The Logic in Philosophy of Science

HANS HALVORSON
Princeton University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107110991
DOI: 10.1017/9781316275603

First published 2019

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Halvorson, Hans, author.
Title: The logic in philosophy of science / Hans Halvorson (Princeton University, New Jersey).
Description: Cambridge ; New York, NY : Cambridge University Press, 2019. |
Includes bibliographical references and index.
Identifiers: LCCN 2018061724 | ISBN 9781107110991 (hardback : alk. paper) |
ISBN 9781107527744 (pbk. : alk. paper)
Subjects: LCSH: Science–Philosophy.
Classification: LCC Q175 .H2475 2019 | DDC 501–dc23
LC record available at https://lccn.loc.gov/2018061724

ISBN 978-1-107-11099-1 Hardback
ISBN 978-1-107-52774-4 Paperback

Contents

Preface

The twentieth century's most interesting philosophers were enthralled by the revolution in mathematical logic, and they accordingly clothed many of their arguments in a formal garb. For example, Hilary Putnam claimed that the Löwenheim-Skølem theorem reduces metaphysical realism to absurdity; Bas van Fraassen claimed that arguments against empiricism presuppose the syntactic view of theories; and W. v. O. Quine claimed that Carnap's notion of an "external question" falls apart because every many-sorted theory is equivalent to a single-sorted theory. These are only a few of the many arguments of twentieth-century philosophy that hinge upon some or other metalogical theorem.

Lack of understanding of the logical theorems can be a huge obstacle to assessing these philosophers' arguments, and this book is my attempt to help remove that obstacle. However, my ideal reader is not the casual tourist of twentieth-century philosophy who wants the minimal amount of logic needed to get the big picture. My ideal reader is the (aspiring) logician-philosopher who wants to strip these arguments down to their logical nuts and bolts.

Although my motivation for writing this book wasn't to get across some particular philosophical point, a few such points emerged along the way. First, the distinction between realism and antirealism really boils down to one's attitude toward theoretical equivalence. Realists are people with a conservative notion of equivalence, and antirealists are people with a liberal notion of equivalence. Second, and relatedly, to give a philosophical account of a relation between theories (e.g., equivalence and reducibility) is tantamount to recommending certain norms of inquiry. For example, if you say that two theories T and T' are equivalent, then you mean (among other things) that any reason for accepting T is also a reason for accepting T'. Hence, you won't bother trying to design an experiment that would test T against T'. Similarly, if you think that T and T' are equivalent, then you'll consider as confused anyone who argues about which of the two is better. In short, to adopt a view on relations between theories is to adopt certain rules about how to use those theories.

I should explain one glaring omission from this book: modal logic. I didn't leave out modal logic because I'm a Quinean extensionalist. To the contrary, I've come to think that the metatheory of first-order logic, and of scientific theories more generally, is chock full of intensional concepts. For example, the models of a scientific theory represent the nomologically possible worlds according to the theory. Furthermore, a scientific theory comes equipped with a notion of "natural property" (in the sense of David Lewis), and these natural properties determine a notion of similarity between possible worlds, which

in turn licenses certain counterfactual inferences. So, while my goal is to theorize about the extensional logic that forms the backbone of the sciences, I believe that doing so calls for the use of intensional concepts.

A final note on how to read this book: Chapters 1–3 are introductory but are not strictly prerequisite for the subsequent chapters. Chapters 1 and 3 treat the metatheory of propositional logic, teaching some Boolean algebra and topology along the way. In Chapter 3, we go through the proof of the Stone duality theorem, because it exemplifies the duality between syntax and semantics that informs the remaining chapters. Chapter 2 covers the basics of both category and set theory in one go, and it's the most technically demanding (and least philosophical) chapter of the book. You don't have to know categorical set theory in order to benefit from the other chapters, it would be enough to know some set theory (e.g., Halmos' *Naive Set Theory*) and to flip back occasionally to look up category-theoretic concepts.

Acknowledgements: Thanks to Bas van Fraassen for the inspiration to pursue philosophy of science both as a science and as an art.

The idea behind this book arose during a year I spent in Utrecht studying category theory. I thank the Mellon New Directions Fellowship for financing that year. Thanks to my Dutch hosts (Klaas Landsman, Ieke Moerdijk, and Jaap van Oosten) for their warm hospitality.

When I returned home, I rediscovered that it's difficult to do two (or fifty) things at once. The project might have foundered, had it not been for the theorem-proving wizardry of Thomas Barrett, Neil Dewar, Dimitris Tsementzis, and Evan Washington. I also found my philosophical views shaped and sharpened by conversations with several students and colleagues, especially John Burgess, Ellie Cohen, Robbie Hirsch, Laurenz Hudetz, Michaela McSweeney, Alex Meehan, Gideon Rosen, Elliot Salinger, David Schroeren, and Jim Weatherall. I probably left somebody out, and I'm sorry about that. For comments and corrections on earlier versions of the manuscript, I thank Thomas Barrett, Gordon Belot, Neil Dewar, Harvey Lederman, Dimitris Tsementzis, Jim Weatherall and Isaac Wilhelm.

Finally, thank you to Hilary and Sophie at CUP for their initial belief in the project and for persevering with me to the end.

Introduction

A New Kind of Philosophy

Some people think that philosophy never makes progress. In fact, professional philosophers might think that more frequently – and feel it more acutely – than anyone else. At the beginning of the twentieth century, some philosophers were so deeply troubled that they decided to cast all previous philosophy on the scrap heap and to rebuild from scratch. "Why shouldn't philosophy be like science?" they asked. "Why can't it also make genuine progress?"

Now, you might guess that these philosophers would have located philosophy's problems in its lack of empirical data and experiments. One advantage of the empirical sciences is that bad ideas (such as "leeches cure disease") can be falsified through experiments. However, this wasn't the diagnosis of the first philosophers of science; they didn't see empirical testability as the sine qua non of a progressive science. Their guiding light was not the empirical sciences, but mathematics, and mathematical physics.

The nineteenth century had been a time of enormous progress in mathematics, not only in answering old questions and extending applications, but but also in clarifying and strengthening the foundations of the discipline. For example, George Boole had clarified the structure of logical relations between propositions, and Georg Cantor had given a precise account of the concept of "infinity," thereby setting the stage for the development of the new mathematical theory of sets. The logician Gottlob Frege had proposed a new kind of symbolic logic that gave a precise account of all the valid argument forms in mathematics. And the great German mathematician David Hilbert, building on a rich tradition of analytic geometry, proposed an overarching axiomatic method in which all mathematical terminology is "de-interpreted" so that the correctness of proofs is judged on the basis of purely formal criteria.

For a younger generation of thinkers, there was a stark contrast between the ever more murky terminology of speculative philosophy and the rising standards of clarity and rigor in mathematics. "What is the magic that these mathematicians have found?" asked some philosophically inclined scientists at the beginning of the twentieth century. "How is it that mathematicians have a firm grip on concepts such as 'infinity' and 'continuous function,' while speculative philosophers continue talking in circles?" It was time, according to this new generation, to rethink the methods of philosophy as an academic discipline.

The first person to propose that philosophy be recreated in the image of nineteenth-century mathematics was Bertrand Russell. And Russell was not at all modest in what he thought this new philosophical method could accomplish. Indeed, Russell cast himself as a direct competitor with the great speculative philosophers, most notably with Hegel. That is, Russell thought that, with the aid of the new symbolic logic, he could describe the fundamental structure of reality more clearly and accurately than Hegel himself did. Indeed, Russell's "logical atomism" was intended as a replacement for Hegel's monistic idealism.

Russell's grand metaphysical ambitions were cast upon the rocks by his student Ludwig Wittgenstein. In essence, Wittgenstein's *Tractatus Logico Philosophicus* was intended to serve as a *reductio ad absurdum* of the idea that the language of mathematical logic is suited to mirror the structure of reality in itself. To the extent that Russell himself accepted Wittgenstein's rebuke, this first engagement of philosophy and mathematical logic came to an unhappy end. In order for philosophy to become wedded to mathematical logic, it took a distinct second movement, this time involving a renunciation of the ambitions of traditional speculative metaphysics. This second movement proposed not only a new method of philosophical inquiry but also a fundamental reconstrual of its aims.

As mentioned before, the nineteenth century was a golden age for mathematics in the broad sense, and that included mathematical physics. Throughout the century, Newtonian physics has been successfully extended to describe systems that had not originally been thought to lie within its scope. For example, prior to the late nineteenth century, changes in temperature had been described by the science of thermodynamics, which describes heat as a sort of continuous substance that flows from one body to another. But then it was shown that the predictions of thermodynamics could be reproduced by assuming that these bodies are made of numerous tiny particles obeying the laws of Newtonian mechanics. This reduction of thermodynamics to statistical mechanics led to much philosophical debate over the existence of unobservable entities, e.g., tiny particles (atoms) whose movement is supposed to explain macroscopic phenomena such as heat. Leading scientists such as Boltzmann, Mach, Planck, and Poincaré sometimes took opposing stances on these questions, and it led to more general reflection on the nature and scope of scientific knowledge.

