Structure and Equivalence

Neil Dewar

October 2, 2020

Die Mathematiker sind eine Art Franzosen: redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsobald ganz etwas anders.

Goethe

Contents

Introduction

This is a book about structure in the representations of physics, about equivalence between such representations, and about the relationship between these two concepts. In a slogan, that relationship is as follows: for two representations to be equivalent is for them to posit the same structure, and the structure of a representation is that which it has in common with equivalent representations. The question of which half of this slogan is primary—that is, whether we should take equivalence to be a derivative notion from structure, or the other way round—is the animating question behind much of what follows. That said, the main aim of this book is not to answer that question, but to introduce students to the tools and ideas that can, I think, be useful in seeking to answer it.

The book is split into four parts, with each part comprising three chapters (with each of the twelve chapters being roughly the same length). Part I looks at issues of structure and equivalence in the context of formal, logical languages. Chapters 1 and 2 introduce notions of definability and translation with regards to (respectively) models and theories of first-order languages, and how these notions can be used to make precise ideas about equivalence, while Chapter 3 looks at whether Ramsey sentences provide a plausible way of explicating the structure of such a theory. The goal is to lay some ideas on the table, in the (admittedly artificial) context of formal languages, that we can use in studying the more physics-oriented structures introduced subsequently. This part of the book presumes familiarity with standard predicate logic.

Parts II and III engage most directly with questions of structure and equivalence as they arise within physics, especially as concerns symmetries in physics. My original intention was to write something much more general about symmetry and equivalence in physics, with appositely chosen case studies to illustrate those general lessons. However, this ambition foundered on three problems: a strict word-count, the desire to make this book even remotely pedagogically accessible, and the inverse relationship between tractability and generality. So instead, I offer two case studies, and leave it to the reader to consider how the lessons drawn from them might (or might not) generalise to other theories.

Thus, Part II is about N-particle Newtonian mechanics, and how its spacetime symmetries can be used as a way of recognising the presence of 'surplus structure' in representations of this theory: Chapter 4 introduces the theory and its symmetries, Chapter 5 discusses the reasons for thinking symmetry-variant structure is surplus structure, and Chapter 6 outlines how to formulate the theory without using such structure. Part III takes up these questions about symmetry and surplus structure, but applied to the symmetries of electromagnetism. These symmetries include both its spacetime symmetries (in Chapter 7) and its internal gauge symmetries (in Chapters 8 and 9). These parts should be mostly accessible to readers provided they have some undergraduate-level knowledge of physics, although some of the tools used are more abstract than one would typically find in an undergraduate physics course. The appendices—on vector and affine spaces, group theory, and differential forms—provide a guide to these tools, although one more suitable for reference or refreshment than introduction.

Finally, Part IV discusses the use of category-theoretic tools to study structure and equivalence: Chapters 10 and 11 introduce (respectively) categories and functors, and Chapter 12 describes how to apply them to categories formed from theories—including both the formal logical theories discussed in Part I, and the physical theories discussed in Parts II and III. So, this part of the book also seeks to bring together the ideas articulated in the earlier parts of the book. It is also perhaps the part of the book that will be most novel to students, at least to those coming from a philosophy or physics background; I hope, however, that it illustrates how the basic concepts, at least, are more readily understandable than one might have expected.

[TO ADD: ACKNOWLEDGMENTS AND THANKS]

Munich, October 2020

Part I.

Logic

1. Models

We begin our investigations by looking at the structure of, and equivalence relations between, models of first-order logic. This is a highly stylised context, especially if our ultimate goal is an inquiry into structure and equivalence in physics: it is a commonplace that for most theories in physics, a presentation of those theories in first-order logic is likely neither possible nor desirable. Nevertheless, as we will see, the first-order case will prove to have complexities enough for us to start with; and it will teach us some lessons that we can use when we turn to physics in the later chapters of this book.

1.1. Review of first-order semantics

We begin with a brief review of the terminology and notation of standard first-order model theory.[1] The foundational concept here is that of a *first-order language*, which consists of a logical vocabulary (common to all first-order languages) and a non-logical vocabulary or signature (different for each first-order language). The logical vocabulary comprises a set Var of *variables*, $x_1, x_2, \ldots, y_1, y_2, \ldots, z_1, z_2, \ldots$; the *equality* symbol $\ulcorner = \urcorner$; the *negation*, *conjunction*, *disjunction* and *implication* symbols $\ulcorner\neg\urcorner$, $\ulcorner\wedge\urcorner$, $\ulcorner\vee\urcorner$ and $\ulcorner\rightarrow\urcorner$; and the *universal and existential quantifiers* $\ulcorner\forall\urcorner$ and $\ulcorner\exists\urcorner$.

In general, a signature consists of both predicate-symbols and function-symbols. However, we will confine ourselves to signatures that only contain predicate-symbols: many of the notions in which we are interested (concerning definability and translatability) are much easier to handle when we exclude function-symbols, and those same notions demonstrate that any theory employing function-symbols is, in a certain sense, equivalent to a theory that uses only relation-symbols. In an ideal world we would have the space to investigate and discuss this notion of equivalence; but in this (as in so many respects), the world is far from ideal. Thus:

[1] Much of the notation and conventions follow Hodges (1997).

Definition 1. A *signature* consists of a set Σ of *predicate-symbols* (denoted by letters such as P, Q, R, etc.), each of which is associated with a natural number known as its *arity*. ♠

So unary predicate-symbols (those which take a single argument) have an arity of 1; binary predicate-symbols (those which take two arguments, also known as binary relation-symbols) have an arity of 2; and so on. Where appropriate, the arity of a symbol will be indicated by adding a parenthetical superscript: to introduce R as a binary predicate-symbol, for instance, we will write its first appearance as $R^{(2)}$.

Given a signature Σ, one can define the set $\mathrm{Form}(\Sigma)$ of well-formed Σ-*formulae*, using the standard compositional rules of predicate logic. A variable in a formula is *free* if it is not bound by any quantifier; we will use $\phi(\xi_1, \ldots, \xi_n)$ to denote a formula with the variables $\xi_1, \ldots, \xi_n$ free, and $\phi(\eta_1/\xi_1, \ldots, \eta_n/\xi_n)$ to denote the result of uniformly substituting η_i for ξ_i throughout such a formula. The set of Σ-*sentences* is the set of closed Σ-formulae (formulae with no free variables).

