are relatively many, or do we choose elementary embeddings, of which there are relatively few? We have opted to play it safe.

DEFINITION 6.4.4 We henceforth use $\mathrm{Mod}(T)$ to denote the category whose objects are models of T, and whose arrows are elementary embeddings between models. As with any category, we say that an arrow $f : M \to N$ in $\mathrm{Mod}(T)$ is an isomorphism just in case there is an arrow $g : N \to M$ such that $g \circ f = 1_M$ and $f \circ g = 1_N$. In this particular case, we say that f is a **Σ-isomorphism**.

There is a clear sense in which elementary embeddings between models of T are structure that is definable in terms of T. In short, elementary embeddings between models should be considered to be part of the semantic content of the theory T. Accordingly, formally equivalent theories ought at least to have equivalent categories of models. We elevate this idea to a definition.

DEFINITION 6.4.5 Let T and T' be theories, not necessarily in the same signature. We say that T and T' are **categorically equivalent** just in case the categories Mod(T) and Mod(T') are equivalent.

Notice that if we had chosen all homomorphisms as arrows, then Mod(T) would have more structure, and it would be more difficult for the categories Mod(T) and Mod(T') to be equivalent. In fact, there are theories T and T' that most mathematicians would consider to be equivalent, but which this criterion would judge to be inequivalent.

DEFINITION 6.4.6 If M is a Σ-structure, we let $Th(M)$ denote the theory consisting of all Σ-sentences ϕ such that $M \models \phi$.

DEFINITION 6.4.7 Let M and N be Σ-structures. We say that M and N are **elementarily equivalent**, written $M \equiv N$, just in case $Th(M) = Th(N)$.

EXERCISE 6.4.8 Show that if $h : M \to N$ is an isomorphism, then M and N are elementarily equivalent.

The converse to this exercise is not true. For example, let T be the empty theory in the signature $\{=\}$. Then for each cardinal number κ, T has a model M with cardinality κ; and if M and N are infinite models of T, then M and N are elementarily equivalent. (The signature $\{=\}$ has no formulas that can discriminate between two different infinite models.) Thus, T has models that are elementarily equivalent but not isomorphic.

6.5 Ultraproducts

The so-called ultraproduct construction is often considered to be a technical device for proving theorems. Here we will emphasize the structural features of ultraproducts, rather than the details of the construction. Note, however, that ultraproducts are not themselves limits or colimits in the sense of category theory. Thus, we cannot give a simple formula relating an ultraproduct to the models from which it is constructed. In one sense, ultraproducts are more like limits in the topological sense than they are in the category-theoretic sense. Indeed, in the case of propositional theories, the ultraproduct of models of a theory *is* the topological limit in the Stone space of the theory.

To see this, it helps to redescribe limits in a topological space X in terms of infinitary operations $X^\infty \to X$. Recall that a point $p \in X$ is said to be a limit point of a subset $A \subseteq X$ just in case every open neighborhood of p intersects A. When X is nice enough (e.g., second countable), these limit points can be detected by sequences. That is, in such cases, p is a limit point of A just in case there is a sequence $a_1, a_2, \ldots$ of elements in A

such that $\lim_i a_i = p$. This last equation is simply shorthand for the statement: for each neighborhood U of p, there is a $n \in \mathbb{N}$ such that $a_i \in U$ for all $i \geq n$.

Suppose now, more specifically, that X is a compact Hausdorff space. Consider the product $\prod_{i\in\mathbb{N}} X$, which consists of infinite sequences of elements of X. We can alternately think of elements of $\prod_{i\in\mathbb{N}} X$ as functions from $\mathbb{N}$ to X. Since $\mathbb{N}$ is discrete, every such function $f : \mathbb{N} \to X$ is continuous, i.e., f^{-1} maps open subsets of X to (open) subsets of $\mathbb{N}$. Of course, f^{-1} also preserves inclusions of subsets. Hence, for each filter $\mathscr{V}$ of open subsets of X, $f^{-1}(\mathscr{V})$ is a filter on $\mathbb{N}$. For each point $p \in X$, let $\mathscr{V}_p$ be the filter of open neighborhoods of p. Now, for each ultrafilter $\mathscr{U}$ on $\mathbb{N}$, we define an operation $\lim_{\mathscr{U}} : \prod_{i\in\mathbb{N}} X \to X$ by the following condition:

$$\lim_{\mathscr{U}} f = p \quad \Longleftrightarrow \quad f^{-1}(\mathscr{V}_p) \subseteq \mathscr{U}.$$

To show that this definition makes sense, we need to check that there is a unique p satisfying the condition on the right. For uniqueness, suppose that $f^{-1}(\mathscr{V}_p)$ and $f^{-1}(\mathscr{V}_q)$ are both contained in $\mathscr{U}$. If $p \neq q$, then there are $U \in \mathscr{V}_p$ and $V \in \mathscr{V}_q$ such that $U \cap V$ is empty. Then $f^{-1}(U) \cap f^{-1}(V)$ is empty, in contradiction with the fact that $\mathscr{U}$ is an ultrafilter. For existence, suppose first that $\mathscr{U}$ is a principal ultrafilter – i.e., contains all sets containing some $n \in \mathbb{N}$. Let $p = f(n)$. Then for each neighborhood V of p, $f^{-1}(V)$ contains $f(n)$, and hence is contained in $\mathscr{U}$. Suppose now that $\mathscr{U}$ is non-principal, hence contains the cofinite filter. Since X is Hausdorff, the sequence $f(1), f(2), \ldots$ has a limit point p. Thus, for each $V \in \mathscr{V}_p$, $f^{-1}(V)$ is a cofinite subset of $\mathbb{N}$, and hence is contained in $\mathscr{U}$. In either case, there is a $p \in X$ such that $f^{-1}(\mathscr{V}_p) \subseteq \mathscr{U}$.

Thus, the topological structure on a compact Hausdorff space X can be described in terms of a family of operations $\lim_{\mathscr{U}} : \prod_i X_i \to X$, where $\mathscr{U}$ runs through all the ultrafilters on $\mathbb{N}$. This result holds in particular when $X = \mathrm{Mod}(T)$ is the Stone space of models of a propositional theory. A limit model $\lim_{\mathscr{U}} M_i$ is called an **ultraproduct** of the models M_i. Thus, in the propositional case, an ultraproduct of models is simply the limit relative to the Stone space topology.

We will now try to carry over this intuition to the case of general first-order theories, modifying details when necessary. To begin with, if T is a first-order theory $\mathrm{Mod}(T)$ is too large to have a topology – it is a class and not a set. What's more, even if we pretend that $\mathrm{Mod}(T)$ is a set, the ultraproduct construction couldn't be expected to yield a topology, but something like a "pseudo-topology" or "weak topology," where limits are defined only up to isomorphism.

The details of the ultraproduct construction run as follows. Let I be an index set, and suppose that for each $i \in I$, M_i is a Σ-structure. If $\mathscr{U}$ is an ultrafilter on I, then we define a Σ-structure $N := \lim_{\mathscr{U}} M_i$ as follows:

- First consider the set $\prod M_i$ of "sequences," where each $a_i \in M_i$. We say that two such sequences are equivalent if they eventually agree in the sense of the ultrafilter $\mathscr{U}$. That is, $(a_i) \sim (b_i)$ just in case $\{i \mid a_i = b_i\}$ is contained in $\mathscr{U}$. We let the domain of N be the quotient of $\prod M_i$ under this equivalence relation.
- For each relation symbol R of Σ, we let $N(R)$ consist of sequences on n-tuples that eventually lie in $M_i(R)$ in the sense of the ultrafilter $\mathscr{U}$. That is, $(a_i) \in N(R)$

just in case $\{i \mid a_i \in M_i(R)\}$ is contained in $\mathscr{U}$. (Here one uses the fact that $\mathscr{U}$ is a ultrafilter to prove that $N(R)$ is well-defined as a subset of N.)

The resulting model $\lim_{\mathscr{U}} M_i$ is said to be an **ultraproduct** of the models M_i. In the special case where each M_i is the same M, we call $\lim_{\mathscr{U}} M_i$ an **ultrapower** of M. In this case, there is a natural elementary embedding $h : M \to \lim_{\mathscr{U}} M_i$ that maps each $a \in M$ to the constant sequence $a, a, \ldots$

We now cite without proof a fundamental theorem for ultraproducts.

THEOREM 6.5.1 (Łos) *Let $\{M_i \mid i \in I\}$ be a family of Σ-structures, let $\mathscr{U}$ be an ultrafilter on I, and let $N = \lim_{\mathscr{U}} M_i$. Then for each Σ-sentence ϕ, $N \models \phi$ iff $\{i \mid M_i \models \phi\} \in \mathscr{U}$.*

Intuitively speaking, $\lim_{\mathscr{U}} M_i$ satisfies exactly those sentences that are eventually validated by M_i as i runs through the ultrafilter $\mathscr{U}$.

We saw before that elementarily equivalent models need not be isomorphic. Indeed, for M and N to be elementarily equivalent, it's sufficient that there is a third model L and elementary embeddings $h : M \to L$ and $j : N \to L$. The following result shows that this condition is necessary as well.