These scientists couldn't have predicted what would happen to physics at the beginning of the twentieth century. The years 1905–1915 saw no fewer than three major upheavals in physics. These upheavals began with Einstein's publication of his special theory of relativity, and continued with Bohr's quantum model of the hydrogen atom, and then Einstein's general theory of relativity. If anything became obvious through these revolutions, it was that we didn't understand the nature of science as well as we thought we did. We had believed we understood how science worked, but people like Einstein and Bohr were changing the rules of the game. It was high time to reflect on the nature of the scientific enterprise as a whole.

The new theories in physics also raised further questions, specifically about the role of mathematics in physical science. All three of the new theories – special and general relativity, along with quantum theory – used highly abstract mathematical notions, the likes

of which physicists had not used before. Even special relativity, the most intuitive of the three theories, uses four-dimensional geometry and a notion of "distance" that takes both positive and negative values. Things only got worse when, in the 1920s, Heisenberg proposed that the new quantum theory make use of non-commutative algebras that had no intuitive connection whatsoever to things happening in the physical world.

The scientists of the early twentieth century were decidedly philosophical in outlook. Indeed, reading the reflections of the young Einstein or Bohr, one realizes that the distinction between "scientist" and "philosopher" had not yet been drawn as sharply as it is today. Nonetheless, despite their philosophical proclivities, Einstein, Bohr, and the other scientific greats were not philosophical system builders, if only because they were too busy publicizing their theories and then working for world peace. Thus, the job of "making sense of how science works" was left to some people who we now consider to be philosophers of science.

If we were to call anybody the first "philosopher of science" in the modern sense of the term, then it should probably be **Moritz Schlick** (1882–1936). Schlick earned his PhD in physics at Berlin under the supervision of Max Planck and thereafter began studying philosophy. During the 1910s, Schlick became one of the first philosophical interpreters of Einstein's new theories, and in doing so, he developed a distinctive view in opposition to Marburg neo-Kantianism. In 1922, Schlick was appointed chair of *Naturphilosophie* in Vienna, a post that had earlier been held by Boltzmann and then by Mach.

When Schlick formulated his epistemological theories, he did so in a conscious attempt to accommodate the newest discoveries in mathematics and physics. With particular reference to mathematical knowledge, Schlick followed nineteenth-century mathematicians – most notably Pasch and Hilbert – in saying that mathematical claims are true by definition and that the words that occur in the axioms are thereby implicitly defined. In short, those words have no meaning beyond that which accrues to them by their role in the axioms.

While Schlick was planting the roots of philosophy of science in Vienna, the young **Hans Reichenbach** (1891–1953) had found a way to combine the study of philosophy, physics, and mathematics by moving around between Berlin, Göttingen, and Munich – where he studied philosophy with Cassirer, physics with Einstein, Planck, and Sommerfeld; and mathematics with Hilbert and Noether. He struggled at first to find a suitable academic post, but eventually Reichenbach was appointed at Berlin in 1926. It was in Berlin that Reichenbach took on a student named Carl Hempel (1905–1997), who would later bring this new philosophical approach to the elite universities in the United States. Hempel's students include several of the major players in twentieth-century philosophy of science, such as Adolf Grünbaum, John Earman, and Larry Sklar. Reichenbach himself eventually relocated to UCLA, where he had two additional students of no little renown: Wesley Salmon and Hilary Putnam.

However, back in the 1920s, shortly before he took the post at Berlin, Reichenbach had another auspicious meeting at a philosophy conference in Erlangen. Here he met a young man named Rudolf Carnap who, like Reichenbach, found himself poised at the intersection of philosophy, physics, and mathematics. Reichenbach introduced Carnap

to his friend Schlick, the latter of whom took an eager interest in Carnap's ambition to develop a "scientific philosophy." A couple of short years later, Carnap was appointed assistant professor of philosophy in Vienna – and so began the marriage between mathematical logic and philosophy of science.

Carnap

Having been a student of Frege's in Jena, Rudolf Carnap (1891–1970) was an early adopter of the new logical methods. He set to work immediately trying to employ these methods in the service of a new style of philosophical inquiry. His first major work – *Der Logische Aufbau der Welt* (1928) – attempted the ultra-ambitious project of constructing all scientific concepts out of primitive (fundamental) concepts. What is especially notable for our purposes was the notion of *construction* that Carnap employed, for it was a nearby relative to the notion of *logical construction* that Russell had employed, and which descends from the mathematician's idea that one kind of mathematical object (e.g., real numbers) can be constructed from another kind of mathematical object (e.g., natural numbers). What's also interesting is that Carnap takes over the idea of *explication*, which arose in mathematical contexts – e.g., when one says that a function f is "continuous" just in case for each $\epsilon > 0$, there is a $\delta > 0$ such that ...

When assessing philosophical developments such as these, which are so closely tied to developments in the exact sciences, we should keep in mind that ideas that are now clear to us might have been quite opaque to our philosophical forebears. For example, these days we know quite clearly what it means to say that a theory T is complete. But to someone like Carnap in the 1920s, the notion of completeness was vague and hazy, and he struggled to integrate it into his philosophical thoughts. We should keep this point in mind as we look toward the next stage of Carnap's development, where he attempted a purely "syntactic" analysis of the concepts of science.

In the late 1920s, the student Kurt Gödel (1906–1978) joined in the discussions of the Vienna circle, and Carnap later credited Gödel's influence for turning his interest to questions about the language of science. Gödel gave the first proof of the completeness of the predicate calculus in his doctoral dissertation (1929), and two years later, he obtained his famous incompleteness theorem, which shows that there is some truth of arithmetic that cannot be derived from the first-order Peano axioms.

In proving incompleteness, Gödel's technique was "metamathematical" – i.e., he employed a theory M *about* the first-order theory T of arithmetic. Moreover, this metatheory M employed purely syntactic concepts – e.g., the length of a string of symbols, or the number of left parentheses in a string, or being the last formula in a valid proof that begins from the axioms of arithmetic. This sort of approach proved to be fascinating for Carnap, in particular, because it transformed questions that seemed hopelessly vague and "philosophical" into questions that were tractable – and indeed tractable by means of the very methods that scientists themselves employed. In short, Gödel's approach indicated the possibility of an exact science of the exact sciences.

And yet, Gödel's inquiry was restricted to one little corner of the exact sciences: arithmetic. Carnap's ambitions went far beyond elementary mathematics; he aspired to

apply these new methods to the entire range of scientific theories, and especially the new theories of physics. Nonetheless, Carnap quickly realized that he faced additional problems beyond those faced by the metamathematician, for scientific theories – unlike their mathematical cousins – purport to say something *contingently true* – i.e., something that could have been otherwise. Hence, the logical approach to philosophy of science isn't finished when one has analyzed a theory T qua mathematical object; one must also say something about how T latches on to empirical reality.

Carnap's first attempts in this direction were a bit clumsy, as he himself recognized. In the 1920s and 1930s, philosophers of science were just learning the basics of formal logic. It would take another forty years until "model theory" was a well-established discipline, and the development of mathematical logic continues today (as we hope to make clear in this book). However, when mathematical logic was still in its infancy, philosophers often tried the "most obvious" solution to their problems – not realizing that it couldn't stand up to scrutiny. Consider, for example, Carnap's attempt to specify the empirical content of a theory T. Carnap proposes that the vocabulary Σ in which a theory T is formulated must include an empirical subvocabulary $O \subseteq \Sigma$, in which case the empirical content of T can be identified with the set $T|_O$ of consequences of T restricted to the vocabulary O. Similarly, in attempting to cash out the notion of "reduction" of one theory to another, Carnap initially said that the concepts of the reduced theory needed to be explicitly defined in terms of the concepts of the reducing theory – not realizing that he was thereby committing to a far more narrow notion of reduction than was being used in the sciences.

In Carnap's various works, however, we do find the beginnings of an approach that is still relevant today. Carnap takes a "language" and a "theory" to be objects of his inquiries, and he notes explicitly that there are choices to be made along the way. So, for example, the classical mathematician chooses a certain language and then adopts certain transformation rules. In contrast, the intuitionistic mathematician chooses a different language and adopts different transformation rules. Thus, Carnap allows himself to ascend semantically – to look at scientific theories from the outside, as it were. From this vantage point, he is no longer asking the "internal questions" that the theorist herself is asking. He is not asking, for example, whether there is a greatest prime number. Instead, the philosopher of science is raising "external questions" – i.e., questions about the theory T, and especially those questions that have precise syntactic formulations. For example, Carnap proposes that the notion of a sentence's being "analytic relative to T" is an external notion that we metatheorists use to describe the structure of T.

The twentieth-century concern with analytic truth didn't arise in the seminar rooms of philosophy departments – or at least not in philosophy departments like the ones of today. In fact, this concern began rather with nineteenth-century geometers, faced with two parallel developments: (1) the discovery of non-Euclidean geometries, and (2) the need to raise the level of rigor in mathematical arguments. Together, these two developments led mathematical language to be disconnected from the physical world. In other words, one key outcome of the development of modern mathematics was the *de-interpretation* of mathematical terms such as "number" or "line." These terms were replaced by symbols that bore no intuitive connection to external reality.