The semantics for first-order model theory is given by *Tarski-models*. A Tarski-model $\mathfrak{A}$ for a language with signature Σ will be referred to as a Σ-*model*:[2]

Definition 2. A Σ-*model* consists of a set $|\mathfrak{A}|$ (the *domain* of $\mathfrak{A}$), equipped with a subset $\Pi_{\mathfrak{A}} \subseteq |\mathfrak{A}|^n$ (the *extension* of Π in $\mathfrak{A}$) for every $\Pi \in \Sigma$. ♠

A Tarski-model $\mathfrak{A}$ determines truth-values for formulae, relative to an assignment of elements of $|\mathfrak{A}|$ to variables in Var, in the standard recursive fashion. If the formula ϕ has the variables $x_1, \ldots, x_n$ free, and if $\mathfrak{A}$ satisfies ϕ relative to the assignment of $a_i \in |\mathfrak{A}|$ to x_i, then we write $\mathfrak{A} \models \phi[a_1, \ldots, a_n]$. If ϕ is a sentence, then the variable-assignment no longer matters, and we write simply $\mathfrak{A} \models \phi$.

1.2. Relationships between models

Most of the above should be familiar if you have taken a standard logic course before. Here, however, we are interested in exploring the use of these ideas to make precise concepts of structure and equivalence. For these purposes, it is very useful to start thinking about the kinds of relationships that Tarski models can bear to one another. First, for a given signature Σ, a *homomorphism* from a Σ-model $\mathfrak{A}$ to another Σ-model $\mathfrak{B}$ is a map which, in a certain sense, maps the structure of $\mathfrak{A}$ onto that of $\mathfrak{B}$. Formally:

[2] In the definition below, $|\mathfrak{A}|^n$ is the n-fold Cartesian product of $|\mathfrak{A}|$ with itself: that is, the set of ordered n-tuples of elements of $|\mathfrak{A}|$.

Definition 3. Let $\mathfrak{A}$ and $\mathfrak{B}$ be two Σ-models. A *homomorphism* $h : \mathfrak{A} \to \mathfrak{B}$ is a function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\text{If } \langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}}, \text{ then } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.1}$$

♠

When two models are homomorphic to one another, there is a certain kind of 'structural resemblance' between them—but it's reasonably weak. However, we can strengthen it in successive degrees. First, the definition of homomorphism requires only a left-to-right implication. If we also require the right-to-left implication to hold, and require that it hold of all atomic formulae (including those using the equality-symbol), then we obtain the notion of *embedding*:[3]

Definition 4. An *embedding* $h : \mathfrak{A} \to \mathfrak{B}$ is an injective function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.2}$$

♠

Finally, recall that $f : X \to Y$ is *surjective* if for any $y \in Y$, there is some $x \in X$ such that $f(x) = y$; a function which is both injective and surjective is *bijective*. This enables us to state the strongest kind of relationship between models that we will be interested in:

Definition 5. An *isomorphism* $h : \mathfrak{A} \to \mathfrak{B}$ is a surjective embedding. That is, it is a bijective function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.3}$$

♠

An isomorphism between a Tarski-model and itself is known as an *automorphism*. For any Tarski-model, the identity map on its domain is an automorphism; but many Tarski-models also possess 'non-trivial' automorphisms, i.e., automorphisms which are not the identity map.

The notion of isomorphism seems to naturally capture a notion of 'structural identity', that is, of what it is for two Tarski-models to have 'the same structure'. (Note that

[3] Recall that a function $f : X \to Y$ is *injective* if for any $x_1, x_2 \in X$, if $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$.

I only say it captures *a* notion of structural identity; as we shall explore, there are other ways of capturing that idea that we should also investigate.) After all, at least intuitively: from the existence of a bijective function, we can infer that the domains of the two models have the same number of elements; and the condition (1.3) indicates that the extensions of the various symbols in Σ are 'distributed' over those elements in the same way.

As a result, if a given formula ϕ holds of a certain n-tuple in $\mathfrak{A}$, then that same formula ϕ will hold of that n-tuple's image under h in $\mathfrak{B}$. In other words, an isomorphism preserves the satisfaction of formulae. However, the converse to this is not true: it is possible for a homomorphism to preserve the satisfaction of formulae without being an isomorphism. Such a homomorphism is known as an *elementary embedding*; formally,

Definition 6. An *elementary embedding* $h : \mathfrak{A} \to \mathfrak{B}$ is a function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for any n-place Σ-formula ϕ and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\text{If } \mathfrak{A} \models \phi[a_1, \ldots, a_n], \text{ then } \mathfrak{B} \models \phi[h(a_1), \ldots, h(a_n)] \tag{1.4}$$

♠

The notion of elementary embedding is also of great model-theoretic significance: indeed, as we will discuss in Chapter 12, there are good reasons for thinking that elementary embeddings, rather than homomorphisms, should be thought of as the structure-preserving mappings in the context of model theory.

However, we can also think about relationships between models of different signatures; indeed, such relationships will be our main concern in the remainder of this chapter. For example, suppose that we have two signatures Σ and Σ^+, such that $\Sigma \subset \Sigma^+$. Then, given any Σ^+-model $\mathfrak{A}$, the *reduct* of $\mathfrak{A}$ to Σ is, intuitively, what we get by 'forgetting' the extensions of all those predicate-symbols that are in Σ^+ but not in Σ. More formally,

Definition 7. Let $\Sigma \subset \Sigma^+$, and let $\mathfrak{A}$ be a Σ^+-model. The *reduct* of $\mathfrak{A}$ to Σ is denoted by $\mathfrak{A}_\Sigma$, and is defined as follows: the domains are identical (i.e. $|\mathfrak{A}_\Sigma| = |\mathfrak{A}|$), and for any $\Pi \in \Sigma$,

$$\Pi_{\mathfrak{A}_\Sigma} = \Pi_{\mathfrak{A}} \tag{1.5}$$

♠

The converse notion to reduct is that of *expansion*.