PROPOSITION 6.5.2 *Let M and N be Σ-structures. Then the following are equivalent.*

1. *$M \equiv N$, i.e., M and N are elementarily equivalent.*
2. *There is a Σ-structure L and elementary embeddings $h : M \to L$ and $j : N \to L$.*
3. *M and N have isomorphic ultrapowers.*

Sketch of proof (3 $\Rightarrow$ 2) Suppose that $j : \lim_{\mathscr{U}_1} M \to \lim_{\mathscr{U}_2} N$ is an isomorphism, and let $L = \lim_{\mathscr{U}_2} N$. Let $h : M \to \lim_{\mathscr{U}_1} M$ be the natural embedding, and similarly for $k : N \to \lim_{\mathscr{U}_2} N$. Then $j \circ h : M \to L$ and $k : N \to L$ are elementary embeddings.

(2 $\Rightarrow$ 1) Since elementary embeddings preserve truth-values of sentences, this result follows immediately.

(1 $\Rightarrow$ 3) This is a difficult result, known as the Keisler–Shelah isomorphism theorem. We omit the proof and refer the reader to Keisler (2010) for further discussion. □

6.6 Relations between Theories

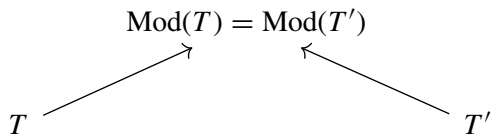
In the previous two chapters, we analyzed theories through a syntactic lens. Thus, to explicate relations between theories – such as equivalence and reduction – we used a syntactic notion, viz. translation. In this chapter, we've taken up the semantic analysis of theories – i.e., thinking about theories in terms of their models. Accordingly, we would like to investigate precise technical relations between categories of models that correspond with our intuitive notions of the relations that can hold between theories. In the best-case scenario, the technical notions we investigate will be useful in honing our intuitions about specific, real-life cases.

This investigation takes on special philosophical significance when we remember that at a few crucial junctures, philosophers claimed a decisive advantage for semantic analyses of relations between theories. Let's recall just a couple of the most prominent such maneuvers.

- van Fraassen (1980) claims that while the empirical content of a theory cannot be isolated syntactically, it can be isolated semantically. Since the notion of empirical content is essential for empiricism, van Fraassen thinks that empiricism requires the semantic view of theories.
- Defenders of various dressed-up versions of physicalism claim that the mental–physical relationship cannot be explicated syntactically, but can be explicated semantically. For example, the non-reductive physicalists of the 1970s claimed that the mental isn't reducible (syntactically) to the physical, but it does supervene (semanatically) on the physical. Similarly, Bickle (1998) claims that the failure of mind–brain reduction can be blamed on the syntactic explication of reduction, and that the problems can be solved by using a semantic explication of reduction.

These claims give philosophers a good reason to investigate the resources of logical semantics.

Let's begin by setting aside some rather flat-footed attempts to use semantics to explicate relations between theories. In particular, there seems to be a common misconception that the models of a theory are language-free, and can provide the standard by which to decide questions of theoretical equivalence. The (mistaken) picture here is that two theories, T and T', in different languages, are equivalent just in case $\mathrm{Mod}(T) = \mathrm{Mod}(T')$. We can illustrate this idea with a picture:



The picture here is that the theory formulations T and T' are language-bound, but the class $\mathrm{Mod}(T) = \mathrm{Mod}(T')$ of models is a sort of thing-in-itself that these different formulations intend to pick out.

If you remember that models are mappings from signatures, then you realize that there is something wrong with this picture. Yes, there are categories $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T')$, but these categories are no more language-independent than the syntactic objects T and T'. In particular, if Σ and Σ' are different signatures, then there is no standard by which one can compare $\mathrm{Mod}(T)$ with $\mathrm{Mod}(T')$. A model of T is a function from Σ to **Sets**, and a model of T' is a function from Σ' to **Sets**. Functions with different domains cannot be equal; but it would also be misleading to say that they are *unequal*. In the world of sets, judgments of equality and inequality only make sense for things that live in the same set.

In a similar fashion, we can't make any progress in analyzing the relations between T and T' by setting their models side by side. One occasionally hears philosophers of science say things like

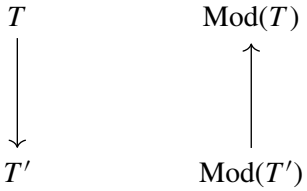
(I) There is a model of T that is not isomorphic to any model of T'; hence, T and T' are not equivalent.

(E) If T is a subtheory of T', then each model of T can be embedded in a model of T'.

However, if T and T' are theories in different signatures, then neither I nor E makes sense. The notions of isomorphism and elementary embedding are signature-relative: a function $h : M \to N$ is an elementary embedding just in case $h(M(\phi)) = N(\phi)$ for each Σ-formula ϕ. If T and T' are written in different signatures, then there is simply no way to compare a model M of T directly with a model N of T'. (And this lesson goes not only for theories in first-order logic, but for any mathematically formalized scientific theory – such as quantum mechanics, general relativity, Hamiltonian mechanics, etc.)

With these flat-footed analyses set aside, we can now raise some serious questions about the relations between Mod(T) and Mod(T'). For example, what mathematical relation between Mod(T) and Mod(T') would be a good explication of the idea that T is equivalent to T'? Is it enough that Mod(T) and Mod(T') are equivalent categories, or should we require something more? Similarly, what mathematical relation between Mod(T) and Mod(T') would be a good explication of the idea that T is reducible to T'? Finally, to return to the issue of empiricism, can the empirical content of a theory T be identified with some structure inside the category Mod(T)?

We will approach these questions from two directions. Our first approach will involve attempting to transfer notions from the syntactic side to the semantic side, as in the following picture:



You will have noticed that we already followed this approach in Chapter 3, with respect to propositional theories. The goal is to take a syntactic relation between theories (such as "being reducible to") and to translate it over to a semantic relation between the models of those theories.

Of course, this first approach won't be at all satisfying to those who would be free from the "shackles of language." Thus, our second angle of attack is to ask directly about relations between Mod(T) and Mod(T'). Where do Mod(T) and Mod(T') live in the mathematical universe, and what are the mathematical relationships between them? Again, it will be no surprise to you that we think Mod(T) should, at the very least, be considered to be a *category*, whose mathematical structure includes not only models, but also arrows between them. Moreover, once we equip Mod(T) with sufficient structure, we will see that these two approaches converge – i.e., that the most interesting relations between Mod(T) and Mod(T') are those that correspond to some syntactic relation between T and T'. It is in this sense that logical semantics is *dual* to logical syntax.

We begin then with the first approach and, in particular, with showing that each translation $F : T \to T'$ gives rise to a functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$. We will also provide a partial translation manual between properties of the translation F and properties of the functor F^*. To the extent that such a translation manual exists, each syntactic relation between T and T' corresponds to a unique semantic relation between $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T')$, and vice versa.

DEFINITION 6.6.1 Suppose that $F : T \to T'$ is a translation, and let M be a model of T'. We define a Σ-structure F^*M as follows:

- Let F^*M have the same domain as M.
- For each relation symbol r of Σ, let
$$(F^*M)(r) = M(Fr).$$
- For each function symbol f of Σ, let $(F^*M)(f)$ be the function with graph $M(Ff)$.

We will now show that $(F^*M)(\phi) = M(F\phi)$ for each Σ-formula ϕ. However, we first need an auxiliary lemma. For this, recall that if $f : X \to Y$ is a function, then its graph is the subset $\{\langle x, f(x)\rangle \mid x \in X\}$ of $X \times Y$.

LEMMA 6.6.2 *For each Σ-term t, $M(Ft)$ is the graph of the function $(F^*M)(t)$.*

Proof We prove this by induction on the construction of t. Recall that if t is a term with n free variables, then Ft is a formula with $n+1$ free variables, and $M(Ft)$ is a subset of S^{n+1}.

- Suppose that $t \equiv x$. Then $Ft \equiv (x = y)$ for some variable $y \not\equiv x$. In this case, $M(Ft)$ is the diagonal of $S \times S$, which is the graph of $1_S = (F^*M)(x)$.
- Now suppose that the result is true for $t_1, \ldots, t_m$, and let $t \equiv f(t_1, \ldots, t_m)$. Recall that Ft is defined as the composite of the relation Ff with the relations $Ft_1, \ldots, Ft_m$. Since M preserves the relevant logical structures, $M(Ft)$ is the composite of the relation $M(Ff)$ with the relations $M(Ft_1), \ldots, M(Ft_m)$. Moreover, $(F^*M)(t)$ is defined to be the composite of the function $(F^*M)(f)$ with the functions $(F^*M)(t_1), \ldots, (F^*M)(t_m)$. In general, the graph of a composite function is the composite of the graphs. Therefore, $M(Ft)$ is the graph of $(F^*M)(t)$.

□

PROPOSITION 6.6.3 *For each Σ-formula ϕ, $(F^*M)(\phi) = M(F\phi)$.*

Proof We prove this by induction on the construction of ϕ.

- Suppose that $\phi \equiv (t_1 = t_2)$. Then $F\phi$ is the formula $\exists y(Ft_1(\vec{x}, y) \wedge Ft_2(\vec{x}, y))$. Here, for simplicity, we write $\vec{x}$ for the canonical context of $F\phi$. Thus, $M(F\phi)$ consists of elements $\vec{a} \in S^n$ such that $\langle \vec{a}, b\rangle \in M(Ft_1)$ and $\langle \vec{a}, b\rangle \in M(Ft_2)$ for some $b \in S$. By the previous lemma, $M(Ft_i)$ is the graph of $(M^*F)(t_i)$. Thus, $M(F\phi)$ is the equalizer of $(M^*F)(t_1)$ and $(M^*F)(t_2)$. That is, $M(F\phi) = (M^*F)(\phi)$.