It was this de-interpretation of mathematical terms that gave rise to the idea that analytic truth is *truth by postulation*, the very idea that was so troubling to Russell, and then to Quine. But in the middle of the nineteenth century, the move that Russell called "theft" enabled mathematicians to proceed with their investigations in absence of the fear that they lacked insight into the meanings of words such as "line" or "continuous function." In their view, it didn't matter what words you used, so long as you clearly explained the rules that governed their use. Accordingly, for leading mathematicians such as Hilbert, mathematical terms such as "line" mean nothing more nor less than what axioms say of them, and it's simply impossible to write down false mathematical postulates. There is no external standard against which to measure the truth of these postulates.

It's against this backdrop that Carnap developed his notion of analytic truth in a framework; and that Quine later developed his powerful critique of the analytic–synthetic distinction. However, Carnap and Quine were men of their time, and their thoughts operated at the level of abstraction that science had reached in the 1930s. The notion of logical metatheory was still in its infancy, and it had hardly dawned on logicians that "frameworks" or "theories" could themselves be treated as objects of investigation.

Quine

If one was a philosophy student in the late twentieth century, then one learned that Quine "demolished" logical positivism. In fact, the errors of positivism were used as classroom lessons in how not to commit the most obvious philosophical blunders. How silly to state a view that, if true, entails that one cannot justifiably believe it!

During his years as an undergraduate student at Oberlin, **Willard van Orman Quine** (1908–2000) had become entranced with Russell's mathematical logic. After getting his PhD from Harvard in 1932, Quine made a beeline for Vienna just at the time that Carnap was setting his "logic of science" program into motion. Quine quickly became Carnap's strongest critic. As the story is often told, Quine was single-handedly responsible for the demise of Carnap's program, and of logical positivism more generally.

Of course, Quine was massively influential in twentieth-century philosophy – not only for the views he held, but also via the methods he used for arriving at those views. In short, the Quinean methodology looks something like this:

1. One cites some theorem ϕ in logical metatheory.
2. One argues that ϕ has certain philosophical consequences, e.g., makes a certain view untenable.

Several of Quine's arguments follow this pattern, even if he doesn't always explicitly mention the relevant theorem from logical metatheory. One case where he is explicit is in his 1940 paper with Nelson Goodman, where he "proves" that every synthetic truth can be converted to an analytic truth. Whatever one may think of Quine's later arguments against analyticity, there is no doubt, historically speaking, that this metatheoretical result played a role in Quine's arriving at the conclusion that there is no

analytic–synthetic distinction. And it would only be reasonable to think that *our* stance on the analytic–synthetic distinction should be responsive to what this mathematical result can be supposed to show.

As the story is typically told, Quine's "Two Dogmas of Empiricism" dealt the death blow to logical positivism. However, Carnap presented Quine with a moving target, as his views continued to develop. In "Empiricism, Semantics, and Ontology" (1950), Carnap further developed the notion of a *framework*, which bears striking resemblances both to the notion of a *scientific theory* and, hence, to the notion of a theory T in first-order logic. Here Carnap distinguishes two types of questions – the questions that are *internal* to the framework and the questions that are *external* to the framework. The internal questions are those that can be posed in the language of the framework and for which the framework can (in theory) provide an answer. In contrast, the external questions are those that we ask *about* a framework.

Carnap's abstract idea can be illustrated by simple examples from first-order logic. If we write down a vocabulary Σ for a first-order language, and a theory T in this vocabulary, then a typical internal question might be something like, "Does anything satisfy the predicate $P(x)$?" In contrast, a typical external question might be, "How many predicate symbols are there in Σ?" Thus, the internal–external distinction corresponds roughly to the older distinction between object language and metalanguage that frames Carnap's discussion in *Logische Syntax der Sprache* (1934).

The philosophical point of the internal–external distinction was supposed to be that one's answers to external questions are not held to the same standards as one's answers to internal questions. A framework includes rules, and an internal question should be answered in accordance with these rules. So, to take one of Carnap's favorite examples, "Are there numbers?" can naturally construed as an external question, since no mathematician is actively investigating that question. This question is *not* up for grabs in mathematical science – instead, it's a presupposition of mathematical science. In contrast, "Is there a greatest prime number?" is internal to mathematical practice; i.e., it is a question to which mathematics aspires to give an answer.

Surely most of us can grasp the intuition that Carnap is trying to develop here. The external questions must be answered in order to set up the game of science; the internal questions are answered in the process of playing the game of science. But Carnap wants to push this idea beyond the intuitive level – he wants to make it a cornerstone of his theory of knowledge. Thus, Carnap says that we may single out a certain special class of predicates – the so-called *Allwörter* – to label a domain of inquiry. For example, the number theorist uses the word "number" to pick out her domain of inquiry – she doesn't investigate whether something falls under the predicate "x is a number." In contrast, a number theorist might investigate whether there are numbers x, y, z such that $x^3 + y^3 = z^3$; and she simply doesn't consider whether some other things, which are not themselves numbers, satisfy this relation.

Quine (1951a, 1960) takes up the attack against Carnap's internal–external distinction. While Quine's attack has several distinct maneuvers, his invocation of hard logical facts typically goes unquestioned. In particular, Quine appeals to the supposedly hard logical fact that every theory in a language that has several distinct quantifiers

(i.e., many-sorted logic) is equivalent to a theory in a language with a single unrestricted quantifier.

> It is evident that the question whether there are numbers will be a category question only with respect to languages which appropriate a separate style of variables for the exclusive purpose of referring to numbers. If our language refers to numbers through variables that also take classes other than numbers as values, then the question whether there are numbers becomes a subclass question … Even the question whether there are classes, or whether there are physical objects becomes a subclass question if our language uses a single style of variables to range over both sorts of entities. Whether the statement that there are physical objects and the statement that there are black swans should be put on the same side of the dichotomy, or on opposite sides, comes to depend upon the rather trivial consideration of whether we use one style of variables or two for physical objects and classes. (Quine, 1976, p. 208)

Thus, suggests Quine, there is a metatheoretical result – that a many-sorted theory is equivalent to a single-sorted theory – that destroys Carnap's attempt to distinguish between *Allwörter* and other predicates in our theories.

We won't weigh in on this issue here, in our introduction. It would be premature to do so, because the entire point of this book is to lay out the mathematical facts in a clear fashion so that the reader can judge the philosophical claims for herself.

In "Two Dogmas of Empiricism," Quine argues that it makes no sense to talk about a statement's admitting of confirming or infirming (i.e., disconfirming) instances, at least when taken in isolation. Just a decade later, **Hilary Putnam**, in his paper "What Theories Are Not" (Putnam, 1962) applied Quine's idea to entire scientific theories. Putnam, student of the ur-positivist Reichenbach, now turns the positivists' primary weapon against them, to undercut the very distinctions that were so central to their program. In this case, Putnam argues that the set $T|_O$ of "observation sentences" does not accurately represent a theory T's empirical content. Indeed, he argued that a scientific theory cannot properly be said to have empirical content and, hence, that the warrant for believing it cannot flow from the bottom (the empirical part) to the top (the theoretical part). The move here is paradigmatic Putnam: a little bit of mathematical logic deftly invoked to draw a radical philosophical conclusion. This isn't the last time that we will see Putnam wield mathematical logic in the service of a far-reaching philosophical claim.

The Semantic Turn

In the early 1930s, the Vienna circle made contact with the group of logicians working in Warsaw, and in particular with **Alfred Tarski** (1901–1983). As far as twentieth-century analytic philosophy is concerned, Tarski's greatest influence has been through his bequest of **logical semantics**, along with his explications of the notions of **structure** and **truth in a structure**. Indeed, in the second half of the twentieth century, analytic philosophy has been deeply intertwined with logical semantics, and ideas from model theory have played a central role in debates in metaphysics, epistemology, philosophy of science, and philosophy of mathematics.

The promise of a purely syntactic metatheory for mathematics fell into question already in the 1930s when Kurt Gödel proved the incompleteness of Peano arithmetic. At the time, a new generation of logicians realized that not all interesting questions about theories could be answered merely by looking at theories "in themselves", and without relation to other mathematical objects. Instead, they claimed, the interesting questions about theories include questions about how they might relate to antecedently understood mathematical objects, such as the universe of sets. Thus was born the discipline of logical semantics. The arrival of this new approach to metatheory was heralded by Alfred Tarski's famous definitions of "truth in a structure" and "model of a theory." Thus, after Tarski, to understand a theory T, we have more than the theory qua syntactic object, we also have a veritable universe $\mathrm{Mod}(T)$ of models of T.

Bas van Fraassen was one of the earliest adopters of logical semantics as a tool for philosophy of science, and he effectively marshaled it in developing an alternative to the dominant outlook of scientific realism. Van Fraassen ceded Putnam's argument that the empirical content of a theory cannot be isolated syntactically. And then, in good philosophical fashion, he transformed Putnam's modus ponens into a modus tollens: the problem is not with empirical content, per se, but with the attempt to explicate is syntactically. Indeed, van Fraassen claimed that one needs the tools of logical semantics in order to make sense of the notion of empirical content; and equipped with this new explication of empirical content, empiricism can be defended against scientific realism. Thus, both the joust and the parry were carried on within an explicitly metalogical framework.