Definition 8. Let $\Sigma \subset \Sigma^+$, let $\mathfrak{A}$ be a Σ^+-model, and let $\mathfrak{B}$ be a Σ-model. $\mathfrak{A}$ is an *expansion* of $\mathfrak{B}$ to Σ^+ if $\mathfrak{A}_\Sigma = \mathfrak{B}$. ♠

1.3. Definability

I remarked earlier that isomorphism provides a certain natural sense of structural identity. On this basis, we might argue that two mathematical representations should be thought of as possessing the same structure just in case they are isomorphic to one another. But taken literally, this criterion is far too restrictive. For, strictly speaking, no two models of different signatures can be isomorphic to one another. So, for example, a strict linear order with five elements that represents the order relation using the symbol $<$ is not isomorphic to a strict linear order with five elements that represents the order relation using the symbol $\prec$. But it would be very strange to think of these two models as having different structures—the difference between them is merely notational.

However, this might seem an unduly uncharitable construal of the proposal. When philosophers talk about isomorphism, they often seem to have a less literal understanding of the notion of isomorphism: one which requires only that the number and distributions of the extensions over the models are the same, independently of what those extensions are labelled. Following Lutz (2015), we will refer to this more liberal notion as *H-isomorphism*:[4]

Definition 9. Let $\mathfrak{A}$ be a Σ_1-model, and let $\mathfrak{B}$ be a Σ_2-model. An *H-isomorphism* consists of a bijection $h : \mathfrak{A} \to \mathfrak{B}$ and a bijection $k : \Sigma_1 \to \Sigma_2$ such that for any $\Pi^{(n)} \in \Sigma_1$ and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in k(\Pi)_{\mathfrak{B}} \tag{1.6}$$

♠

Nevertheless, a little reflection suggests that this is still too restrictive a notion of isomorphism. Consider first the standard model of the natural numbers, equipped with extensions for zero, successor, addition, and multiplication. Call this model $\mathfrak{M}$. Now consider the standard model of the natural numbers, equipped with extensions for zero, successor, addition, multiplication, and *evenness*. Call this model $\mathfrak{N}$. $\mathfrak{M}$ and $\mathfrak{N}$ are not H-isomorphic: there is no bijection between their signatures, since those contain four and five symbols respectively. And yet, it is natural to feel that $\mathfrak{M}$ and $\mathfrak{N}$ have the same structure. After all, it is not as though the notion of evenness is somehow 'missing' in $\mathfrak{M}$, just because $\mathfrak{M}$ does not come equipped with a special label for it. The only difference, we want to say, between $\mathfrak{M}$ and $\mathfrak{N}$ is that some piece of structure which is implicitly present in $\mathfrak{M}$ has been bestowed with a specific name in $\mathfrak{N}$.

[4] So-called since it plays a role in Halvorson (2012)'s argument against the semantic view of theories.

We can make this notion of 'implicit structure' precise through the concept of *definability*. Intuitively, a certain collection of elements, or of tuples, is definable if it consists of precisely those elements that match a certain description. More formally:

Definition 10. Let $\mathfrak{A}$ be a Σ-structure. A set $X \subseteq |\mathfrak{A}|^n$ is *definable in* $\mathfrak{A}$ if there is some Σ-formula $\phi(x_1, \ldots, x_n)$ such that

$$\langle a_1, \ldots, a_n \rangle \in X \text{ iff } \mathfrak{A} \models \phi[a_1, \ldots, a_n] \tag{1.7}$$

♠

For example, the set of even numbers is definable in $\mathfrak{N}$, being definable by the (one-place) formula

$$\exists y(y + y = x) \tag{1.8}$$

If we accept the idea that the definable sets should be considered an (implicit) part of a Tarski-model's 'structure', then this also suggests regarding two models as having the same structure when the extensions of one model are definable in the other. We make this precise via the notion of *codetermination* of models:[5]

Definition 11. Let $\mathfrak{A}$ be a Σ-model and let $\mathfrak{B}$ be a T-model. $\mathfrak{A}$ and $\mathfrak{B}$ are *codeterminate* if:

- $|\mathfrak{A}| = |\mathfrak{B}|$;
- for every $\Pi \in \Sigma$, $\Pi_{\mathfrak{A}}$ is definable in $\mathfrak{B}$; and
- for every $\Omega \in \mathrm{T}$, $\Omega_{\mathfrak{B}}$ is definable in $\mathfrak{A}$.

♠

We can shed further light on definability by reflecting on its relationship to *invariance*. A set of elements (or tuples) in a Tarski-model is invariant if it is 'fixed' by all automorphisms of that model, that is:

Definition 12. Let $\mathfrak{A}$ be a Σ-structure. A set $X \subseteq |\mathfrak{A}|^n$ is *invariant in* $\mathfrak{A}$ if, for any automorphism $h : \mathfrak{A} \to \mathfrak{A}$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in X \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in X \tag{1.9}$$

♠

[5] See Barrett (nd), Winnie (1986).

Thus, for example, the extension of any predicate in Σ is guaranteed to be invariant (by the definition of automorphism). More generally, it turns out that the extension of any formula whatsoever is invariant: that is,

Theorem 1. *Let $\mathfrak{A}$ be a Σ-structure. For any set $X \subseteq |\mathfrak{A}|^n$, if X is definable then X is invariant.*

Proof. Left as exercise. □

However, the converse is not true: not all invariant sets are definable. For example, consider again the natural numbers (whether in the form of the model $\mathfrak{M}$ or the model $\mathfrak{N}$). This model is *rigid*, in that it possesses no non-trivial automorphisms. As a result, every subset in the domain is invariant. Since there are $\aleph_0$-many natural numbers, there are $2^{\aleph_0}$-many such subsets. But the signature is finite, and so there can only be at most $\aleph_0$-many formulae (since each formula is itself a finite construction); and hence, only at most $\aleph_0$-many definable subsets. That said, although in general a model might contain sets that are both indefinable and invariant, there is a partial result:

Theorem 2. *Let $\mathfrak{A}$ be a* finite *Σ-structure. For any set $X \subseteq |\mathfrak{A}|^n$, if X is invariant then X is definable.*

Proof. Left as exercise. □

Note a corollary: if $\mathfrak{A}$ is finite and rigid, then every subset of $|\mathfrak{A}|$ (or of $|\mathfrak{A}|^n$, for any n) is definable.