- Suppose that $\phi \equiv p(t_1, \ldots, t_m)$. Then $F\phi$ is the formula

$$\exists z_1 \cdots \exists z_m (Fp(y_1, \ldots, y_m) \wedge Ft_1(\vec{x}, y_1) \wedge \cdots \wedge Ft_m(\vec{x}, y_m)).$$

  Hence $M(F\phi)$ consists of those $\vec{a} \in S^n$ such that there are $b_1, \ldots, b_m \in S$ with $\langle \vec{a}, b_i \rangle \in M(Ft_i)$ and $\vec{b} \in M(Fp)$. By the previous lemma, $M(Ft_i)$ is the graph of $(F^*M)(t_i)$. Hence, $M(F\phi)$ consists of those $\vec{a} \in S^n$ such that

$$\langle (F^*M)(t_1), \ldots, (F^*M)(t_m) \rangle (\vec{a}) \in M(Fp) = (F^*M)(p).$$

  In other words, $M(F\phi) = (F^*M)(\phi)$. (Here we have ignored the fact that the terms $t_1, \ldots, t_m$ might have different free variables. In that case, we need simply to prefix the $(F^*M)(t_i)$ with the appropriate projections to represent them on the same domain S^n.)
- Suppose that $\phi \equiv (\phi_1 \wedge \phi_2)$, and the result is true for ϕ_1 and ϕ_2. Now, $F(\phi_1 \wedge \phi_2) = F\phi_1 \wedge F\phi_2$. Hence $M(F(\phi_1 \wedge \phi_2))$ is the pullback of $M(F\phi_1)$ and $M(F\phi_2)$ along the relevant projections (determined by the contexts of ϕ_1 and ϕ_2). Since F preserves contexts of formulas, and $M(F\phi_i) = (F^*M)(\phi_i)$, it follows that $M(F(\phi_1 \wedge \phi_2)) = (F^*M)(\phi_1 \wedge \phi_2)$.
- We now deal with the existential quantifier. For simplicity, suppose that ϕ has free variables x and y. We suppose that the result is true for ϕ; that is,

$$(F^*M)(\phi) = M(F(\phi)),$$

  and we show that

$$(F^*M)(\exists x \phi) = M(F(\exists x \phi)).$$

  By definition, $(F^*M)(\exists x \phi)$ is the image of $(F^*M)(\phi)$ under the projection $\pi : X \times Y \to Y$. Moreover, $F(\exists x \phi) = \exists x F(\phi)$, which means that $M(F(\exists x \phi))$ is the image of $M(F(\phi))$ under the projection π.

□

PROPOSITION 6.6.4 *Suppose that $F : T \to T'$ is a translation. If M is a model of T' then F^*M is a model of T.*

Proof Suppose that $T \vdash \phi$. Since F is a translation, $T' \vdash F\phi$. Since M is a model of T', $M(F\phi) = S^n$. Therefore, $(F^*M)(\phi) = S^n$. Since ϕ was an arbitrary Σ-formula, we conclude that F^*M is a model of T. □

DEFINITION 6.6.5 Let $F : T \to T'$ be a translation. We now extend the action of F^* from models of T' to elementary embeddings between these models. Let M and N be models of T' with corresponding domains X and Y. Let $h : M \to N$ be an elementary embedding. Since F^*M has the same domain as M, and similarly for F^*N and N, this h is a candidate for being an elementary embedding of F^*M into F^*N. We need only check that the condition of Defn. 6.4.2 holds – i.e., that for each Σ-formula ϕ, the following diagram is a pullback:

$$\begin{array}{ccc} (F^*M)(\phi) & \longrightarrow & (F^*N)(\phi) \\ \downarrow & & \downarrow \\ X^n & \xrightarrow{h^n} & Y^n \end{array}$$

But $(F^*M)(\phi) = M(F\phi)$ and $(F^*N)(\phi) = N(F\phi)$. Since $h : M \to N$ is elementary, the corresponding diagram is a pullback. Therefore, $h : F^*M \to F^*N$ is elementary.

Now, the underlying function of $F^*h : F^*M \to F^*N$ is the same as the underlying function of $h : M \to N$. Thus, F^* preserves composition of functions, as well as identity functions; and $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is a functor.

We have shown that each translation $F : T \to T'$ corresponds to a functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$. Now we would like to compare properties of F with properties of F^*. The fundamental result here is that if F is a homotopy equivalence, then F^* is an equivalence of categories.

PROPOSITION 6.6.6 *If T and T' are intertranslatable, then* $\mathrm{Mod}(T)$ *and* $\mathrm{Mod}(T')$ *are equivalent categories. In particular, if $F : T \to T'$ and $G : T' \to T$ form a homotopy equivalence, then F^* and G^* are inverse functors.*

Sketch of proof In the following chapter, we prove a stronger result: if T and T' are Morita equivalent, then $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T')$ are equivalent categories. In order to avoid duplicating work, we will just sketch the proof here. One shows that $(FG)^* = G^*F^*$, and that for any two translations F and G, if $F \simeq G$, then $F^* = G^*$. Since $GF \simeq 1_T$, it follows that

$$F^*G^* = (GF)^* = 1_T^* = 1_{\mathrm{Mod}(T)}.$$

Similarly, $G^*F^* = 1_{\mathrm{Mod}(T')}$, and therefore $\mathrm{Mod}(T)$ and $\mathrm{Mod}(T')$ are equivalent categories. □

COROLLARY 6.6.7 *If T and T' are intertranslatable, then T and T' are categorically equivalent.*

One upshot of this result is that categorical properties of $\mathrm{Mod}(T)$ are **invariants** of intertranslatability. For example, if $\mathrm{Mod}(T)$ has all finite products and $\mathrm{Mod}(T')$ does not, then T and T' are not intertranslatable. We should think a bit, then, about which features of a category are invariant under categorical equivalence.

Recall that the identity of a category $\mathbf{C}$ has nothing to do with the identity of its objects. All that matters is the relations that these objects have to each other. Thus, if we look at $\mathrm{Mod}(T)$ qua category, then we are forgetting that its objects are models. Instead, we are focusing exclusively on the arrows (elementary embeddings) that relate these models, including the symmetries (automorphisms) of models. Here are some of the properties that can be expressed in the language of category theory:

1. $\mathbf{C}$ has products.
2. $\mathbf{C}$ has coproducts.
3. $\mathbf{C}$ has all small limits.

The list could go on, but the real challenge is to say which of the properties of the category Mod(T) corresponds to an interesting feature of the theory T. For example, might it be relevant that Mod(T) has products – i.e., that for any two models M, N of T, there is a model $M \times N$, with the relevant projections, etc.? Keep in mind that these mathematical statements don't have an obvious interpretation in terms of what the theory T might be saying about the world. For example, to say that Mod(T) has products doesn't tell us that there is an operation that takes two possible worlds and returns another possible world.

Recall, in addition, that category theorists ignore properties that are not invariant under categorical equivalence. For example, the property "**C** has exactly two objects," is not invariant under all categorical equivalences. Although the notion of a "categorical property" is somewhat vague, the practicing category theorist knows it when he sees it – and fortunately, work is in progress in explicating this notion more precisely (see Makkai, 1995; Tsementzis, 2017a).

To be clear, we don't mean to say that Mod(T) should be seen *merely* as a category. If we did that, then we would lose sight of some of the most interesting information about a theory. Consider, in particular, the following fact:

PROPOSITION 6.6.8 *If T is a propositional theory, then* Mod(T) *is a discrete category – i.e., the only arrows in* Mod(T) *are identity arrows.*

This result implies that for any two propositional theories T and T', if they have the same number of models, then they are categorically equivalent. But don't let this make you think that the space Mod(T) of models of a propositional theory T has no interesting structure. We saw in Chapter 3 that it has interesting topological structure, which represents a notion of "closeness" of models.

At the time of writing, there is no canonical account of the structure that is possessed by Mod(T) for a general first-order theory T. However, there has been much interesting mathematical research in this direction. The first main proposal, due to Makkai (1985), defines the "ultraproduct structure" on Mod(T) – i.e., which models are ultraproducts of which others. Interestingly, as we saw in the previous section, the ultraproduct construction looks like a topological limiting construction – and the coincidence is exact for the case of propositional theories. The second proposal for identifying the structure of Mod(T) is due originally to Butz and Moerdijk (1998), and has been recently developed by Awodey and Forssell (2013). According to this second proposal, Mod(T) is a topological groupoid, i.e., a groupoid in the category of topological spaces. Thus, according to both proposals, Mod(T) is like a category with a topology on it, where neither bit of structure – categorical or topological – is dispensable.