Since the 1970s, philosophical discussions of science have been profoundly influenced by this little debate about the place of syntax and semantics. Prior to the criticisms – by Putnam, van Fraassen, et al. – of the "syntactic view of theories" philosophical discussions of science frequently drew upon new results in mathematical logic. As was pointed out by van Fraassen particularly, these discussions frequently degenerated, as philosophers found themselves hung up on seemingly trivial questions, e.g., whether the observable consequences of a recursively axiomatized theory are also recursively axiomatizable. Part of the shift from syntactic to semantic methods was supposed to be a shift toward a more faithful construal of science in practice. In other words, philosophers were supposed to start asking the questions that arise in the practice of science, rather than the questions that were suggested by an obsessive attachment to mathematical logic.

The move away from logical syntax has had some healthy consequences in terms of philosophers engaging more closely with actual scientific theories. It is probably not a coincidence that since the fall of the syntactic view of theories, philosophers of science have turned their attention to specific theories in physics, biology, chemistry, etc. As was correctly pointed out by van Fraassen, Suppes, and others, scientists themselves don't demand first-order axiomatizations of these theories – and so it would do violence to those theories to try to encode them in first-order logic. Thus, the demise of the syntactic view allowed philosophers to freely draw upon the resources of set-theoretic structures, such as topological spaces, Riemannian manifolds, Hilbert spaces, C^*-algebras, etc.

Nonetheless, the results of the semantic turn have not been uniformly positive. For one, philosophy of science has seen a decline in standards of rigor, with the unfortunate consequence that debating parties more often than not talk past each other. For example, two philosophers of science might take up a debate about whether isomorphic models represent the same or different possibilities. However, these two philosophers of science may not have a common notion of "model" or of "isomorphism." In fact, many philosophers of science couldn't even tell you a precise formal explication of the word "isomorphism" – even though they rely on the notion in many of their arguments. Instead, their arguments rely on some vague sense that isomorphisms preserve structure, and an even more vague sense of what structure is.

In this book, we'll see many cases in point, where a technical term from science (physics, math, or logic) has made its way into philosophical discussion but has then lost touch with its technical moorings. The result is almost always that philosophers add to the stock of confusion rather than reducing it. How unfortunate it is that philosophy of science has fallen into this state, given the role we could play as prophets of clarity and logical rigor. One notable instance where philosophers of science could help increase clarity is the notion of *theoretical equivalence*. Scientists, and especially physicists, frequently employ the notion of two theories being equivalent. Their judgments about equivalence are not merely important for their personal attitudes toward their theories, but also for determining their actions – e.g., will they search for a crucial experiment to determine whether T_1 or T_2 is true? For example, students of classical mechanics are frequently told that the Lagrangian and Hamiltonian frameworks are equivalent, and on that basis, they are discouraged from trying to choose between them.

Now, it's not that philosophers don't talk about such issues. However, in my experience, philosophers tend to bring to bear terminology that is alien to science, and which sheds no further light on the problems. For example, if an analytic philosopher is asked, "when do two sentences ϕ and ψ mean the same thing?" then he is likely to say something like, "if they pick out the same proposition." Here the word "proposition" is alien to the physicist; and what's more, it doesn't help to solve real-life problems of synonymy. Similarly, if an analytic philosopher is asked, "when do two theories T_1 and T_2 say the same thing?" then he might say something like, "if they are true in the same possible worlds." This answer may conjure a picture in the philosopher's head, but it won't conjure any such picture in a physicist's head – and even if it did, it wouldn't help decide controversial cases. We want to know whether Lagrangian mechanics is equivalent to Hamiltonian mechanics, and whether Heisenberg's matrix mechanics is equivalent to Schrödinger's wave mechanics. The problem here is that space of possible worlds (if it exists) cannot be surveyed easily, and the task of comparing the subset of worlds in which T_1 is true with the subset of worlds in which T_2 is true is hardly tractible. Thus, the analytic philosopher's talk about "being true in the same possible worlds" doesn't amount to an *explication* of the concept of equivalence. An explication, in the Carnapian sense, should supply clear guidelines for how to use a concept.

Now, don't get me wrong. I am not calling for a Quinean ban on propositions, possible worlds, or any of the other concepts that analytic philosophers have found so interesting. I only want to point out that these concepts are descendants, or cousins, of similar

concepts that are used in the exact sciences. Thus, it's important that analytic philosophers – to the extent that they want to understand and/or clarify science – learn to tie their words back down into their scientific context. For example, philosophers' possible worlds are the descendant of the logician's "models of a theory," the mathematician's "solutions of a differential equation," and the physicist's "points in state space." Thus, it's fine to talk about possible worlds, but it would be advisable to align our usage of the concept with the way it's used in the sciences.

As we saw before, Carnap had self-imposed the constraint that a philosophical explication of a concept must be *syntactic*. So, for example, to talk about "observation sentences," one must construct a corresponding predicate in the language of syntactic metalogic – a language whose primitive concepts are things like "predicate symbol" and "binary connective." Carnap took a swing at defining such predicates, and Quine, Putnam, and friends found his explications to be inadequate. There are many directions that one could go from here – and one of these directions remains largely unexplored. First, one can do as Quine and Putnam themselves did: stick with logical syntax and change one's philosophical views. Second, one can do as van Fraassen did: move to logical semantics and stick with Carnap's philosophical views. (To be fair, van Fraassen's philosophical views are very different than Carnap's – I only mean to indicate that there are certain central respects in which van Fraassen's philosophical views are closer to Carnap's than to Quine's.) The third option is to say perhaps logical syntax had not yet reached a fully mature stage in 1950, and perhaps new developments will make it more feasible to carry out syntactic explications of philosophical concepts. That third option is one of the objectives of this book – i.e. to raise syntactic analysis to a higher level of nuance and sophistication.

Model Theoretic Madness

By the 1970s, scientific realism was firmly entrenched as the dominant view in philosophy of science. Most the main players in the field – Boyd, Churchland, Kitcher, Lewis, Salmon, Sellars, etc. – had taken up the realist cause. Then, with a radical about-face, Putnam again took up the tools of mathematical logic, this time to argue for the incoherence of realism. In his famous "model-theoretic argument," Putnam argued that logical semantics – in particular, the Löwenheim-Skølem theorem – implies that any consistent theory is true. In effect, then, Putnam proposed a return to a more liberal account of theoretical equivalence, indeed, something even more liberal than the logical positivists' notion of empirical equivalence. Indeed, in the most plausible interpretation of Putnam's conclusion, it entails that any two consistent theories are equivalent to each other.

Whatever you might think of Putnam's radical claim, there is no doubt that it stimulated some interesting responses. In particular, Putnam's claim prompted the arch-realist David Lewis to clarify the role that *natural properties* play in his metaphysical system. According to Lewis, the defect in Putnam's argument is the assumption that a predicate P can be assigned to any subset of objects in the actual world. This assumption is mistaken, says Lewis, because not every random collection of things corresponds to

some natural class, and we should only consider interpretations in which predicates that occur in T are assigned to natural classes of objects in the actual world. Even if T is consistent, there may be no such interpretation relative to which T is actually true.

There are mixed views on whether Lewis' response to Putnam is effective. However, for our purposes, the important point is that the upshot of Lewis' response would be to move in the direction of a more conservative account of theoretical equivalence. And now the question is whether the notion of theoretical equivalence that Lewis is proposing goes too far in the other direction. On one interpretation of Lewis, his claim is that two theories T and T' are equivalent only if they share the same "primitive notions." If we apply that claim literally to first-order theories, then we might think that theories T and T' are equivalent only if they are written with the same symbols. However, this condition wouldn't even allow notationally variant theories to be equivalent.

While Lewis was articulating the realist stance, Putnam was digging up more arguments for a liberal and inclusive criterion of theoretical equivalence. Here he drew on his extensive mathematical knowledge to find examples of theories that mathematicians call equivalent, but which metaphysical realists would call inequivalent. One of Putnam's favorite examples here was axiomatic Euclidean geometry, which some mathematicians formulate with points as primitives, and other mathematicians formulate with lines as primitives — but they never argue that one formulation is more correct than the other. Thus, Putnam challenges the scientific credentials of realism by giving examples of theories that scientists declare to be equivalent, but which metaphysical realists would declare to be inequivalent.

At the time when Putnam put forward these examples, analytic philosophy was unfortunately growing more distant from its logical and mathematical origins. What this meant, in practice, is that while Putnam's examples were extensively discussed, the discussion never reached a high level of logical precision. For example, nobody clearly explained how the word "equivalence" was being used.

These exciting, and yet imprecise, discussions continued with reference to a second example that Putnam had given. In this second example, Putnam asks how many things are on the following line:

* *

There are two schools of metaphysicians who give different answers to this question. According to the mereological nihilists, there are exactly two things on the line, and both are asterisks. According to the mereological universalists, there are three things on the line: two individual asterisks, and one composite of two asterisks. Putnam, however, declares that the debate between these two schools of metaphysicians is a "purely verbal dispute", and neither party is more correct than the other.