2. Theories

In the previous section, we looked at structure and equivalence between Tarski-models. In this section, we turn our attention to these concepts within the realm of *theories*. 'Theory' here will be meant in the usual sense of model theory:

Definition 13 (First-order theory)**.** Let Σ be a signature. A Σ*-theory* T is a set of Σ-sentences. ♠

Definition 14 (Model of a theory)**.** Let T be a Σ-theory. A Σ-model $\mathfrak{A}$ is a *model of* T if, for every sentence $\phi \in T$, $\mathfrak{A} \models \phi$. ♠

Definition 15 (Consequence)**.** Let T be a Σ-theory. A Σ-sentence ϕ is a *consequence* of T if, for every model $\mathfrak{A}$ of T, $\mathfrak{A} \models \phi$. In such a case, we say that T *entails* ϕ, and write $T \vDash \phi$; if ψ is a consequence of $\{\phi\}$, then we write $\phi \vDash \psi$.[1] ♠

We will denote the class of models of T by $\mathrm{Mod}(T)$. Within philosophy of science, there has been a great deal of discussion of the respective merits of the *syntactic view of theories* (roughly, that theories are sets of sentences) and the *semantic view of theories* (roughly, that theories are classes of models). The definition of a theory as a set of sentences is not intended to take a side on this debate, which may not be trading on quite such a sharp distinction as its protagonists suppose.[2] Any set of sentences brings a class of models in its wake, so any theory on the syntactic view will correspond to some theory on the semantic view; and although not any class of Tarski-models will be the class of models of some theory, many of the most interesting such classes are.[3]

[1] Note that the symbols for satisfaction and consequence are unfortunately similar: the former is $\models$, whilst the latter is $\vDash$. The easiest way to distinguish them is to look at what is on the left-hand-side of the symbol: if it is a Σ-model, then the relation is satisfaction; if it is a theory (or sentence), the relation is consequence.

[2] See, for instance, Lutz (2015).

[3] More specifically, if a class K of Σ-models is closed under isomorphism (i.e. if a model $\mathfrak{A}$ is in K, then so is any model isomorphic to $\mathfrak{A}$) and also closed under both 'ultraproducts' and 'ultraroots' (whose definition is too complex for a footnote), then there is some Σ-theory T such that $K = \mathrm{Mod}(T)$. See (Hodges, 1993, §9.5).

2.1. Translations between theories

What is it for two theories to be equivalent, i.e., to posit the same structure? The strictest criterion of equivalence that one might consider is that of *identity*: two theories are equivalent if they consist of the same sentences. Although this is surely a sufficient condition for equivalence, it seems overly restrictive. For instance, this criterion would consider the theories $\{\exists x Px\}$ and $\{\exists x \neg\neg Px\}$ to be inequivalent.[4]

A more relaxed condition is that of *logical equivalence*. For our purposes, this means having the same models:

Definition 16 (Logical equivalence)**.** Let T_1 and T_2 be Σ-theories. T_1 and T_2 are *logically equivalent* if $\mathrm{Mod}(T_1) = \mathrm{Mod}(T_2)$: that is, if for every Σ-model $\mathfrak{A}$, $\mathfrak{A}$ is a model of T_1 iff $\mathfrak{A}$ is a model of T_2. ♠

There's a natural relationship between isomorphism (as a criterion of equivalence between models) and logical equivalence (as a criterion of equivalence between theories), expressed by the following proposition.

Proposition 1. Let T_1 and T_2 be Σ-theories. T_1 and T_2 are logically equivalent iff for every model $\mathfrak{A}_1$ of T_1, there is an isomorphic model $\mathfrak{A}_2$ of T_2.

Proof. Left as exercise. □

In the previous section, we discussed the weaker notion of H-isomorphism, as a less language-dependent version of isomorphism. The corresponding notion for theories would be two theories that are logically equivalent 'up to a choice of notation'; we shall say that two theories related in this fashion are *notational variants* of one another.

Definition 17. Let T_1 be a Σ_1-theory, and let T_2 be a Σ_2-theory. T_1 and T_2 are *notational variants* of one another if there is an arity-preserving bijection $k : \Sigma_1 \to \Sigma_2$ such that $k(T_1)$ is logically equivalent to T_2, where $k(T_1)$ is the result of replacing every occurrence of any $P \in \Sigma_1$ in T_1 with $k(P)$. ♠

However, as a criterion of equivalence, notational variance is still very strict. In the previous chapter, we considered a further weakening from H-isomorphism, namely codetermination. This motivates us to consider a weaker kind of relationship between theories: that we can offer a *translation* between them.[5]

[4] If we had required that a theory be a set of sentences *closed under entailment* (so that if $T \vDash \phi$ then $\phi \in T$), then the identity criterion would coincide with the criterion of logical equivalence.

[5] For more detail on translations between theories, see (Halvorson, 2019, chap. 4).

The basic idea of translating one theory into another is that we can systematically replace expressions of the first theory's language by expressions of the second theory's language, in such a way that all theorems of the first theory are converted into theorems of the second theory. More precisely,

Definition 18 (Translation between theories)**.** Let T_1 be a Σ_1-theory, and T_2 a Σ_2-theory. A *translation* from T_1 to T_2 is a map $\tau : \mathrm{Form}(\Sigma_1) \to \mathrm{Form}(\Sigma_2)$, which:

1. Preserves variables: if the Σ_1-formula ϕ has exactly the variables $\xi_1, \ldots, \xi_n$ free, then $\tau(\phi)$ has exactly $\xi_1, \ldots, \xi_n$ free.

2. Commutes with substitution: for any Σ_1-formula ϕ with the variables $\xi_1, \ldots, \xi_n$ free, and any variables $\eta_1, \ldots, \eta_n$,

$$\tau(\phi(\eta_1/\xi_1, \ldots, \eta_n/\xi_n)) = \tau(\phi)(\eta_1/\xi_1, \ldots, \eta_n/\xi_n) \tag{2.1}$$

3. Commutes with the logical connectives: for any Σ_1-formulae ϕ and ψ, and any variable ξ,

$$\tau(\neg\phi) = \neg\tau(\phi) \tag{2.2}$$
$$\tau(\phi \wedge \psi) = \tau(\phi) \wedge \tau(\psi) \tag{2.3}$$
$$\tau(\forall\xi\phi) = \forall\xi\tau(\phi) \tag{2.4}$$

   etc.