In the case of predicate logic theories, the categorical structure of Mod(T) does occasionally tell us something about T. We first show that the completeness or incompleteness of a theory can be detected by its category of models. Recall that a theory T in signature Σ is said to be **complete** just in case for each Σ-sentence ϕ, either $T \vdash \phi$ or $T \vdash \neg\phi$. Obviously every inconsistent theory is incomplete. So when we talk about a complete theory T, we usually mean a complete, consistent theory. In this case, the following conditions are equivalent:

1. T is complete.
2. $\mathrm{Cn}(T) = \mathrm{Th}(M)$ for some Σ-structure M.
3. T has a unique model, up to elementary equivalence – i.e., if M, N are models of T, then $M \equiv N$.
4. $\mathrm{Mod}(T)$ is directed in the sense that for any two models M_1, M_2 of T, there is a model N of T and elementary embeddings $h_i : M_i \to N$.

EXERCISE 6.6.9 Prove that the four conditions are equivalent. Hint: use Prop. 6.5.2.

The last property is a categorical property: if **C** and **D** are categorically equivalent, then **C** is directed iff **D** is directed. Therefore, completeness of theories is an invariant of categorical equivalence.

Now, it is well known that complete theories can nonetheless have many non-isomorphic models. It has occasionally been thought that an ideal theory T would be **categorical** in the sense that every two models of T are isomorphic. (The word "categorical" here has nothing to do with category theory.) However, the Löwenheim–Skølem theorem destroys any hope of finding a nontrivial categorical theory: if T has an infinite model, then it has models of other infinite cardinalities, and these models cannot be isomorphic. For the purposes, then, of classifying more of less "nice" theories, logicians found it useful to weaken categoricity in the following way:

DEFINITION 6.6.10 Let κ be a cardinal number, and let T be a theory in signature Σ. We say that T is κ-**categorical** just in case any two models M and N of T, if $|M| = |N| = \kappa$, then there is an isomorphism $h : M \to N$.

Example 6.6.11 Let T be the empty theory in signature $\{=\}$. A model of T is simply a set, and two models of T are isomorphic if they have the same cardinality. Therefore, T is κ-categorical for each cardinal number κ. ⌟

Example 6.6.12 Let $\Sigma = \{<\}$, where $<$ is a binary relation symbol. Let T be the theory in Σ that says that $<$ is a discrete linear order without endpoints. Then T is not $\aleph_0$-categorical. For example, the set $\mathbb{N}$ of natural numbers (with its standard ordering) is a model of T, but so is the disjoint union $\mathbb{N} \amalg \mathbb{N}$, where every element of the second copy is greater than every element of the first. ⌟

Thus, if T is categorical for all cardinal numbers, then $\mathrm{Mod}(T)$ has a relatively simple structure as a category: it is like a tower, with a unique (up to isomorphism) model M_κ for each cardinal number κ. (A generalization of the Löwenheim–Skølem theorem shows that for each infinite model M of T, there is a model N of T of higher cardinality and an elementary embedding $h : M \to N$.) Nonetheless, it is well known that there are many inequivalent categorical theories, and these theories are differentiated by the topological groups of symmetries of their models.

We now set aside the discussion of equivalence to look at other types of relations between theories. Recall that a translation $F : T \to T'$ is said to be **essentially surjective** just in case for each Σ-sentence ψ there is a Σ-sentence ϕ such that $T' \vdash \psi \leftrightarrow F\phi$.

A paradigmatic case of an essentially surjective translation is the translation from a theory T to a theory T' with some new axioms in the same signature. Recall also that a functor $F^* : \mathbf{C} \to \mathbf{D}$ is said to be **full** just in case for any objects M, N of $\mathbf{C}$, and for any arrow $f : F^*M \to F^*N$, there is an arrow $g : M \to N$ such that $F^*g = f$. In the special case of groups (i.e., categories with only one object, and only isomorphisms), a functor is full iff it is a surjective homomorphism.

PROPOSITION 6.6.13 *If $F : T \to T'$ is essentially surjective then $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is full.*

Proof Let $h : F^*M \to F^*N$ be a Σ-elementary embedding. We need to show that $h = F^*j$ where $j : M \to N$ is a Σ'-elementary embedding. Finding the function j is easy, since h is already a function from the domain of M to the domain of N. Thus, we need only show that h is Σ'-elementary – i.e., that for any Σ'-formula ψ, the following is a pullback:

$$\begin{array}{ccc} M(\psi) & \longrightarrow & N(\psi) \\ \downarrow & & \downarrow \\ X^n & \xrightarrow{h^n} & Y^n \end{array}$$

Since F is eso, there is a Σ-formula ϕ such that $T' \vdash \psi \leftrightarrow F\phi$. Since M and N are models of T', $M(\psi) = M(F\phi) = F^*M(\phi)$ and $N(\psi) = N(F\phi) = (F^*N)(\phi)$. Since h is Σ-elementary, the diagram is a pullback. Therefore, $j : M \to N$ is Σ'-elementary, and F^* is full. □

The preceding result can be quite useful in showing that there is no essentially surjective translation from T to T'.

Example 6.6.14 Let T be the theory in signature $\{=\}$ that says there are exactly two things. Let T' be the theory in signature $\{=, c\}$ that says there are exactly two things. These two theories consist of exactly the same sentences; and yet, we will now see that they are not intertranslatable.

The theory T is categorical: i.e., it has a unique model $M = \{*, \star\}$ up to isomorphism, and $\mathrm{Aut}(M) = \mathbb{Z}_2$ is the permutation group on two elements. Thus, $\mathrm{Mod}(T)$ is equivalent to the group $\mathbb{Z}_2$. The theory T' is also categorical; however, its models are rigid, i.e., have no nontrivial automorphisms. Hence, $\mathrm{Mod}(T')$ is equivalent to the group (e). Clearly there is no full functor $G : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$, and, therefore, Prop. 6.6.13 entails that there is no essentially surjective translation $F : T \to T'$.

It would hardly to make sense to think of either T or T' as an actual scientific theory. However, in the spirit of constructing toy models, we could raise a fanciful question: if Jack accepted T and Jill accepted T', then what would be the locus of their disagreement? They both assert precisely the same sentence: there are exactly two things. We cannot say that they disagree about whether there are constant symbols, because symbols aren't things "in the world," but are devices used to speak about things in the world. So perhaps Jack and Jill disagree about whether the two things in the world are interchangeable? ⌟

The next pair of results derive properties of F from properties of F^*. We first recall the syntactic notion of a conservative extension.

DEFINITION 6.6.15 A translation $F : T \to T'$ is said to be **conservative** just in case $T' \vdash F\phi$ only if $T \vdash \phi$, for each Σ-formula ϕ.

Thus, a conservative translation $F : T \to T'$ is one that does not create new consequences for T. Paradigm examples of this kind of translation can be generated by the inclusion $I : \Sigma \to \Sigma'$ where $\Sigma \subseteq \Sigma'$. Adding this new vocabulary to Σ does not generate new consequences for a theory T in Σ.

Now let's consider how the notion of a conservative extension might be formulated semantically. Recall that a functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is said to be **essentially surjective** just in case for each model M of T, there is a model N of T' and an isomorphism $h : M \to F^*N$. In the case of an inclusion $I : \Sigma \to \Sigma'$, the functor I^* is essentially surjective iff each model of T can be expanded to a model of T'.

It's fairly easy to see that if F^* is essentially surjective, then F is conservative. In fact, we can weaken the condition on F^* as follows.

DEFINITION 6.6.16 Let $F : T \to T'$ be a translation. We say that $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is **covering** just in case for each $M \in \mathrm{Mod}(T)$, there is an $N \in \mathrm{Mod}(T')$ and an elementary embedding $h : M \to F^*(N)$.

PROPOSITION 6.6.17 *Let $F : T \to T'$ be a translation. If F^* is covering then F is conservative.*

Proof Suppose that $T' \vdash F\phi$. Let M be an arbitrary model of T, and let $h : M \to F^*(N)$ be the promised elementary embedding. Since $N \models F\phi$, we have $F^*(N) \models \phi$, and since h is elementary, $M \models \phi$. Since M was an arbitrary model of T, it follows that $T \vdash \phi$. □

COROLLARY 6.6.18 *Let $F : T \to T'$ be a translation. If F^* is essentially surjective, then F is conservative.*

The following example shows that the condition of F^* being essentially surjective is strictly stronger than F being conservative. Thus, a translation $F : T \to T'$ may be conservative even though not every model of T can be expanded to a model of T'.

Example 6.6.19 Let $\Sigma = \{c_q \mid q \in \mathbb{Q}\}$, and let $\Sigma' = \{c_r \mid r \in \mathbb{R}\}$. Let T' be the theory with axioms $c_r \neq c_s$ when $r \neq s$, and let T be the restriction of T' to Σ. Obviously, for each model M of T, there is a model $M' = M \sqcup N$ of T' and an elementary embedding $h : M \to I^*(M')$. By Prop. 6.6.17, T' is a conservative extension of T. However, a countable model M of T cannot be isomorphic to $I^*(M')$, for any model M' of T'. Therefore, I^* is not essentially surjective. ⌟

DISCUSSION 6.6.20 We have given a relatively weak condition on $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$, which implies that $F : T \to T'$ is conservative. Unfortunately, we do not know if these conditions are equivalent. It seems, in fact, that F being conservative is

equivalent to a slightly weaker (and more complicated) condition on F^*, as described by Breiner (2014).

The dual functor $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ has many additional uses. For example, we can now complete the proof that two theories T_1 and T_2 have a common definitional extension iff they are intertranslatable (i.e., homotopy equivalent).