Again, what's important for us here it that Putnam's claim amounts to a proposal to liberalize the standards of theoretical equivalence. By engaging in this dispute, metaphysicians have implicitly adopted a rather conservative standard of equivalence – where it matters whether you think that a pair of asterisks is something more beyond the individuals that constitute it. Putnam urges us to adopt a more liberal criterion of

theoretical equivalence, according to which it simply doesn't matter whether we say that the pair "really exists", or whether we don't.

From Reduction to Supervenience

The logical positivists – Schlick, Carnap, Neurath, etc. – aspired to uphold the highest standards of scientific rationality. Most of them believed that commitment to scientific rationality demands a commitment to physicalism, i.e. the thesis that physical science is the final arbiter on claims of ontology. In short, they said that we ought to believe that something exists only if physics licenses that belief.

Of course, we don't much mind rejecting claims about angels, demons, witches, and fairies. But what are we supposed to do with the sorts of statements that people make in the ordinary course of life – about each other, and about themselves? For example, if I say, "Søren is in pain," then I seem to be committed to the existence of some object denoted by "Søren", that has some property "being in pain." How can physical science license such a claim, when it doesn't speak of an object Søren or the property of being in pain?

The general thesis of physicalism, and the particular thesis that a person is his body, were not 20th century novelties. However, it was a 20th century novelty to attempt to explicate these theses using the tools of symbolic logic. To successfully explicate this concept would transform it from a vague ideological stance to a sharp scientific hypothesis. (There is no suggestion here that the hypothesis would be empirically verifiable – merely that it would be clear enough to be vulnerable to counterargument.)

For example, suppose that $r(x)$ denotes the property of being in pain. Then it would be natural for the physicalist to propose either (1) that statements using $r(x)$ are actually erroneous, or (2) that there is some predicate $\phi(x)$ in the language of fundamental physics such that $\forall x(r(x) \leftrightarrow \phi(x))$. In other words, if statements using $r(x)$ are legitimate, then $r(x)$ must actually pick out some underlying physical property $\phi(x)$.

The physicalist will want to clarify what he means by saying that $\forall x(r(x) \leftrightarrow \phi(x))$, for even a Cartesian dualist could grant that this sentence is contingently true. That is, a Cartesian dualist might say that there is a physical description $\phi(x)$ which 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. He wants to say that there is a more thick connection between pain experiences and happenings in the physical world. At the very least, a reductionist would say that

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

where T is our most fundamental theory of the physical world. That is, to the extent that ordinary language ascriptions are correct, they can be translated into true statements of fundamental physics.

This sort of linguistic reductionism seems to have been the favored view among early-twentieth-century analytic philosophers – or, at least among the more scientifically inclined of them. Certainly, reductionism had vocal proponents, such as U.T. Place and Herbert Feigl. Nonetheless, by the third quarter of the twentieth century, this view

had fallen out of fashion. In fact, some of the leading lights in analytic philosophy – such as Putnam and Fodor – had 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.

There was another movement afoot in analytic philosophy – a movement away from the formal mode, back toward the material mode, i.e., from a syntactic point of view, to a semantic point of view. What this movement entailed in practice was a shift from syntactic explications of concepts to semantic explications of concepts. Thus, it is only natural that having discarded the syntactic explication of mind–body reduction, analytic philosophers would cast about for a semantic explication of the idea. Only, in this case, the very word "reduction" had so many negative associations that a new word was needed. To this end, analytic philosophers co-opted the word "supervenience." Thus Donald Davidson:

> Mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect. (Davidson, 1970)

Davidson's prose definition of supervenience is so clear that it is begging for formalization. Indeed, as we'll later see, when the notion of supervenience is formalized, then it is none other than the model theorist's notion of implicit definability.

It must have seemed to the 1970s philosophers that significant progress had been made in moving from the thin syntactic concept of reduction to the thick semantic concept of supervenience. Indeed, by the 1980s, the concept of supervenience had begun to play a major role in several parts of analytic philosophy. However, with the benefit of hindsight, we ought to be suspicious if we are told that an implausible philosophical position can be converted into a plausible one merely by shifting from a syntactic to a semantic explication of the relevant notions. In this case, there is a worry that the concept of supervenience is nothing but a reformulation, in semantic terms, of the notion of reducibility. As we will discuss in Section 6.7, if supervenience is cashed out as the notion of implicit definability, then **Beth's theorem** shows that supervenience is equivalent to reducibility.

Why did philosophers decide that mind-brain reductionism was implausible? We won't stop here to review the arguments, as interesting as they are, since that has been done in many other places (see Bickle, 2013). We are interested rather in claims (see, e.g., Bickle (1998)) that the arguments against reduction are only effective against syntactic accounts thereof – and that semantics permits a superior account of reduction that is immune to these objections.

Throughout this book, we argue for a fundamental duality between logical syntax and semantics. To the extent that this duality holds, it is mistaken to think that semantic accounts of concepts are more intrinsic, or that they allow us to transcend the human reliance on representations, or that they provide a bridge to the "world" side of the mind-world divide.

To the contrary, logical semantics is … wait for it … just more mathematics. As such, while semantics can be used to represent things in the world, including people and their practice of making claims about the world, its means of representation are no different than those of any other part of mathematics. Hence, every problem and puzzle and confusion that arises in logical syntax – most notably, the problem of language dependence – will rear its ugly head again in logical semantics. Thus, for example, if scientific antirealism falls apart when examined under a syntactic microscope, then it will also fall apart when examined under a semantic microscope. Similarly, if mind-body reductionism isn't plausible when explicated syntactically, then it's not going to help to explicate it semantically.

What I am saying here should not be taken as a blanket criticism of attempts to explicate concepts semantically. In fact, I'll be the first to grant that model theory is not only a beautiful mathematical theory, but is also particularly useful for philosophical thinking. However, we should be suspicious of any claims that a philosophical thesis (e.g. physicalism, antirealism, etc.) is untenable when explicated syntactically, but becomes tenable when explicated semantically. We should also be suspicious of any claims that semantic methods are any less prone to creating pseudoproblems than syntactic methods.

Realism and Equivalence

As we have seen, many of these debates in twentieth-century philosophy ultimately turn on the question of how one theory is related to another. For example, the debate about the mind-body relation can be framed as a question about how our folk theory of mind is related to the theories of the brain sciences, and ultimately to the theories of physics.

If we step up a level of abstraction, then even the most general divisions in 20th century philosophy have to do with views on the relations of theories. Among the logical positivists, the predominant view was a sort of antirealism, certainly about metaphysical claims, but also about the theoretical claims of science. Not surprisingly, the preferred view of theoretical equivalence among the logical positivists was empirical equivalence: two theories are equivalent just in case they make the same predictions. That notion of equivalence is quite liberal in that it equates theories that intuitively seem to be inequivalent.

If we leap forward to the end of the twentieth century, then the outlook had changed radically. Here we find analytic metaphysicians engaged in debates about mereological nihilism versus universalism, or about presentism versus eternalism. We also find philosophers of physics engaged in debates about Bohmian mechanics versus Everettian interpretations of quantum mechanics, or about substantivalism versus relationalism about spacetime. The interesting point here is that there obviously had been a radical change in the regnant standard of theoretical equivalence in the philosophical community. Only seventy years prior, these debates would have been considered pseudo-debates, for they attempt to choose between theories that are empirically equivalent. In short, the philosophical community as a whole had shifted from a more liberal to a more conservative standard of theoretical equivalence.

There have been, however, various defections from the consensus view on theoretical equivalence. The most notable example here is the Hilary Putnam of the 1970s. At this time, almost all of Putnam's efforts were devoted to liberalizing standards of theoretical equivalence. We can see this not only in his model-theoretic argument, but also in the numerous examples that he gave of theories with "different ontologies," but which he claimed are equivalent. Putnam pointed to different formulations of Euclidean geometry, and also the famous example of "Carnap and the mereologist," which has since become a key example of the quantifier variance debate. We discuss the geometry example in Section 7.4, and the mereology example in Section 5.4.

One benefit of the formal methods developed in this book is a sort of taxonomy of views in twentieth-century philosophy. The realist tendency is characterized by the adoption of more conservative standards of theoretical equivalence; and the antirealist tendency is characterized by the adoption of more liberal standards of theoretical equivalence. Accordingly, we shouldn't think of "realism versus antirealism" on the model of American politics, with its binary division between Republicans and Democrats. Indeed, philosophical opinions on the realism–antirealism question lie on a continuum, corresponding to a continuum of views on theoretical equivalence. (In fact, views on theoretical equivalence really form a multidimensional continuum; I'm merely using the one-dimensional language for intuition's sake.) Most of us will find ourselves with a view of theoretical equivalence that is toward the middle of the extremes, and many of the philosophical questions we consider are questions about whether to move – if ever so slightly – in one direction or the other.