4. Preserves consequence: for any Σ_1-formula ϕ,

$$\text{If } T_1 \vDash \phi \text{ then } T_2 \vDash \tau(\phi) \tag{2.5}$$

♠

When τ is a translation from T_1 to T_2, we will write $\tau : T_1 \to T_2$. Since a translation is required to commute with substitution and the logical connectives, we can specify such a translation just by specifying, for every $\Pi^{(n)} \in \Sigma_1$, how to translate $\Pi x_1 \ldots x_n$. In what follows, this is how we will usually specify translations.

How does the existence of a translation from one theory to another relate to the structures posited by the two theories? We can get some insight here by reflecting on how it is reflected in the relationships between the theories' classes of models. The key observation here is that a translation τ from one theory to another induces a 'dual map' τ^*

from the models of the latter theory to those of the former (so the dual map goes 'in the other direction' from the translation).

Definition 19 (Dual map to a translation)**.** Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. Given any Σ_2-model $\mathfrak{A}$, we define the Σ_1-model $\tau^*(\mathfrak{A})$ as follows. First, the domain of $\tau^*(\mathfrak{A})$ is the same as that of $\mathfrak{A}$, that is, $|\tau(\mathfrak{A})| = |\mathfrak{A}|$. Second, for any $\Pi^{(n)} \in \Sigma_1$, we define the extension of Π in $\tau^*(\mathfrak{A})$ as follows: for any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\tau^*(\mathfrak{A})} \text{ iff } \mathfrak{A} \models \tau(\Pi)[a_1, \ldots, a_n] \tag{2.6}$$

For any model $\mathfrak{A}$ of T_2, $\tau^*(\mathfrak{A})$ is a model of T_1 (see Proposition 3 below). So τ^* is a function from $\mathrm{Mod}(T_1)$ to $\mathrm{Mod}(T_2)$, which we refer to as the *dual map* to the translation τ. ♠

To prove the claim used in this definition—that if $\mathfrak{A}$ is a model of T_2, then $\tau^*(\mathfrak{A})$ is a model of T_1—we need the following useful proposition.

Proposition 2. Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. For any Σ_2-model $\mathfrak{A}$, and any Σ_1-sentence ϕ,

$$\tau^*(\mathfrak{A}) \models \phi \text{ iff } \mathfrak{A} \models \tau(\phi) \tag{2.7}$$

Proof. By induction on the length of formulae; left as exercise. □

Given this proposition, the proof that the dual map to a translation preserves modelhood is straightforward.

Proposition 3. Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. If $\mathfrak{A}$ is a T_2-model, then $\tau^*(\mathfrak{A})$ is a T_1-model.

Proof. Suppose, for *reductio*, that $\tau^*(\mathfrak{A})$ is not a T_1-model. Then there must be some sentence $\phi \in T_1$ such that $\tau^*(\mathfrak{A}) \not\models \phi$. Then by Proposition 2, $\mathfrak{A} \not\models \tau(\phi)$. Since $\mathfrak{A}$ is a T_2-model, it follows that $T_2 \nvDash \tau(\phi)$. But by the definition of a translation, $T_2 \vDash \phi$; so by contradiction, $\tau^*(\mathfrak{A})$ must be a T_1-model. □

Now, let us consider the question of how the notion of translation could be used to articulate a criterion of equivalence. The mere existence of a translation (as defined here) would be a very weak condition, and would have some very counter-intuitive consequences: it would mean, for example, that any theory would be equivalent to any strictly stronger theory (since inclusions are always translations)—indeed, that any theory is equivalent to some inconsistent theory! A more plausible criterion is to require the existence of a *pair* of translations. This criterion is known as *mutual interpretability*.

Definition 20. Let T_1 and T_2 be theories of signatures Σ_1 and Σ_2 respectively. T_1 and T_2 are *mutually interpretable* if there exist translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$. ♠

However, mutual interpretability is still a relatively weak notion, as the following examples indicate.

Example 1. Let $\Sigma_1 = \{P^{(1)}\}$, and let $\Sigma_2 = \{Q^{(1)}, R^{(1)}\}$. Let

$$T_1 = \varnothing \tag{2.8}$$

$$T_2 = \{\forall x(Qx \to Rx)\} \tag{2.9}$$

One would expect that T_1 and T_2 should not be regarded as equivalent: intuitively, T_2 says something non-trivial, whereas T_1 does not. Yet T_1 and T_2 are mutually interpretable, since

$$\tau(Px) = Qx \tag{2.10}$$

is a translation from T_1 to T_2, and

$$\sigma(Qx) = Px \tag{2.11}$$

$$\sigma(Rx) = Px \tag{2.12}$$

is a translation from T_1 to T_2.

In light of this, we introduce a yet stronger condition: not just that there exist a pair of translations, but that those translations be, in a certain sense, inverse to one another. The intuition here is that if we take some expression of our first theory's language, translate it into the second language, and then translate it back into the first language, we should—if the pair of translations really express an equivalence between the theories—get an expression with the same meaning as the expression with which we began. Formally, we cash out this condition of 'having the same meaning' as 'equivalent modulo the ambient theory'; the resulting criterion is known as *intertranslatability*.[6]

Definition 21. Let T_1 and T_2 be theories of signatures Σ_1 and Σ_2 respectively. T_1 and T_2 are *intertranslatable* if there exist translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$, such that for

[6] Barrett and Halvorson (2016a)

any Σ_1-formula $\phi(x_1, \ldots, x_n)$ and any Σ_2-formula $\psi(y_1, \ldots, y_m)$,

$$T_1 \vDash \forall x_1 \ldots \forall x_n(\phi \leftrightarrow \sigma(\tau(\phi))) \tag{2.13}$$

$$T_2 \vDash \forall y_1 \ldots \forall y_m(\psi \leftrightarrow \psi(\sigma(\psi))) \tag{2.14}$$

In such a case, we will say that τ and σ are *inverse translations* to one another. ♠

Where we are dealing with a pair of inverse translations of this kind, we will often express them by writing

$$\Pi x_1 \ldots x_n \equiv \tau(\Pi x_1 \ldots x_n) \tag{2.15}$$

for every $\Pi \in \Sigma_1$, and

$$\Omega y_1 \ldots y_m \equiv \sigma(\Omega y_1 \ldots y_m) \tag{2.16}$$

for every $\Omega \in \Sigma_2$. Thus, the symbol $\equiv$ will typically have expressions from two different languages on either side of it.