THEOREM 6.6.21 (Barrett) *Suppose that T_i is a theory in Σ_i, where Σ_1 and Σ_2 are disjoint signatures. If T_1 and T_2 are intertranslatable, then T_1 and T_2 have a common definitional extension.*

Proof Suppose that T_1 and T_2 are intertranslatable, with $F : T_1 \to T_2$ and $G : T_2 \to T_1$ the relevant translations. We begin by defining definitional extensions T_1^+ and T_2^+ of T_1 and T_2 to the signature $\Sigma_1 \cup \Sigma_2$.

We define $T_1^+ = T_1 \cup \{\delta_s : s \in \Sigma_2\}$, where for each symbol $s \in \Sigma_2$ the Σ_2-sentence δ_s is an explicit definition of s. If $q \in \Sigma_2$ is an n-ary predicate symbol, then we let the definition $\delta_q \equiv \forall\vec{x}(q \leftrightarrow Gq)$. If $g \in \Sigma_2$ is an n-ary function symbol, then we let the definition $\delta_g \equiv \forall\vec{x}\forall y(g(\vec{x}) = y \leftrightarrow Gg(\vec{x}, y))$. It is straightforward to verify that T_1 satisfies the admissibility condition for δ_g.

We define $T_2^+ = T_2 \cup \{\delta_t : t \in \Sigma_1\}$ in the same manner. If $p \in \Sigma_1$ is an n-ary predicate symbol, then we let $\delta_p \equiv \forall\vec{x}(p \leftrightarrow Fp)$. If $f \in \Sigma_1$ is an n-ary function symbol, then we let $\delta_f \equiv \forall\vec{x}\forall y(f(\vec{x}) = y \leftrightarrow Ff(\vec{x}, y))$. It is also straightforward to verify that T_2 satisfies the admissibility condition for δ_f.

We show now that T_1^+ and T_2^+ are logically equivalent. Without loss of generality, we show that every model of T_2^+ is a model of T_1^+. The converse follows via an analogous argument. Let M be a model of T_2^+. We show that M is a model of T_1^+. There are two cases that need checking.

First, we show that $M(\phi) = 1$ when $T_1 \vdash \phi$. Since F^*M is a model of T_1, we have $1 = (F^*M)(\phi) = M(F\phi)$. One can then verify by induction that for every Σ_1 formula ψ, and for every model M of T_2^+, $M(\psi) = M(F\psi)$. Therefore, $M(\phi) = 1$.

Second, we show that $M(\delta_s) = 1$ for every $s \in \Sigma_2$. Let $q \in \Sigma_2$ be an n-ary predicate symbol. Then

$$M(q(\vec{x})) \;=\; M(FGq(\vec{x})) \;=\; M(Gq(\vec{x})).$$

The first equality follows from the fact that F and G are quasi-inverse and the fact that M is a model of T_2^+. The second equivalence follows from the argument of the previous paragraph. Thus, $M(\delta_q) = 1$. In a similar manner one can verify that $M(\delta_g) = 1$ for every function symbol $g \in \Sigma_2$.

We have therefore shown that each model of T_1^+ is a model T_2^+. Thus, T_1^+ and T_2^+ are logically equivalent, and T_1 and T_2 are definitionally equivalent. □

Example 6.6.22 Let $\Sigma \equiv \{=\}$, let T_1 be the theory in Σ that says there is exactly one thing, and let T_2 be the theory in Σ that says there are exactly two things. In one important sense, T_1 and T_2 have the same number of models: one (up to isomorphism).

Since T_1 and T_2 should not be considered to be equivalent, having the same number of models is not an adequate criterion for equivalence.

Perhaps we can strengthen that criterion by saying that two theories are equivalent if the models of the one can be *constructed* from the models of the other? But that criterion seems also to say that T_1 and T_2 are equivalent. From each model $\{*\}$ of T_1, we can construct a corresponding model $\{*, \{*\}\}$ of T_2; and we can recover the original model $\{*\}$ from the model $\{*, \{*\}\}$.

This criterion is alluring, but it is still far too liberal. We will need to do something to capture its intuition, but without making the criterion of equivalence too liberal.

One natural suggestion here is to consider $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ as categories, and to consider functors between them. There are then two proposals to consider:

1. Each functor $F : \mathrm{Mod}(T_1) \to \mathrm{Mod}(T_2)$ represents a genuine theoretical relation between T_1 and T_2.
2. Every genuine theoretical relation between T_1 and T_2 is represented by a functor $F : \mathrm{Mod}(T_1) \to \mathrm{Mod}(T_2)$.

There is immediate reason to question the first proposal. For example, in the case of propositional theories T_1 and T_2, the categories $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are discrete. Hence, functors $F : \mathrm{Mod}(T_1) \to \mathrm{Mod}(T_2)$ correspond one-to-one with functions on objects (in this case, models). But we have seen cases where intuitively inequivalent propositional theories have categories with the same number of models. Thus, it seems that not every functor (or function) between $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ represents a legitimate relation between the theories.

There's another, more concrete, worry about the first proposal. Consider the case where T_1 and T_2 are fairly expressive theories in first-order logic. For example, T_1 might be Peano arithmetic, and T_2 might be ZF set theory. Setting aside worries about the size of sets, a function from $\mathrm{Mod}(T_1)$ to $\mathrm{Mod}(T_2)$ is simply a pairing $\langle M, N\rangle$ of models of T_2 with models of T_1. But there need not be any "internal" relation between M and N. This goes against an intuition that for theories T_1 and T_2 to be equivalent, there needs to be relations between their individual models, and not just their categories of models qua categories. In the case at hand, we want to say that for any model M of T_1, there is a model N of T_2, and some relation $\Phi(M, N)$ between M and N. But what relations Φ are permitted? And does the *same* relation Φ need to hold for every model M and the corresponding N, or can the relation itself depend on the input model M? ⌟

6.7 Beth's Theorem and Implicit Definition

> [T]here is an argument, based on an application of Beth's renowned definability theorem, which might appear to render simultaneous support for physicalism and anti-reductionism impossible. (Hellman and Thompson, 1975)

The logical positivists vacillated between being metaphysically neutral and being committed to metaphysical naturalism. One particular instance of the latter commitment was their view on the mind–body problem. With the new symbolic logic as their tool, they

had a clear story to tell about how the mental is related to the physical: it is **reducible** to it. For example, suppose that $r(x)$ denotes some kind of mental property, say the property of being in pain. In this case, the reductionist says that there is a predicate $\phi(x)$ in the language of basic physics such that $\forall x(r(x) \leftrightarrow \phi(x))$ – i.e., something is in pain iff it instantiates the physical property ϕ.

Of course, we should be clearer when we say that $\forall x(r(x) \leftrightarrow \phi(x))$, for even a Cartesian dualist might say that this sentence is contingently true. That is, a Cartesian dualist might say that there is a purely physical description $\phi(x)$ that happens, as a matter of contingent fact, to pick out exactly those things that are in pain. The reductionist, in contrast, wants to say more – that there is some sort of lawlike connection between being in pain and being in a certain physical state. At the very least, a reductionist would say that

$$T \vdash \forall x(r(x) \leftrightarrow \phi(x)),$$

where T is our best scientific theory (perhaps the ideal future scientific theory). That is, according to the best theory, to be in pain is nothing more or less than to instantiate the physical property ϕ.

By the third quarter of the twentieth century, this sort of hard-core reductionism had fallen out of fashion. In fact, some of the leading lights in analytic philosophy – such as Hilary Putnam – had devised master arguments which were taken to demonstrate the utter implausibility of the reductionist point of view. Nonetheless, what had not fallen out of favor among analytic philosophers was the naturalist stance that had found its precise explication in the reductionist thesis. Thus, analytic philosophers found themselves on the hunt for a new, more plausible way to express their naturalistic sentiments.

In the 1970s, philosophers with naturalistic sentiments often turned to the concept of "supervenience" in order to describe the relationship between the mental and the physical. Now, there has been much debate in the ensuing years about how to cash out the notion of supervenience, and we don't have anything to add to that debate. Instead, we'll opt for the most obvious explication of supervenience in the context of first-order logic, in which case supervenience amounts to the model theorist's notion of implicit definability:

> Given a fixed background theory T, a predicate r is implicitly definable in terms of others $p_1, \ldots, p_n$ just in case for any two models M, N of T, if M and N agree on the extensions of $p_1, \ldots, p_n$, then M and N agree on the extension of r.

Now, there is a relevant theorem from model theory, viz. **Beth's theorem**, which shows that if T implicitly defines r in terms of $p_1, \ldots, p_n$, then T explicitly defines r in terms of $p_1, \ldots, p_n$; that is

$$T \vdash \forall x(r(x) \leftrightarrow \phi(x)),$$

where ϕ is a formula built from the predicates $p_1, \ldots, p_n$. In other words, if r supervenes on $p_1, \ldots, p_n$, then r is reducible to $p_1, \ldots, p_n$. According to Hellman and Thompson, this result "might appear to render simultaneous support for physicalism and anti-reductionism impossible."

We begin the technical exposition with a description of the background assumptions of Beth's theorem. To be clear, philosophers can take exception with these background assumptions. They might say that we have stacked the deck against non-reductive physicalism by means of these assumptions, and that a different account of supervenience will permit it to be distinguished from reducibility. Although such a response is completely reasonable, it suggests that physicalism isn't a sharp hypothesis but a stance that can be held "come what may."