In this book, we will develop three moderate views of theoretical equivalence. The first two views say that theories are equivalent just in case they are intertranslatable – only they operate with slightly different notions of "translation." The first, and more conservative, view treats quantifier statements as an invariant, so that a good translation must preserve them intact. (We also show that this first notion of intertranslatability corresponds to "having a common definitional extension." See Theorems 4.6.17 and 6.6.21.) The second, and more liberal, view allows greater freedom in translating one language's quantifier statements into a complex of the other language's quantifier statements. (We also show that this second notion of intertranslatability corresponds to "having a common Morita extension." See Theorems 7.5.3 and 7.5.5.) The third view of equivalence we consider is the most liberal, and is motivated not by linguistic considerations, but by scientific practice. In particular, scientists seem to treat theories as equivalent if they can "do the same things with them." We will explicate this notion of what a scientific theory can do in terms of its "category of models." We then suggest that two theories are equivalent in this sense if their categories of models are equivalent in the precise, category-theoretic sense.

Summary and Prospectus

The following seven chapters try to accomplish two things at once: to introduce some formal techniques, and to use these techniques to gain philosophical insight. Most of

the philosophical discussions are interspersed between technical results, but there is one concluding chapter that summarizes the major philosophical themes. We include here a chart of some of the philosophical issues that arise in the course of these chapters. The left column states a technical result, the middle column states the related philosophical issue, and the right column gives the location (section number) where the discussion can be found. To be fair, I don't mean to say that the philosophers mentioned on the right explicitly endorse the argument from the metalogical result to the philosophical conclusion. In some cases they do; but in other cases, the philosopher seems rather to presuppose the metalogical result.

Logic	Philosophy	Location
Translate into empty theory	Analytic–synthetic distinction (Quine)	**3.7.10**
Translate into empty theory	Implicit definition (Quine)	**3.7.10**
Eliminate sorts	Ontological monism (Quine)	**5.3**
Eliminate sorts	No external questions (Quine)	**5.4.17**
Eliminate sorts	Against quantifier variance	**5.4.4, 5.4.16**
Indivisible vocabulary	Against empiricism (Putnam, Boyd)	**4.4**
Beth's theorem	Supervenience implies reduction	**6.7**
Löwenheim–Skølem	Against realism (Putnam)	**8.3**
Equivalent geometries	Against realism (Putnam)	**7.4**
Ramsefication	Structural realism	**8.1**
Ramsefication	Functionalism	**8.1**

Notes

- In this chapter, our primary objective was to show the philosopher-in-training some of the payoffs for learning the metatheory of first-order logic: the better she understands the logic, the better she will understand twentieth-century philosophy, and the options going forward. Although we've tried to be reasonably faithful to the historical record, we've focused on just one part of this history. The curious reader should consult more detailed studies, such as Coffa (1993); Friedman (1999); Hylton (2007); Soames (2014).
- For Russell's program for rebuilding philosophy on the basis of formal logic, see Russell (1901, 1914a).
- Carnap's personal recollections can be found in Carnap and Schilpp (1963).
- Frege and Russell were early critics of Hilbert's view of implicit definition (see, e.g., Blanchette, 2012). In contrast, Schlick (1918, I.7) explicitly endorses Hilbert's view. For Carnap's view, see Park (2012). The discussion later got muddled up in discussions of Ramsey sentences (see, e.g., Winnie, 1967; Lewis,

1970), which we will discuss in Chapter 8. For an extended discussion of implicit definition and its relation to 20th century philosophical issues, see Ben-Menahem (2006).

- For more on the 19th century backdrop to analyticity, see Coffa (1986).
- For overviews of logical methods in philosophy of science, see van Benthem (1982); Winnie (1986); Van Fraassen (2011); Leitgeb (2011). The primary novelty of the present book is our use of category-theoretic methods. We have tried not to mention category theory more than necessary, but we use it frequently.

1 Invitation to Metatheory

This chapter is meant to serve as a preview, and for motivation to work through the chapters to come. In the next chapter, we'll move quickly into "categorical set theory" – which isn't all that difficult, but which is not yet well known among philosophers. For the past fifty years or so, it has almost been mandatory for analytic philosophers to know a little bit of set theory. However, it has most certainly not been mandatory for philosophers to know a little bit of category theory. Indeed, most analytic philosophers are familiar with the words "subset" and "powerset" but not the words "natural transformation" or "equivalence of categories." Why should philosophers bother learning these unfamiliar concepts?

The short answer is that is that category theory (unlike set theory) was designed to explicate *relations* between mathematical structures. Since philosophers want to think about relations between theories (e.g., equivalence, reducibility) and since theories can be modeled as mathematical objects, philosophers' aims will be facilitated by gaining some fluency in the language of category theory. At least that's one of the main premises of this book. So, in this chapter, we'll review some of the basics of the metatheory of propositional logic. We will approach the issues from a slightly different angle than usual, placing less emphasis on what individual theories say and more emphasis on the relations between these theories.

To repeat, the aim of **metatheory** is to theorize about theories. For simplicity, let's use M to denote this hypothetical theory about theories. Thus, M is not the *object* of our study; it is the *tool* we will use to study other theories and the relations between them. In this chapter, I will begin using this tool M to talk about theories – without explicitly telling you anything about M itself. In the next chapter, I'll give you the user's manual for M.

1.1 Logical Grammar

DEFINITION 1.1.1 A **propositional signature** Σ is a collection of items, which we call **propositional constants**. Sometimes these propositional constants are also called **elementary sentences**. (Sometimes people call them atomic sentences, but we will be using the word "atomic" for a different concept.)

These propositional constants are assumed to have no independent meaning. Nonetheless, we assume a primitive notion of identity between propositional constants; the fact

that two propositional constants are equal or non-equal is not explained by any more fundamental fact. This assumption is tantamount to saying that Σ is a **bare set** (and it stands in gross violation of Leibniz's principle of the identity of indiscernibles).

ASSUMPTION 1.1.2 The **logical vocabulary** consists of the symbols $\neg$, $\wedge$, $\vee$, $\rightarrow$. We also use two further symbols for punctuation: a left and a right parenthesis.

DEFINITION 1.1.3 Given a propositional signature Σ, we define the set $\mathsf{Sent}(\Sigma)$ of Σ-sentences as follows:

1. If $\phi \in \Sigma$, then $\phi \in \mathsf{Sent}(\Sigma)$.
2. If $\phi \in \mathsf{Sent}(\Sigma)$, then $(\neg\phi) \in \mathsf{Sent}(\Sigma)$.
3. If $\phi \in \mathsf{Sent}(\Sigma)$ and $\psi \in \mathsf{Sent}(\Sigma)$, then $(\phi \wedge \psi) \in \mathsf{Sent}(\Sigma)$, $(\phi \vee \psi) \in \mathsf{Sent}(\Sigma)$, and $(\phi \rightarrow \psi) \in \mathsf{Sent}(\Sigma)$.
4. Nothing is in $\mathsf{Sent}(\Sigma)$ unless it enters via one of the previous clauses.

The symbol ϕ here is a variable that ranges over finite strings of symbols drawn from the alphabet that includes Σ; the connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$; and (when necessary) left and right parentheses "(" and ")". We will subsequently play it fast and loose with parentheses, omitting them when no confusion can result. In particular, we take a negation symbol $\neg$ always to have binding precedence over the binary connectives.

Note that each sentence is, by definition, a finite string of symbols and, hence, contains finitely many propositional constants.

Since the set $\mathsf{Sent}(\Sigma)$ is defined inductively, we can prove things about it using "proof by induction." A proof by induction proceeds as follows:

1. Show that the property of interest, say P, holds of the elements of Σ.
2. Show that if P holds of ϕ, then P holds of $\neg\phi$.
3. Show that if P holds of ϕ and ψ, then P holds of $\phi \wedge \psi$, $\phi \vee \psi$, and $\phi \rightarrow \psi$.

When these three steps are complete, one may conclude that all things in $\mathsf{Sent}(\Sigma)$ have property P.

DEFINITION 1.1.4 A **context** is essentially a finite collection of sentences. However, we write contexts as sequences, for example $\phi_1, \ldots, \phi_n$ is a context. But ϕ_1, ϕ_2 is the same context as ϕ_2, ϕ_1, and is the same context as ϕ_1, ϕ_1, ϕ_2. If Δ and Γ are contexts, then we let Δ, Γ denote the union of the two contexts. We also allow an empty context.

1.2 Proof Theory

We now define the relation $\Delta \vdash \phi$ of derivability that holds between contexts and sentences. This relation is defined **recursively** (aka **inductively**), with base case $\phi \vdash \phi$ (Rule of Assumptions). Here we use a horizontal line to indicate that if $\vdash$ holds between the things above the line, then $\vdash$ also holds for the things below the line.