In general, if we have a theory T_1, then its image $\tau[T_1]$ under a map $\tau : \text{Form}(\Sigma_1) \to \text{Form}(\Sigma_2)$ is not intertranslatable with T_1, even if we suppose that τ preserves free variables, and that it commutes with substitution and the logical connectives.[7] However, if τ is 'suitably invertible', then this does hold. More precisely:

Proposition 4. Suppose that the translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$ are inverse to one another. Then T_2 is logically equivalent to $\tau[T_1]$, and T_1 is logically equivalent to $\sigma[T_2]$.

Proof. Suppose that there were some model $\mathfrak{B}$ of T_2 which was was not a model of $\tau[T_1]$. Then for some $\phi \in T_1$, $fB \not\vDash \tau(\phi)$; but then it follows that $T_2 \not\vDash \tau(\phi)$, which contradicts the assumption that τ is a translation. The other case is proven similarly. □

In particular, take as given some Σ_1-theory T_1, and suppose that $\tau : \text{Form}(\Sigma_1) \to \text{Form}(\Sigma_2)$ and $\sigma : \text{Form}(\Sigma_2) \to \Sigma_1)$ preserve free variables, commute with substitution and the logical connectives. If $\sigma(\tau(\phi))$ is logically equivalent to ϕ and $\tau(\sigma(\psi))$ is logically equivalent to ψ (for every $\phi \in \text{Form}(\Sigma_1)$ and $\psi \in \text{Form}(\Sigma_2)$), then T_1 is intertranslatable with $\tau[T_1]$; this also holds if we have equivalence with respect to some background theory, rather than full logical equivalence. We will employ this observation in Part II.

Finally, we observe that intertranslatability is associated with codetermination between classes of models in a natural way.

[7] See Barrett and Halvorson (2016a).

Proposition 5. If $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$ are inverse translations, then:

- for any $\mathfrak{A} \in \mathrm{Mod}(T_1)$, $\mathfrak{A}$ is codeterminate with $\sigma^*(\mathfrak{A})$, and $\tau^*(\sigma^*(\mathfrak{A})) = \mathfrak{A}$; and
- for any $\mathfrak{B} \in \mathrm{Mod}(T_2)$, $\mathfrak{B}$ is codeterminate with $\tau^*(\mathfrak{B})$, and $\sigma^*(\tau^*(\mathfrak{B})) = \mathfrak{B}$.

Proof. Left as exercise. □

This suggests that intertranslatability is a fairly natural criterion for equivalence between theories. That said, we should bear in mind that all we have discussed here are criteria of *formal* equivalence: roughly, of two theories having the same form, independently of their content. However, merely being of the same form is manifestly insufficient for two theories to be equivalent in the full sense of 'saying the same thing'. For example, as Sklar (1982) famously observes, the two theories 'all lions have stripes' and 'all tigers have stripes' are intertranslatable but do not say the same thing. So we should bear in mind that formal criteria like those discussed in this chapter (and, to some extent, the whole of this book) can only be a partial guide to theoretical equivalence.

3. Ramsey sentences

We've now seen some of the ways in which we can use the resources of model theory to articulate different senses of equivalence between models and theories. In the course of doing this, I have made occasional remarks about how these different notions of equivalence might be thought to capture different notions of 'structure', in the sense that they point to different ways of understanding the claim that two models, or two theories, have the same structure. However, there is an alternative way of approaching the relationship between the notions of structure and equivalence: rather than using equivalences to reveal structure, one can seek to articulate a notion of 'structure' directly, and then use that to formulate a criterion of equivalence. In this section, we consider one well-known proposal for the 'structural content' of a theory: that this content can be identified with the theory's *Ramsey sentence*.

3.1. Second-order logic

As we shall see, the Ramsey sentence of a first-order theory is a second-order sentence; so we begin by reviewing the formalism of second-order logic. In second-order logic, we can—as people say—*quantify into predicate position*. Intuitively, this means that we can make quantified claims about properties (and relations): so rather than being limited to saying things like 'all whales are mammals', we can now say things like 'anything which is true of all mammals is true of all whales', or 'there are some properties which whales and dolphins both have'.

More formally, then, second-order logic is distinguished from first-order logic by having not only a stock Var of first-order variables $x, y, z, \dots$, but also a stock $\textsc{Var}$ of second-order variables $X, Y, Z, \dots$. Like predicates, every second-order variable has an associated arity $n \in \mathbb{N}$. And as with predicates, we will indicate the arity of a second-order variable (where helpful) by a parenthesised superscript, thus: $X^{(n)}$. The subset of $\textsc{Var}$ containing all the n-ary variables will be denoted $\textsc{Var}^n$.

Other than this, the symbolic vocabulary of second-order logic is the same as that of first-order logic: we have the equality-symbol, the logical connectives, the quantifiers, and a signature Σ consisting of predicates (of various arities). The rules for forming well-formed formulae are the same as for first-order logic, but with two additional clauses:

- If $X^{(n)} \in \textsc{Var}$, and if $x_1, \ldots, x_n \in \mathrm{Var}$, then $Xx_1 \ldots x_n$ is a formula
- If ψ is a formula, and $X \in \textsc{Var}$, then $\forall X\psi$ is a formula

Respectively, these clauses tell us that the new variables can go into predicate position, and that they can be quantified over. As with the first-order case, we will use $\ulcorner\vee\urcorner$, $\ulcorner\rightarrow\urcorner$ and $\ulcorner\exists\urcorner$ as abbreviations (so $\exists X\psi$ abbreviates $\neg\forall X\psi$).

We now turn to the standard semantics of second-order logic.[1] As with the first-order case, we use Tarski-models: for signature Σ, a Σ-model $\mathfrak{A}$ consists of a set $\mathfrak{A}$ equipped with extensions for all predicates in Σ. Given a Σ-model $\mathfrak{A}$, a *second-order variable-assignment* G for $\mathfrak{A}$ consists of a map $g : \mathrm{Var} \rightarrow |\mathfrak{A}|$, and for every $n \in \mathbb{N}$, a map $G^n : \textsc{Var}^n \rightarrow \mathcal{P}(|\mathfrak{A}|^n)$. Here, $\mathcal{P}(|\mathfrak{A}|^n)$ is the *power set* of $|\mathfrak{A}|^n$, i.e. the set containing all subsets of $|\mathfrak{A}|^n$; thus, for any $\Xi^{(n)} \in \textsc{Var}$, $G^n(\Xi)$ is some set of n-tuples from $|\mathfrak{A}|$.