Fixed Assumptions of Svenonius' and Beth's Theorems

- T is a theory in signature Σ.
- $\Sigma^+ = \Sigma \cup \{r\}$, where r is an n-ary relation symbol.
- T^+ is a theory in Σ^+.
- T^+ is a conservative extension of T.

Svenonius' and Beth's theorems are closely related. Svenonius' theorem begins with an assumption about symmetry and invariance:

In each model M of T^+, the subset $M(r)$ is invariant under Σ-automorphisms.

It then shows that for each model M of T^+, there is a Σ-formula ϕ such that $M(r) = M(\phi)$. The formula ϕ may differ from model to model. Beth's theorem begins with the assumption that T^+ implicitly defines r in terms of Σ.

DEFINITION 6.7.1 We say that T^+ **implicitly defines** r in terms of Σ just in case for any two models M, N of T^+, if $M|_\Sigma = N|_\Sigma$, then $M = N$.

(Here $M|_\Sigma$ is the Σ-structure that results from "forgetting" what M assigns to the relation symbol $r \in \Sigma^+ \backslash \Sigma$.) Beth's theorem then shows that T^+ explicitly defines r in terms of Σ – i.e., there is a single Σ-formula ϕ such that $T^+ \vdash \forall \vec{x}(r(\vec{x}) \leftrightarrow \phi(\vec{x}))$, hence, in every model M of T^+, the relation r is coextensive with ϕ.

There are a variety of ways that one can prove the theorems of Beth and Svenonius. The reader may like, for example, to study the fairly straightforward proof of Beth's theorem in Boolos et al. (2002, chapter 20). However, our goal here is not merely to convince you that these theorems are true. We want to give you a feel for why they are true and to help you see that they are instances of certain general mathematical patterns. To achieve these ends, it can help to expand the mathematical context – even if that requires a bit more work. Accordingly, we will present a proof of Beth's theorem with a more topological slant.

We begin with the notion of a **type** in a model M of a theory T. (The terminology here is not particularly intuitive, but it has become standard. A better phrase might be been "ideal element.") As a quick overview, each element $a \in M$ corresponds to a family Γ of formulas $\phi(x)$, viz. those formulas that it satisfies. That is,

$$\Gamma = \{\phi(x) \mid a \in M(\phi(x))\}.$$

In fact, this set Γ is a filter relative to implication in M. That is, if $\phi(x) \in \Gamma$ and $M \models \forall x(\phi(x) \rightarrow \psi(x))$, then $\psi(x) \in \Gamma$. Similarly, if $\phi(x), \psi(x) \in \Gamma$, then $\phi(x) \wedge \psi(x) \in \Gamma$. Finally, for any $\phi(x)$, either $\phi(x) \in \Gamma$ or $\neg\phi(x) \in \Gamma$.

However, it's also possible to have an ultrafilter Γ of formulas for which there is no corresponding element $a \in M$. The obvious cases here are where the formulas "run off to infinity." For example, consider the family of formulas

$$\Gamma = \{r < x \mid r \in \mathbb{R}\},$$

with $\mathbb{R}$ the real numbers. Then Γ is a filter, but no real number a satisfies all formulas in Γ. Intuitively speaking, the filter Γ is satisfied by an ideal point at infinity that is greater than any real number. (While Γ is a filter, it is not an ultrafilter. It is contained in infinitely many distinct ultrafilters, each of which corresponds to a point at infinity.)

It's also possible for a model M to have ideal points "in the interstices" between the real points. For example, in the case of the real numbers $\mathbb{R}$, let's say that a filter Γ of formulas is *centered on* 0 just in case Γ contains each formula $-\delta < x < \delta$. Then a simple counting argument (with infinite cardinalities) shows that there are infinitely many incompatible filters, all of which are centered on 0. Each such filter corresponds to an ideal element that is smaller than any finite real number.

DEFINITION 6.7.2 Let M be a Σ-structure, and let p be a set of Σ-formulas in context $\vec{x} = x_1, \ldots, x_n$. We call p an ***n*-type** if $p \cup \mathrm{Th}(M)$ is satisfiable. We say that p is a **complete *n*-type** if $\phi \in p$ or $\neg\phi \in p$ for all Σ-formulas ϕ in context $\vec{x}$. We let S_n^M be the set of all complete n-types.

Each element a in a model gives rise to a complete 1-type:

$$\mathrm{tp}^M(a) = \{\phi(x) \mid a \in M(\phi(x))\}.$$

Similarly, each n-tuple $\vec{a} = a_1, \ldots, a_n$ gives rise to a complete n-type $\mathrm{tp}^M(\vec{a}) \in S_n^M$. We say that $\vec{a}$ **realizes the type** $p \in S_1^M$ when $p = \mathrm{tp}^M(\vec{a})$.

DEFINITION 6.7.3 We now equip the set S_n^M of complete n-types with a topology, and we show that this topology makes S_n^M a Stone space. For each Σ-formula ϕ in context $\vec{x}$, let

$$E_\phi = \{p \in S_n^M \mid \phi \in p\}.$$

The definition here is similar to that which we used in defining the Stone space of a propositional theory. In that case, E_ϕ was the set of models of the sentence ϕ. In the present case, S_n^M are not quite models of a theory. But if M is a model of a theory T, then the types S_n^M are essentially all elements of M^n along with ideal elements.

In order to show that S_n^M is a compact topological space, we will need to adduce a central theorem of model theory – the so-called realizing types theorem. We cite the result without proof, referring the interested reader to Marker (2006, chapter 4).

THEOREM 6.7.4 (Realizing types) *Suppose that F is a finite subset of S_n^M. Then there is an elementary extension N of M such that each $p \in F$ is realized in N.*

PROPOSITION 6.7.5 *S_n^M is a compact topological space.*

Proof Recall that a topological space is compact just in case any family of closed sets with the finite intersection property (fip) has nonempty intersection. Suppose then that $\mathcal{F}$ is a collection of closed subsets of S_n^M that has the fip. It will suffice to consider the case where the elements of $\mathcal{F}$ are each of the form E_ϕ for some Σ-formula ϕ. Let $\mathcal{F}_0$ denote the corresponding family of formulas. Since $\mathcal{F}$ has the fip, for each $\phi_1, \ldots, \phi_n \in \mathcal{F}_0$, there is some $p \in S_n^M$ such that $\phi_1, \ldots, \phi_n \in p$, hence $\phi_1 \wedge \cdots \wedge \phi_n \in p$. By the realizing types theorem, there is an elementary extension N of M and $a \in N(\phi_1 \wedge \cdots \wedge \phi_n)$. Thus, $\mathrm{Th}(M) \cup \mathcal{F}_0$ is finitely satisfiable. By the compactness theorem, $\mathrm{Th}(M) \cup \mathcal{F}_0$ is satisfiable, and hence $\mathcal{F}_0$ is an n-type. Since each n-type is contained in a complete n-type, we are done. □

We now look at the relationship between types and symmetries of models.

DEFINITION 6.7.6 Let M be a Σ-structure, and let $a, b \in M$. We say that a and b are **indiscernible** in M just in case $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$. In other words, for every Σ-formula ϕ, $a \in M(\phi)$ iff $b \in M(\phi)$.

DEFINITION 6.7.7 Let $a, b \in M$. We say that a and b are **co-orbital** just in case there is an automorphism $h : M \to M$ such that $h(a) = b$.

Since automorphisms are invertible and closed under composition, being co-orbital is an equivalence relation on M, and it partitions M into a family of equivalence classes. We call these equivalence classes the **orbits** under the symmetry group $\mathrm{Aut}(M)$.

PROPOSITION 6.7.8 *Let $h : M \to N$ be an elementary embedding. Then $\mathrm{tp}^M(a) = \mathrm{tp}^N(h(a))$.*

Proof Since h is an elementary embedding $a \in M(\phi)$ iff $h(a) \in N(\phi)$, for all Σ-formulas ϕ. □

The preceeding result leads immediately to the following.