Rule of Assumptions $\dfrac{}{\phi \vdash \phi}$

$\wedge$ elimination $\dfrac{\Gamma \vdash \phi \wedge \psi}{\Gamma \vdash \phi}$ $\qquad$ $\dfrac{\Gamma \vdash \phi \wedge \psi}{\Gamma \vdash \psi}$

$\wedge$ introduction $\dfrac{\Gamma \vdash \phi \qquad \Delta \vdash \psi}{\Gamma, \Delta \vdash \phi \wedge \psi}$

$\vee$ introduction $\dfrac{\Gamma \vdash \phi}{\Gamma \vdash \phi \vee \psi}$ $\qquad$ $\dfrac{\Gamma \vdash \psi}{\Gamma \vdash \phi \vee \psi}$

$\vee$ elimination $\dfrac{\Gamma \vdash \phi \vee \psi \qquad \Delta, \phi \vdash \chi \qquad \Theta, \psi \vdash \chi}{\Gamma, \Delta, \Theta \vdash \chi}$

$\rightarrow$ elimination $\dfrac{\Gamma \vdash \phi \rightarrow \psi \qquad \Delta \vdash \phi}{\Gamma, \Delta \vdash \psi}$

$\rightarrow$ introduction $\dfrac{\Gamma, \phi \vdash \psi}{\Gamma \vdash \phi \rightarrow \psi}$

RA $\dfrac{\Gamma, \phi \vdash \psi \wedge \neg\psi}{\Gamma \vdash \neg\phi}$

DN $\dfrac{\Gamma \vdash \neg\neg\phi}{\Gamma \vdash \phi}$

The definition of the turnstyle $\vdash$ is then completed by saying that $\vdash$ is the smallest relation (between sets of sentences and sentences) such that

1. $\vdash$ is closed under the previously given clauses, and
2. If $\Delta \vdash \phi$ and $\Delta \subseteq \Delta'$, then $\Delta' \vdash \phi$.

The second property here is called **monotonicity**.

There are a variety of ways that one can explicitly generate pairs Δ, ϕ such that $\Delta \vdash \phi$. A method for doing such is typically called a **proof system**. We will not explicitly introduce any proof system here, but we will adopt the following definitions.

DEFINITION 1.2.1 A pair Δ, ϕ is called a **sequent** or **proof** just in case $\Delta \vdash \phi$. A sentence ϕ is said to be **provable** just in case $\vdash \phi$. Here $\vdash \phi$ is shorthand for $\emptyset \vdash \phi$. We use $\top$ as shorthand for a sentence that is provable – for example, $p \rightarrow p$. We could then add as an inference rule "$\top$ introduction," which allowed us to write $\Delta \vdash \top$. It can be proven that the resulting definition of $\vdash$ would be the same as the original definition. We also sometimes use the symbol $\bot$ as shorthand for $\neg\top$. It might then be convenient to restate RA as a rule that allows us to infer $\Delta \vdash \neg\phi$ from $\Delta, \phi \vdash \bot$. Again, the resulting definition of $\vdash$ would be the same as the original.

DISCUSSION 1.2.2 The rules we have given for $\vdash$ are sometimes called the **classical propositional calculus** or just the **propositional calculus**. Calling it a "calculus" is

meant to indicate that the rules are purely formal and don't require any understanding of the meaning of the symbols. If one deleted the DN rule and replaced it with Ex Falso Quodlibet, the resulting system would be the **intuitionistic propositional calculus**. However, we will not pursue that direction here.

1.3 Semantics

DEFINITION 1.3.1 An **interpretation** (sometimes called a **valuation**) of Σ is a function from Σ to the set $\{\text{true}, \text{false}\}$, i.e., an assignment of truth-values to propositional constants. We will usually use 1 as shorthand for "true" and 0 as shorthand for "false."

Clearly, an interpretation v of Σ extends naturally to a function $v : \mathsf{Sent}(\Sigma) \to \{0, 1\}$ by the following clauses:

1. $v(\neg\phi) = 1$ if and only if $v(\phi) = 0$.
2. $v(\phi \wedge \psi) = 1$ if and only if $v(\phi) = 1$ and $v(\psi) = 1$.
3. $v(\phi \vee \psi) = 1$ if and only if either $v(\phi) = 1$ or $v(\psi) = 1$.
4. $v(\phi \to \psi) = v(\neg\phi \vee \psi)$.

DISCUSSION 1.3.2 The word "interpretation" is highly suggestive, but it might lead to confusion. It is sometimes suggested that elements of $\mathsf{Sent}(\Sigma)$ are part of an uninterpreted calculus without intrinsic meaning, and that an intepretation $v : \Sigma \to \{0, 1\}$ endows these symbols with meaning. However, to be clear, $\mathsf{Sent}(\Sigma)$ and $\{0, 1\}$ are both mathematical objects; neither one of them is more linguistic than the other, and neither one of them is more "concrete" than the other.

This point becomes even more subtle in predicate logic, where we might be tempted to think that we can interpret the quantifiers so that they range over all actually existing things. To the contrary, the domain of a predicate logic interpretation must be a *set*, and a set is something whose existence can be demonstrated by ZF set theory. Since the existence of the world is not a consequence of ZF set theory, it follows that the world is not a set. (Put slightly differently: a set is an abstract object, and the world is a concrete object. Therefore, the world is not a set.)

DEFINITION 1.3.3 A **propositional theory** T consists of a signature Σ, and a set Δ of sentences in Σ. Sometimes we will simply write T in place of Δ, although it must be understood that the identity of a theory also depends on its signature. For example, the theory consisting of a single sentence p is different depending on whether it's formulated in the signature $\Sigma = \{p\}$ or in the signature $\Sigma' = \{p, q\}$.

DEFINITION 1.3.4 (Tarski truth) Given an interpretation v of Σ and a sentence ϕ of Σ, we say that ϕ is **true** in v just in case $v(\phi) = 1$.

DEFINITION 1.3.5 For a set Δ of Σ sentences, we say that v is a **model** of Δ just in case $v(\phi) = 1$, for all ϕ in Δ. We say that Δ is **consistent** if Δ has at least one model, and that Δ is **inconsistent** if it has no models.

Any time we define a concept for sets of sentences (e.g., consistency), we can also extend that concept to theories, as long as it's understood that a theory is technically a pair consisting of a signature and a set of sentences in that signature.

DISCUSSION 1.3.6 The use of the word "model" here has its origin in consistency proofs for non-Euclidean geometries. In that case, one shows that certain non-Euclidean geometries can be translated into models of Euclidean geometry. Thus, if Euclidean geometry is consistent, then non-Euclidean geometry is also consistent. This kind of maneuver is what we now call a **proof of relative consistency**.

In our case, it may not be immediately clear what sits on the "other side" of an interpretation, because it's certainly not Euclidean geometry. What kind of mathematical thing are we interpreting our logical symbols into? The answer here – as will become apparent in Chapter 3 – is either a Boolean algebra or a fragment of the universe of sets.

DEFINITION 1.3.7 Let Δ be a set of Σ sentences, and let ϕ be a Σ sentence. We say that Δ **semantically entails** ϕ, written $\Delta \models \phi$, just in case ϕ is true in all models of Δ. That is, if v is a model of Δ, then $v(\phi) = 1$.

EXERCISE 1.3.8 Show that if $\Delta, \phi \models \psi$, then $\Delta \models \phi \to \psi$.

EXERCISE 1.3.9 Show that $\Delta \models \phi$ if and only if $\Delta \cup \{\neg\phi\}$ is inconsistent. Here $\Delta \cup \{\neg\phi\}$ is the theory consisting of $\neg\phi$ and all sentences in Δ.

We now state three main theorems of the metatheory of propositional logic.

THEOREM 1.3.10 (Soundness) *If $\Delta \vdash \phi$, then $\Delta \models \phi$.*

The soundness theorem can be proven by an argument directly analogous to the substitution theorem in Section 1.4. We leave the details to the reader.

THEOREM 1.3.11 (Completeness) *If $\Delta \models \phi$, then $\Delta \vdash \phi$.*

The completeness theorem can be proven in various ways. In this book, we will give a topological proof via the Stone duality theorem (see Chapter 3).

THEOREM 1.3.12 (Compactness) *Let Δ be a set of sentences. If every finite subset Δ_F of Δ is consistent, then Δ is consistent.*

The compactness theorem can be proven in various ways. One way of proving it – although not the most illuminating – is as a corollary of the completeness theorem. Indeed, it's not hard to show that if $\Delta \vdash \phi$, then $\Delta_F \vdash \phi$ for some finite subset Δ_F of Δ. Thus, if Δ is inconsistent, then $\Delta \vdash \bot$, hence $\Delta_F \vdash \bot$ for a finite subset Δ_F of Δ. But then Δ_F is inconsistent.

DEFINITION 1.3.13 A theory T, consisting of axioms Δ in signature Σ, is said to be **complete** just in case Δ is consistent and for every sentence ϕ of Σ, either $\Delta \models \phi$ or $\Delta \models \neg\phi$.

Be careful to distinguish between the completeness of our proof system (which is independent of any theory) and completeness of some particular theory T. Famously, Kurt Gödel proved that the theory of Peano arithmetic is incomplete – i.e., there is a sentence ϕ of the language of arithmetic such that neither $T \vdash \phi$ nor $T \vdash \neg\phi$. However, there are much simpler examples of incomplete theories. For example, if $\Sigma = \{p, q\}$, then the theory with axiom $\vdash p$ is incomplete in Σ.

DEFINITION 1.3.14 Let T be a theory in Σ. The **deductive closure** of T, written $\mathrm{Cn}(T)$, is the set of Σ sentences that is implied by T. If $T = \mathrm{Cn}(T)$, then we say that T is **deductively closed**.