Now let ϕ be some second-order Σ-formula, let $\mathfrak{A}$ be a Σ-model, and let G be a second-order variable-assignment. Truth is then defined in the same way as in the first-order case, but with two extra clauses (corresponding to the two new clauses for formulae):

- For any $\Xi^{(n)} \in \textsc{Var}$ and any $\xi_1, \ldots, \xi_n \in \mathrm{Var}$,

$$\mathfrak{A} \models_G X^{(n)}x_1 \ldots x_n \text{ iff } \langle g(x_1), \ldots, g(x_n)\rangle \in G^n(\Xi) \tag{3.1}$$

- $\mathfrak{A} \models_G \forall X^{(n)}\phi$ iff for all $A \subseteq |\mathfrak{A}|^n$, $\mathfrak{A} \models_{G^X_A} \phi$

where the variable-assignment G^X_A is defined by the condition that

$$G^X_A(Y) = \begin{cases} G(Y) & \text{if } Y \neq X \\ A & \text{if } Y = X \end{cases} \tag{3.2}$$

We then say that a sentence ϕ is true relative to a Tarski-model $\mathfrak{A}$ if, for every variable-assignment G over $\mathfrak{A}$, $\mathfrak{A} \models_G \phi$. In this case, we write $\mathfrak{A} \models \phi$. Consequence is defined and denoted as in the first-order case.

[1] See (Shapiro, 1991, §4.2) or Manzano (1996) for more detailed treatments.

3.2. Ramseyfication

We can now turn our attention to the Ramsey sentence itself. The intuitive idea is that the 'structural core' of a theory T will make the same structural claims about the world as T, but without committing itself to which properties or relations it is that instantiate that structure. Thus, if a theory says something like 'positively charged particles repel one another', the structural claim thereby expressed is merely 'there is a property, such that any two particles possessing that property will repel one another'. One might object that even this does not go far enough, since it still speaks of 'repulsion'; or, one might distinguish between charge and repulsion on the basis that the notion of repulsion, unlike that of positive charge, is associated with a direct empirical content. In the first instance we will take the latter attitude, since the former (more extreme) view can be recovered as a special case.

Thus, suppose that our non-logical vocabulary Σ is divided into two classes, Ω and Θ: intuitively speaking, we suppose that Ω is the collection of 'observational' predicates (like 'repulsion'), while Θ is the collection of 'theoretical' predicates (like 'positive charge'). Suppose further that the theory T we are interested in (which is formulated in Σ) consists only of finitely many sentences; without loss of generality, we can suppose that T consists of a single sentence.[2]

We first form a 'skeleton' theory T^*, by replacing all the theoretical predicates that occur in T by second-order variables (of the appropriate arity): that is, if only $R_1, \ldots, R_p \in \Theta$ occur in T, then

$$T^* = T[X_1/R_1, \ldots, X_p/R_p] \tag{3.3}$$

where for each i, X_i is of the same arity as R_i. We then form the Ramsey sentence of T by existentially quantifying over all of these predicates:

$$T^R = \exists X_1 \exists X_2 \ldots \exists X_p T^* \tag{3.4}$$

We take the signature of the Ramsey sentence to be Σ (even though the sentence itself only contains predicates from Ω).

We can now ask the question: how much of a theory's structure does the Ramsey sentence capture? To answer this, first define the *observational reduct* of any Σ-model $\mathfrak{A}$

[2] In principle, we could apply the Ramseyfication procedure to a theory which consisted of infinitely many sentences. However, we would need the second-order language of T^R to be an *infinitary* second-order language: if T contained κ-many sentences, and if λ-many predicates from Θ occur in T, then T^R must be in a language that permits κ-size conjunction, and which admits the introduction of λ-many second-order quantifiers. In order to not have to deal with these complications, we will suppose that the original theory is finite.

to be its reduct $\mathfrak{A}_\Omega$ to Ω. Second, let $\mathfrak{W}$ be a first-order model of signature Σ which is a faithful representation of the world: i.e., which has the observational and theoretical predicates distributed over its elements in just the way that the corresponding observational and theoretical properties are distributed over the objects of the world. If this formulation makes you uncomfortable (which it probably should), then just think of $\mathfrak{W}$ as a 'preferred model', without worrying about in virtue of what it is preferred. We'll say that a Σ-theory is *true* if $\mathfrak{W}$ is one of its models; that it is *observationally adequate* if it has some model whose observational reduct is identical to $\mathfrak{W}_\Omega$; and that it is *numerically adequate* if it has some model whose domain coincides with $|\mathfrak{W}|$. Intuitively: a theory which is true admits a model which matches the actual number of objects, and the actual distribution of observational and theoretical properties over those objects; a theory which is observationally adequate admits a model which matches the actual number of objects, and the actual distribution of observational properties over those objects; and a theory which is numerically adequate admits a model which matches the actual number of objects[3].

This enables us to now make the following observation: for any Σ-theory T, its Ramsey sentence T^R is true just in case T is observationally adequate.[4] More formally:

Proposition 6. Let T be a theory of signature Σ, and let T^R be the Ramsey sentence of T. Then $\mathfrak{W} \models T^R$ if and only if T is observationally adequate (i.e., there is some model $\mathfrak{A}$ of T such that $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$).

Proof. First, suppose that $\mathfrak{W} \models T^R$: that is, that

$$\mathfrak{W} \models \exists X_1 \ldots \exists X_p T^* \tag{3.5}$$

This is true just in case there is some second-order variable-assignment G for $\mathfrak{W}$ such that

$$\mathfrak{W} \models_G T^* \tag{3.6}$$

[3] Bear in mind here that $\mathfrak{W}$ is merely supposed to be a 'faithful representative' of the world, not 'the world itself' (whatever, exactly, these terms might mean). So it's harmless to define observational adequacy as the theory admitting a model identical to $\mathfrak{W}_\Omega$ (not just isomorphic to it), and to define numerical adequacy as the theory admitting a model with the same domain identical to $|\mathfrak{W}|$ (not just equinumerous with it)

[4] (Ketland, 2004, Theorem 2)

But now consider the model $\mathfrak{A}$, defined as follows:

$$\begin{aligned}|\mathfrak{A}| &= |\mathfrak{W}| \\ P^{\mathfrak{A}} &= P^{\mathfrak{W}}, \text{ for every } P \in \Omega \\ R_i^{\mathfrak{A}} &= G(X_i), \text{ for every } R_i \in \Theta\end{aligned}$$

A proof by induction shows that

$$\mathfrak{A} \models T \tag{3.7}$$

But by construction, $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$. So T is observationally adequate.