PROPOSITION 6.7.9 *If two elements a, b are co-orbital, then they are indistinguishable. That is, if there is an automorphism $h : M \to M$ such that $h(a) = b$, then $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$.*

Example 6.7.10 We now show that the converse to the previous proposition is not generally true – i.e., indistinguishable elements are not necessarily co-orbital. Let $\Sigma = \{<, c_1, c_2, \ldots, d_1, d_2, \ldots\}$, where $<$ is a binary relation, and the c_i and d_j are constant symbols. Define a Σ-structure M as follows: the domain of M is the rational numbers $\mathbb{Q}$; $<$ is given its standard interpretation on $\mathbb{Q}$; $M(c_i) = -\frac{1}{i}$ and $M(d_i) = 1 + \frac{1}{i}$ for $i = 1, 2, \ldots$

- We claim first that $[0,1]$ is invariant under all automorphisms of M. Indeed, for each $i \in \mathbb{N}$, let $(c_i, d_i) = M(c_i < x < d_i)$. Then

$$[0,1] = \bigcap_{i=1}^{\infty} (c_i, d_i).$$

  If $h : M \to M$ is an automorphism, then (c_i, d_i) is invariant under h, hence $[0,1]$ is invariant under h.
- We claim that there is no Σ-formula ϕ such that $[0,1] = M(\phi)$. Indeed, it's easy to see that for any formula ϕ, if $1 \in M(\phi)$, then there is a $\delta > 0$ such that $1 + \delta \in M(\phi)$.
- We claim that $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$ for all $a, b \in [0,1]$. For this, we can argue in two steps. First, for any $a, b \in (0,1)$, there is an automorphism $h : M \to M$ such that $h(a) = b$. Second, choose $a \in (0,1)$, and show that $\mathrm{tp}^M(a) = \mathrm{tp}^M(1)$. Let $\phi \in \mathrm{tp}^M(1)$. By an argument similar to the preceding one, there is some $\delta > 0$ and some $c \in (1-\delta, 1)$ such that $\phi \in \mathrm{tp}^M(c)$. Since there is an automorphism h such that $h(a) = c$, it follows that $\phi \in \mathrm{tp}^M(a)$. Therefore, $\mathrm{tp}^M(1) \subseteq \mathrm{tp}^M(a)$.
- We claim that there is no automorphism $h : M \to M$ such that $h(0) = 1$. Suppose, to the contrary, that h is such an automorphism, and let $a \in (0,1)$. Since $0 < a$ and h is order preserving, $1 = h(0) < h(a)$. Thus, there is an $i \in \mathbb{N}$ such that $h(a) \in (d_i, \infty)$. But $\mathrm{tp}^M(a) = \mathrm{tp}^M(h(a))$, and, therefore, $a \in (d_i, \infty)$ – a contradiction.
- Notice, finally, that the element $1 \in M$ has the following feature: for every formula ϕ, if $M \vDash \phi(1)$, then $M \vdash \phi(a)$ for some $a > 1$.

⌟

The previous considerations show that M has a partition $\mathbb{O}$ into orbits and a partition $\mathbb{I}$ into indiscernables, and that $\mathbb{I} \subseteq \mathbb{O}$.

We will also need the following result, which shows that indiscernibles in M always lie on the same orbit in some elementary extension N of M. We again refer the reader to Marker (2006, chapter 4) for a proof.

PROPOSITION 6.7.11 *Let M be a Σ-structure, and suppose that $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$. Then there is a Σ-structure N, an elementary embedding $h : M \to N$, and an automorphism $s : N \to N$ such that $s(h(a)) = h(b)$.*

In the case of finite structures, most of these subtle distinctions evaporate. For example, in finite structures, indistinguishable elements are automatically co-orbital.

PROPOSITION 6.7.12 *Let M be a finite Σ-structure, and suppose that $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$. Then there is an automorphism $k : M \to M$ such that $k(a) = b$.*

Proof Suppose that $\mathrm{tp}^M(a) = \mathrm{tp}^M(b)$. By Prop 6.7.11, there is an elementary embedding $h : M \to N$ and an automorphism $j : N \to N$ such that $j(h(a)) = h(b)$. Since M is finite, h is an isomorphism. Define $k = h^{-1} \circ j \circ h$. Then $k(a) = h^{-1}(j(h(a))) = h^{-1}(h(b)) = b$. □

Let's talk now about **invariant subsets** of a model M. A subset $A \subseteq M$ is said to be **invariant** just in case $h(A) = A$ for every automorphism $h : M \to M$. By definition, the automorphisms of a Σ-structure preserve the extensions of Σ formulas. That is, if ϕ is a Σ-formula (with a single free variable), then $M(\phi)$ is invariant under all automorphisms of M. The converse, however, is not true – i.e., not all invariant subsets are extensions of some formula. We already saw one example of this situation in 6.7.10. Other examples are easy to come by. Consider, for example, the natural numbers $\mathbb{N}$ as a model of Peano arithmetic. This model is **rigid** – i.e., there are no nontrivial automorphisms. Hence, every subset of $\mathbb{N}$ is invariant under automorphisms. Nonetheless, the language Σ of Peano arithmetic only has a countable number of formulas. Thus, there are many invariant subsets of $\mathbb{N}$ that are not of the form $\mathbb{N}(\phi)$ for some formula ϕ.

Once again, finite structures don't have as much subtlety. Indeed, in finite structures, all invariant subsets are definable.

THEOREM 6.7.13 (finite Svenonius) *If M is a finite Σ-structure, and A is an invariant subset of M, then there is a Σ-formula θ such that $A = M(\theta)$.*

Proof Let $\mathscr{B}$ be the Boolean algebra of representable subsets of M, i.e., sets of the form $M(\phi)$ for some formula ϕ. For each $a \in M$, the set

$$\{M(\phi) \mid \phi \in \mathrm{tp}^M(a)\} \;=\; \{M(\phi) \mid a \in M(\phi)\},$$

is an ultrafilter on $\mathscr{B}$. Thus, if X is the Stone space of $\mathscr{B}$, there is a map $\pi \equiv \mathrm{tp}^M : M \to X$ such that $\pi(a)[M(\phi)] = 1$ iff $a \in M(\phi)$. In this case, since $\mathscr{B}$ is finite, each ultrafilter is principal, i.e., is the up-set of some $M(\phi)$. Hence $\pi : M \twoheadrightarrow X$ is surjective.

Since A is invariant under $\mathrm{Aut}(M)$, Prop. 6.7.12 entails that $\mathrm{tp}^M(a) \neq \mathrm{tp}^M(b)$ whenever $a \in A$ and $b \notin A$. Thus, A descends along π, i.e., $\pi^{-1}[\pi(A)] = A$. Since X is finite, $\pi(A)$ is clopen – i.e., there is a formula θ such that $A = M(\theta)$. □

The following key result will lead very quickly to a proof of Svenonius' theorem.

PROPOSITION 6.7.14 *Let M be a Σ^+-structure, and suppose that for every elementary extension N of M, any automorphism of $N|_\Sigma$ preserves $N(r)$. Then there is a Σ-formula ϕ such that $M \vDash \forall x(r(x) \leftrightarrow \phi(x))$.*

Proof We first claim that in every elementary extension N of M, if $a, b \in N$ such that $\mathrm{tp}^N(a)|_\Sigma = \mathrm{tp}^N(b)|_\Sigma$, then $a \in N(r)$ iff $b \in N(r)$. Suppose not, i.e., that there is an elementary extension N of M with $a, b \in N$ satisfying the same Σ-formulas, but $a \in N(r)$ and $b \notin N(r)$. By using an argument similar to the realizing types theorem, we can show that there is an elementary extension $i : N \to N'$, and an automorphism s of $N'|_\Sigma$ such that $s(i(a)) = i(b)$. Thus, s does not leave $N'(r)$ invariant, contradicting the assumptions of the proposition.

Now since any finite subset of S_1^M is realized in some elementary extension of M (Prop 6.7.4), it follows that for all $p, q \in S_1^M$, if $p|_\Sigma = q|_\Sigma$, then $p \in E_r$ iff $q \in E_r$. Conversely, if $p \in E_r$ and $q \notin E_r$, then there is some Σ-formula ϕ such that $p \in E_\phi$ and $q \notin E_\phi$. From this, it follows that the intersection of all E_ϕ such that $p \in E_\phi$ lies

in E_ϕ. By the compactness of S_1^M, there are finitely many Σ-formulas $\phi_1, \ldots, \phi_n$ such that $p \in E_{\phi_i}$ and

$$E_{\phi_i} \cap \cdots \cap E_{\phi_n} \subseteq E_r.$$

If we let $\psi_p \equiv \phi_1 \wedge \cdots \wedge \phi_n$, then $p \in E_{\psi_p}$ and $E_{\psi_p} \subseteq E_r$. The family $\{E_{\psi_p} \mid p \in E_r\}$ covers E_r, hence by compactness again has a finite subcover. Taking the conjunction of the corresponding formulas gives an explicit definition of $r(x)$ in terms of Σ. □

THEOREM 6.7.15 (Svenonius) *Suppose that in every model M of T^+, the set $M(r)$ is invariant under all Σ-automorphisms. Then there are Σ-formulas $\phi_1, \ldots, \phi_n$ such that*

$$T \vdash \forall x(r(x) \leftrightarrow \phi_1(x)) \vee \cdots \vee \forall x(r(x) \leftrightarrow \phi_n(x)).$$

Proof By the previous proposition, for each model M of T, there is a Σ-formula ϕ_M such that $M \vDash \forall x(r(x) \leftrightarrow \phi_M(x))$. Let $\psi_M \equiv \forall x(r(x) \leftrightarrow \phi_M(x))$, and let Δ be the set of $\neg\psi_M$, where M runs over all models of T. (To deal with size issues, we could consider isomorphism classes of models bounded by a certain size, depending on the signature Σ.) Then $T \cup \Delta$ is inconsistent. By compactness, there is a finite subset $\neg\psi_1, \ldots, \neg\psi_n$ of Δ such that $T \vdash \psi_1 \vee \cdots \vee \psi_n$. □

If, in addition, the theory T is complete, then the assumptions of Svenonius' theorem entail that T explicitly defines r in terms of Σ. Beth's theorem derives the same conclusion, without the completeness assumption, but with a stronger notion of implicit definability. Consider the following explications of the notion of definability:

IE Invariance under elementary embeddings: For any models M and N of T, and for any Σ-elementary embedding $h : M \to N$, $h(M(r)) = N(r)$.