Example 1.3.15 Let $\Sigma = \{p\}$, and let $T = \{p\}$. Let $\Sigma' = \{p, q\}$, and let $T' = \{p\}$. Here we must think of T and T' as different theories, even though they consist of the same sentences – i.e., $T = T'$. One reason to think of these as different theories: T is complete, but T' is incomplete. Another reason to think of T and T' as distinct is that they have different deductive closures. For example, $q \vee \neg q$ is in the deductive closure of T', but not of T.

The point here turns out to be philosophically more important than one might think. Quine argued (correctly, we think) that choosing a theory is not just choosing axioms, but axioms in a particular language. Thus, one can't tell what theory a person accepts merely by seeing a list of the sentences that she believes to be true. ⌟

EXERCISE 1.3.16 Show that the theory T' from the previous example is not complete.

EXERCISE 1.3.17 Show that $\mathrm{Cn}(\mathrm{Cn}(T)) = \mathrm{Cn}(T)$.

EXERCISE 1.3.18 Consider the signature $\Sigma = \{p\}$. How many complete theories are there in this signature? (We haven't been completely clear on the identity conditions of theories and, hence, on how to count theories. For this exercise, assume that theories are deductively closed, and two theories are equal just in case they contain exactly the same sentences.)

1.4 Translating between Theories

Philosophers constantly make claims about relations between theories – that they are equivalent, or inequivalent or one is reducible to the other, or one is stronger than another. What do all these claims mean? Now that we have a formal notion of a theory, we can consider how we might want to represent relations between theories. In fact, many of the relations that interest philosophers can be cashed out in terms of the notion of a **translation**.

There are many different kinds of translations between theories. Let's begin with the most trivial kind of translation – a change of notation. Imagine that at Princeton, a scientist is studying a theory T. Now, a scientist at Harvard manages to steal a copy of the Princeton scientist's file, in which she has been recording all the consequences

of T. However, in order to avoid a charge of plagiarism, the Harvard scientist runs a "find and replace" on the file, replacing each occurence of the propositional constant p with the propositional constant h. Otherwise, the Harvard scientist's file is identical to the Princeton scientist's file.

What do you think: is the Harvard scientist's theory the same or different from the Princeton scientist's theory?

Most of us would say that the Princeton and Harvard scientists have the same theory. But it depends on what we mean by "same." These two theories aren't the same in the strictest sense, since one of the theories contains the letter "p," and the other doesn't. Nonetheless, in this case, we're likely to say that the theories are the same in the sense that they differ only in ways that are incidental to how they will be used. To borrow a phrase from Quine, we say that these two theories are **notational variants** of each other, and we assume that notational variants are equivalent.

Let's now try to make precise this notion of "notational variants" or, more generally, of **equivalent theories**. To do so, we will begin with the more general notion of a translation from one theory into another.

DEFINITION 1.4.1 Let Σ and Σ' be propositional signatures. A **reconstrual** from Σ to Σ' is a function from the set Σ to the set $\mathsf{Sent}(\Sigma')$.

A reconstrual f extends naturally to a function $\overline{f} : \mathsf{Sent}(\Sigma) \to \mathsf{Sent}(\Sigma')$, as follows:

1. For p in Σ, $\overline{f}(p) = f(p)$.
2. For any sentence ϕ, $\overline{f}(\neg\phi) = \neg\overline{f}(\phi)$.
3. For any sentences ϕ and ψ, $\overline{f}(\phi \circ \psi) = \overline{f}(\phi) \circ \overline{f}(\psi)$, where $\circ$ stands for an arbitrary binary connective.

When no confusion can result, we use f for $\overline{f}$.

THEOREM 1.4.2 (Substitution) *For any reconstrual $f : \Sigma \to \Sigma'$, if $\phi \vdash \psi$ then $f(\phi) \vdash f(\psi)$.*

Proof Since the family of sequents is constructed inductively, we will prove this result by induction.

(rule of assumptions) We have $\phi \vdash \phi$ by the rule of assumptions, and we also have $f(\phi) \vdash f(\phi)$.

($\wedge$ intro) Suppose that $\phi_1, \phi_2 \vdash \psi_1 \wedge \psi_2$ is derived from $\phi_1 \vdash \psi_1$ and $\phi_2 \vdash \psi_2$ by $\wedge$ intro, and assume that the result holds for the latter two sequents. That is, $f(\phi_1) \vdash f(\psi_1)$ and $f(\phi_2) \vdash f(\psi_2)$. But then $f(\phi_1), f(\phi_2) \vdash f(\psi_1) \wedge f(\psi_2)$ by $\wedge$ introduction. And since $f(\psi_1) \wedge f(\psi_2) = f(\psi_1 \wedge \psi_2)$, it follows that $f(\phi_1), f(\phi_2) \vdash f(\psi_1 \wedge \psi_2)$.

($\to$ intro) Suppose that $\theta \vdash \phi \to \psi$ is derived by conditional proof from $\theta, \phi \vdash \psi$. Now assume that the result holds for the latter sequent, i.e., $f(\theta), f(\phi) \vdash f(\psi)$. Then conditional proof yields $f(\theta) \vdash f(\phi) \to f(\psi)$. And since $f(\phi) \to f(\psi) = f(\phi \to \psi)$, it follows that $f(\theta) \vdash f(\phi \to \psi)$.

(reductio) Suppose that $\phi \vdash \neg\psi$ is derived by RAA from $\phi, \psi \vdash \bot$, and assume that the result holds for the latter sequent, i.e., $f(\phi), f(\psi) \vdash f(\bot)$. By the properties of f, $f(\bot) \vdash \bot$. Thus, $f(\phi), f(\psi) \vdash \bot$, and by RAA, $f(\phi) \vdash \neg f(\psi)$. But $\neg f(\psi) = f(\neg\psi)$, and, therefore, $f(\phi) \vdash f(\neg\psi)$, which is what we wanted to prove.

($\vee$ elim) We leave this step, and the others, as an exercise for the reader. □

DEFINITION 1.4.3 Let T be a theory in Σ, let T' be a theory in Σ', and let $f : \Sigma \to \Sigma'$ be a reconstrual. We say that f is a **translation** or **interpretation** of T into T', written $f : T \to T'$, just in case:

$$T \vdash \phi \quad \Longrightarrow \quad T' \vdash f(\phi).$$

Note that we have used the word "interpretation" here for a mapping from one theory to another, whereas we previously used that word for a mapping from a theory to a different sort of thing, viz. a set of truth values. However, there is no genuine difference between the two notions. We will soon see that an interpretation in the latter sense is just a special case of an interpretation in the former sense. We believe that it is a mistake to think that there is some other (mathematically precise) notion of interpretation where the targets are concrete (theory-independent) things.

DISCUSSION 1.4.4 Have we been too liberal by allowing translations to map elementary sentences, such as p, to complex sentences, such as $q \wedge r$? Could a "good" translation render a sentence that has no internal complexity as a sentence that does have internal complexity? Think about it.

We will momentarily propose a definition for an equivalence of theories. However, as motivation for our definition, consider the sorts of things that can happen in translating between natural languages. If I look up the word "car" in my English–German dictionary, then I find the word "Auto." But if I look up the word "Auto" in my German–English dictionary, then I find the word "automobile." This is as it should be – the English words "car" and "automobile" are synonymous and are equally good translations of "Auto." A good round-trip translation need not end where it started, but it needs to end at something that has the *same meaning* as where it started.

But how are we to represent this notion of "having the same meaning"? The convicted Quinean might want to cover his eyes now, as we propose that a theory defines its own internal notion of sameness of meaning. (Recall what we said in the preface: that first-order metatheory is chalk full of intensional concepts.) In particular, ϕ and ψ have the same meaning relative to T just in case $T \vdash \phi \leftrightarrow \psi$. With this notion in mind, we can also say that two translations $f : T \to T'$ and $g : T \to T'$ are synonymous just in case they agree up to synonymy in the target theory T'.

DEFINITION 1.4.5 (equality of translations) Let T and T' be theories, and let both f and g be translations from T to T'. We write $f \simeq g$ just in case $T' \vdash f(p) \leftrightarrow g(p)$ for each atomic sentence p in Σ.

With this looser notion of equality of translations, we are ready to propose a notion of an equivalence between theories.

DEFINITION 1.4.6 For each theory T, the identity translation $1_T : T \to T$ is given by the identity reconstrual on Σ. If $f : T \to T'$ and $g : T' \to T$ are translations, we let gf denote the translation from T to T given by $(gf)(p) = g(f(p))$, for each atomic sentence p of Σ. Theories T and T' are said to be **homotopy equivalent**, or simply **equivalent**, just in case there are translations $f : T \to T'$ and $g : T' \to T$ such that $gf \simeq 1_T$ and $fg \simeq 1_{T'}$.

EXERCISE 1.4.7 Prove that if v is a model of T', and $f : T \to T'$ is a translation, then $v \circ f$ is a model of T. Here $v \circ f$ is the interpretation of Σ obtained by applying f first, and then applying v.

EXERCISE 1.4.8 Prove that if $f : T \to T'$ is a translation, and T' is consistent, then T is consistent.