Second, suppose that T is observationally adequate: i.e., that there is some model $\mathfrak{A}$ of T such that $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$. Consider any second-order variable-assignment G for $\mathfrak{A}$ satisfying the condition that for every $R_i \in \Theta$,

$$G(X_i) = R_i^{\mathfrak{A}} \tag{3.8}$$

Since $|\mathfrak{A}| = |\mathfrak{W}|$, we can regard G as a variable-assignment for $\mathfrak{W}$. Then, again, a proof by induction shows that

$$\mathfrak{W} \models_G T^* \tag{3.9}$$

from which it follows immediately that $\mathfrak{W} \models T^R$. □

We also have the following corollary, which applies to the more radical view canvassed above (that the use of *all* predicates, not just the 'theoretical' ones, should be converted to existential quantifications).

Corollary 1. Suppose that $\Theta = \varnothing$ (equivalently, that $\Sigma = \Omega$); that is, consider the case where we Ramsefy away *all* the vocabulary. Then $\mathfrak{W} \models T^R$ if and only if T is numerically adequate (i.e., there is some model $\mathfrak{A}$ of T such that $|\mathfrak{A}| = |\mathfrak{W}|$).

Philosophically, this observation is usually taken as a problem for the proposal that a theory's structure is captured by its Ramsey sentence: simply put, the concern is that Proposition 6 shows that the Ramsey sentence fails to capture anything about a theory beyond its empirical or observational content.[5] So if we do indeed take the 'structure' of a theory to be that which is captured by its Ramsey sentence, then we appear to

[5] In the literature, this objection is referred to as 'Newman's objection', since a version of this objection was discussed in Newman (1928). (Note that this is *before* the introduction of the Ramsey sentence in Ramsey (1931)—even allowing for the fact that Ramsey's essay was written in 1929. The reason for this is that Newman's objection was, originally, offered as a criticism of Russell (1927), and only later applied to the Ramsey-sentence approach to theories.)

have the corollary that a theory simply has no non-observational structure. Moreover, there is something faintly paradoxical to this, insofar as the observational predicates were precisely the ones that we did not Ramseyfy. So the Ramsey-sentence approach to structure seems to hold that although the Ramsey sentence of a theory articulates that theory structure, the only structure a theory in fact possesses is expressed by that part of the theory's language which is not subject to Ramseyfication!

If the Ramsey sentence is taken as expressing a theory's structure, then it is natural to take two theories as equivalent if they have logically equivalent Ramsey sentences. A further way of thinking about Newman's objection is to observe that, if we use standard second-order semantics, then this criterion of equivalence degenerates into observational equivalence. More precisely, let us say that two first-order Σ-theories, T_1 and T_2, are *observationally equivalent* if for every model $\mathfrak{A}$ of T_1, there is some model $\mathfrak{B}$ of T_2 such that $\mathfrak{A}_\Omega$ is isomorphic to $\mathfrak{B}_\Omega$, and vice versa. Then:

Proposition 7. T_1 and T_2 are observationally equivalent if and only if, with respect to standard second-order semantics, their Ramsey sentences are logically equivalent.

Proof. Left as exercise. □

What can be said in response? One option is to bite the bullet, and argue that—in fact—it is *good* that a theory's structure should turn out to be exhausted by its observational structure. In other words, the Ramsey sentence can be regarded as a useful vehicle for expressing a (fairly strong) form of empiricism about scientific theories: it offers one way of making precise the idea that the real content of a scientific theory is its observational or empirical 'core'. This is, roughly speaking, the attitude that Carnap (1958) took in advocating the Ramsey sentence as expressing the 'synthetic part' of a theory, with the 'analytic part' expressed by the so-called *Carnap sentence*, $T^C = (T^R \to T)$.[6]

For non-empiricists, it is less clear what the best response is. One option is to argue that the way we have formalised the Ramsey sentence failed to capture the intuitive idea. In particular, note that our intuitive gloss above quantified over *properties*, whereas the standard semantics for second-order logic permits the second-order variables to range over *arbitrary subsets of the domain*. So one might argue that the Ramsey-sentence approach to structural content should use some other semantics for second-order logic, where the range of the second-order quantifiers is somehow restricted.

A natural way of doing this is to use so-called *Henkin semantics*. In a *Henkin model* $\mathfrak{H}$, for each $n \in \mathbb{N}$ a subset of $\mathcal{P}(|\mathfrak{H}|^n)$ is picked out as the permitted range for the

[6] For commentary and discussion, see Psillos (2000) or Andreas (2017).

second-order n-ary variables to range over.[7] This suffices to avoid the Newman objection; however, it turns out that this still captures a relatively weak notion of structural content. In particular, Dewar (2019a) shows that two theories T_1 and T_2 have logically equivalent Ramsey sentences under Henkin semantics if there exist translations from T_1 to T_2 and vice versa (without any requirement that these translations are inverse to one another); and as Example 1 showed, this is a fairly weak notion of equivalence.

In sum, then, we've seen that the notion of 'the structure of a theory' is slippier than one might expect, and admits of a variety of different formal explications. In particular, we now have a hierarchy of criteria of equivalence, in descending order of strictness:

- Logical equivalence
- Notational variance
- Intertranslatability
- Mutual translatability
- Logically equivalent Ramsey sentences (on Henkin semantics)
- Logically equivalent Ramsey sentences (on standard semantics)

We now move away from logic, and turn to theories of physics; we will bear in mind the lessons from these chapters, however, as guides to these more complex and interesting cases.

[7] Moreover, we require that these privileged subsets are, in an appropriate sense, closed under definability; see Manzano (1996) for details.

Part II.

Newtonian mechanics