IA Invariance under automorphisms: For any model M of T, and for any Σ-automorphism $h : M \to M$, $h(M(r)) = M(r)$.

IS For any models M and N of T, if $M|_\Sigma = N|_\Sigma$ then $M = N$. (This version is very close to the metaphysician's notion of global supervenience.)

ID Let T' be the result of uniformly replacing r in T with r'. Then $T \cup T' \vdash \forall x(r(x) \leftrightarrow r'(x))$.

The implication IE $\Rightarrow$ IA is immediate. To see that IE $\Rightarrow$ IS, suppose that $M|_\Sigma = N|_\Sigma$, and let $h : M \to N$ be the identity function. Then h is a Σ-elementary embedding, hence IE implies that $h(M(r)) = N(r)$, that is, $M(r) = N(r)$. To see that IS $\Rightarrow$ ID, let M be a model of $T \cup T'$. Define a $\Sigma \cup \{r\}$ structure N to agree with M on Σ, and such that $N(r) = M(r')$. Because M is a model of T', it follows that N is a model of T. Hence $M(r) = N(r) = M(r')$, and $M \vDash \forall x(r(x) \leftrightarrow r'(x))$.

We now show that $\neg$IE $\Rightarrow$ $\neg$ID. If $\neg$IE, then there are models M and N of T, and an elementary embedding $h : M \to N$ such that $h(M(r)) \neq N(r)$. We use N to define a $\Sigma \cup \{r, r'\}$ structure N': let N' agree with N on $\Sigma \cup \{r\}$, and let $N'(r') = h(M(r))$. Obviously N' is a model of T. To see that N' is a model of T', first let M' be the $\Sigma \cup \{r'\}$ structure that looks just like M except that $M'(r') = M(r)$. Then $M' \vDash T'$, and $N'(r') = h(M(r)) = h(M'(r'))$. That is, N' is the push-forward of M', and hence

$N' \vDash T'$. Finally, $N'(r) \neq N'(r')$, and hence $T \cup T' \nvdash \forall x(r(x) \leftrightarrow r'(x))$. This completes the proof that $\neg\text{IE} \Rightarrow \neg\text{ID}$.

All told, we have the following chain of implications:

$$\begin{array}{ccccc} \text{IE} & \Longleftrightarrow & \text{ID} & \Longleftrightarrow & \text{IS} \\ & & \Downarrow & & \\ & & \text{IA} & & \end{array}$$

What's more, the implication ID $\Rightarrow$ IA cannot be reversed.

We now sketch the proof that the stronger notion of implicit definability (IE,ID,IS) implies explicit definability.

THEOREM 6.7.16 (Beth's theorem) *If T implicitly defines r in terms of Σ, then T explicitly defines r in terms of Σ.*

Proof We follow the outlines of the proof by Poizat (2012, 185). Assume that T implicitly defines r in terms of Σ. Since IS $\Rightarrow$ IA, Svenonius' theorem implies that there are Σ-formulas $\phi_1, \ldots, \phi_n$ such that

$$T \vdash \forall x(r(x) \leftrightarrow \phi_1(x)) \vee \cdots \vee \forall x(r(x) \leftrightarrow \phi_n(x)).$$

If T were inconsistent, or consistent with only a single one of these disjuncts, then T would explicitly define r in terms of Σ. So suppose that $n > 1$, and T is consistent with all n disjuncts. For each disjunct $\forall x(r(x) \leftrightarrow \phi_i(x))$, let T_i be the theory that results from replacing r in T with ϕ_i. Implicit definability then yields $T_i \cup T_j \vdash \forall x(\phi_i(x) \leftrightarrow \phi_j(x))$. Notice that r does not occur in $T_i \cup T_j$. Using the compactness theorem, we can then use Σ-formulas $\theta_1, \ldots, \theta_m$ to divide the space of models of T into cells with the feature that for each k, we have $T, \theta_k \vdash \forall x(r(x) \leftrightarrow \phi_{i(k)})$ for some $i(k)$. One can then use the formulas $\theta_1, \ldots, \theta_m$ to construct an explicit definition of r in terms of Σ. □

Example 6.7.17 Petrie (1987) argues that global supervenience does not entail reducibility. We first state his definition verbatum:

> Let $\mathscr{A}$ and $\mathscr{B}$ be sets of properties. We say that $\mathscr{A}$ **globally supervenes** on $\mathscr{B}$ just in case worlds which are indiscernible wih regard to $\mathscr{B}$ are also indiscernible with regard to $\mathscr{A}$.

Switching to the formal mode – i.e., speaking of predicates rather than properties – and restricting to the context of first-order logic, it appears that global supervenience is just another name for implicit definability. We will use $\Sigma = \{p\}$ for the subvenient predicate symbol, and we let $\Sigma^+ = \Sigma \cup \{r\}$. Petrie describes his example as follows (with notation adapted):

> There are two structures M and N, both of which have domain $\{a, b\}$. In M, the extension of p is $\{a, b\}$ and the extension of r is $\{a\}$. In N, the extension of p is $\{a\}$ and the extension of r is empty.

Petrie points out that this example does not satisfy strong supervenience. However, since $M|_\Sigma \neq N|_\Sigma$, it trivially satisfies global supervenience – i.e., r is implicitly defined in terms of p.

Here we need to slow down: implicit definability is defined in terms of some background theory T. In this case, however, there can be no theory T such that M and N are models of T, and such that T implicitly defines r in terms of p. Indeed, for any theory T in Σ^+, if M is a model of T, then so is M' where $M'(p) = \{a, b\}$ and $M'(r) = \{b\}$. But then $M|_\Sigma = M'|_\Sigma$, whereas $M(r) \neq M'(r)$. Therefore, T does not implicitly define r in terms of Σ.

Thus, Petrie's space of possible worlds is not of the form $\mathrm{Mod}(T)$ for any theory T. One of the key assumptions of formal logic is that possibilities are specified only up to isomorphism – i.e., if M is possible, and M' results from permuting some of the (featureless) elements of M, then M' is also possible. (Why would it be possible that a is an r, but not possible that b is an r?) Thus, for one to grant the force of Petrie's counterexample, one has to abandon a key assumption of formal logic. Is it worth sacrificing formal logic in order to defend non-reductive physicalism? ⌟

6.8 Notes

- Within mathematics, the study of logical semantics is called **model theory**, and there are several excellent textbooks. Some of our favorites: Hodges (1993); Marker (2006); Poizat (2012).
- The classic sources on the semantic view of theories are Suppe (1974, 1989).
- The completeness of the predicate calculus was first proven by Kurt Gödel in his PhD thesis (Gödel, 1929).

  The typical textbook proof of the theorem proceeds as follows: supposing that Γ is proof-theoretically consistent, show that Γ can be expanded to a maximally consistent set Γ^*. This expansion invokes Zorn's lemma, which is a variant of the axiom of choice. The resulting set Γ^* is then used to construct a model of Γ.

  The topological proof in this chapter has several advantages over the typical textbook proof. For example, the topological theorem makes it clear that completeness doesn't require the full axiom of choice. It is known that the Baire category theorem is strictly weaker than AC (see Herrlich and Keremedis, 2000; Herrlich, 2006). The topological completeness proof was first given by Rasiowa and Sikorski (1950). See also (Rasiowa and Sikorski, 1963).

  An even more elegant proof of completeness is provided by Deligne's embedding theorem for coherent categories (see Makkai and Reyes, 1977). If $T \nvdash \bot$, then T corresponds to a (Boolean) coherent category C_T. By Deligne's theorem, there is an embedding $F : C_T \to \mathbf{Sets}$, which yields a model of T.
- The category **Sets** has tons of structure (limits, colimits, exponentials, etc.), and so is adequate to represent all syntactic structures of a first-order theory. If we're only interested in a fragment of first-order logic, it can also be interesting to look at representations in less structured categories. For example, Cartesian categories have enough structure to represent algebraic theories (such as the theory of groups). For more details, see Johnstone (2003).

- In Section 6.5, we redescribed topological structure on X as a family of operations $X^\infty \to X$, i.e., functions from infinite sequences to points of X. This description is not merely heuristic: the category **CHaus** of compact Hausdorff spaces is equivalent to the category of algebras for the ultrafilter monad on **Sets**. Thus, **CHaus** is the category of models of an (infinitary) algebraic theory. For more details, see (Mac Lane, 1971, VI.9) and (Manes, 1976, 1.5.24).
- For an interesting analysis of supervenience and reduction in terms of ultraproducts, see Dewar (2018b).
- Beth's theorem first appeared in (1956), and Svenonius' in (1959). In recent work, Makkai (1991); Zawadowski (1995); Moerdijk and Vermeulen (1999) show that Beth's theorem is equivalent to a result about *effective descent morphisms*, a notion of central importance in mainstream mathematics. This kind of unifying result shows that there is no clear boundary between mathematics and metamathematics.
- Our discussion of Beth's theorem draws from Barrett (2018b). The relevance of Beth's theorem to the prospects of non-reductive physicalism was first pointed out by Hellman and Thompson (1975), and has been subsequently discussed by Bealer (1978); Hellman (1985); Tennant (1985, 2015). According to Hellman (personal communication), the issue was first brought up by Hilary Putnam in a graduate seminar at Harvard. For more on supervenience and its history in analytic philosophy, see McLaughlin and Bennett (2018).